LANGUAGE AND LINGUISTICS MONOGRAPH SERIES A25

Linguistic Patterns in Spontaneous Speech

Edited by

Shu-Chuan Tseng

Institute of Linguistics, Academia Sinica, Taipei, Taiwan
2009

Table of Contents

I. Spontaneity: definition and standard

II. Variation: allophones and registers

III. Prosody: feature and processing

IV. Disfluency: pattern and detection

V. Spoken dialogue: communication and recognition

List of Contributors

Sarah Borys
ECE Department and Beckman Institute, University of Illinois at Urbana-Champaign, Urbana, USA.

Ken Chen
The Genome Sequencing Center, Washington University School of Medicine, St. Louis, USA.

Jennifer Cole
Linguistics Department and Beckman Institute, University of Illinois at Urbana-Champaign, Urbana, USA.

Sébastien Cuendet
Speech Group, International Computer Science Institute, Berkeley, USA.

Yasuharu Den
Department of Cognitive and Information Sciences, Faculty of Letters, Chiba University, Chiba, Japan.

Benoit Favre
Speech Group, International Computer Science Institute, Berkeley, USA.

James Fung
Speech Group, International Computer Science Institute, Berkeley, USA.

Dafydd Gibbon
Faculty of Linguistics and Literary Studies, University of Bielefeld, Bielefeld, Germany.

Dilek Hakkani-Tür
Speech Group, International Computer Science Institute, Berkeley, USA.

Mark Hasegawa-Johnson
ECE Department and Beckman Institute, University of Illinois at Urbana-Champaign, Urbana, USA.

Amit Juneja
Senior Development Team, Think-a-Move, Ltd., Beachwood, USA.

Partha Lal
Centre for Speech Technology Research, University of Edinburgh, Edinburgh, UK.

Lin-Shan Lee
College of Electrical Engineering and Computer Science, National Taiwan University, Taipei, Taiwan.

Adrian Leemann
Hirose & Minematsu Laboratory, School of Engineering, University of Tokyo, Tokyo, Japan.
Department of Linguistics, University of Berne, Berne, Switzerland.

Che-Kuang Lin
Graduate Institute of Communication Engineering, National Taiwan University, Taipei, Taiwan.

Yi-Fen Liu
Institute of Information Systems and Applications, National Tsing Hua University, Hsinchu, Taiwan.

Kikuo Maekawa
Department of Language Research, The National Institute for Japanese Language, Tokyo, Japan.

Akira Shimazu
School of Information Science, Japan Advanced Institute of Science and Technology, Nomi, Ishikawa, Japan.

Elizabeth Shriberg
Speech Technology and Research Laboratory, SRI International, Menlo Park, USA.
Speech Group, International Computer Science Institute, Berkeley, USA.

Beat Siebenhaar
Institute of German Studies, University of Leipzig, Leipzig, Germany.

Kenji Takano
School of Information Science, Japan Advanced Institute of Science and Technology, Nomi, Ishikawa, Japan.

Jen Ting
Department of English, National Taiwan Normal University, Taipei, Taiwan.

Shu-Chuan Tseng
Institute of Linguistics, Academia Sinica, Taipei, Taiwan.

Annie Wenhui Yang
Faculty of English for International Business, Guangdong University of Foreign Studies, Guangzhou, P.R. China.

Tae-Jin Yoon
Department of Linguistics, University of Victoria, Victoria, British Columbia, Canada.

Xiaodan Zhuang
ECE Department and Beckman Institute, University of Illinois at Urbana-Champaign, Urbana, USA.

Acknowledgements

I would like to thank a number of people for their significant contribution to the preparation of this book. The majority of the papers have been selected from the papers presented at the International Symposium on Linguistic Patterns in Spontaneous Speech held in the Institute of Linguistics, Academia Sinica, in 2006. I am grateful to all the invited speakers and authors who submitted papers to the symposium as well as authors who accepted my invitation to contribute to this book. They all made the symposium a successful event and made this book possible. I want to express my gratitude to the reviewers of *Language and Linguistics* and to Professor Kathleen Ahrens and Dr. Elizabeth Zeitoun who gave useful and constructive comments on the original manuscript. I would also like to thank Ms. Joyce Kuo for all her careful and graceful work on reformatting the manuscript.

Surely, all shortcomings remain my own responsibility.

Shu-Chuan Tseng
August 2008

Overview of this volume

An introduction to this book presupposes an understanding of what spontaneous speech is. As mentioned in **Gibbon**'s chapter, spontaneous speech is a type of natural and authentic speech produced without any external influence, out of a momentary impulse of the speaker. In the literature on linguistic studies of spoken language, we have observed a shift of data types in question from isolated to connected, from solicited to non-solicited, from imitative to non-imitative, and from prepared/planned to non-prepared/ non-planned. In the field of phonetic analysis, observations and measurements were done on isolated words first, then on connected discourse (Lehiste 1972, Klatt 1975, Beckman & Edwards 1990, O'Shaughnessy 1995, Kohler 1996). Both segmental and suprasegmental characteristics have been shown to be different in isolated context and in natural speech.

Psycholinguists have long noticed the importance of natural speech, as various experiments (perception and production) and observations have been done on spontaneous speech (Goldman-Eisler 1968, Levelt 1989). Early linguistic development of normal-hearing and hearing-impaired children has been investigated on both imitative and spontaneous speech data (Musselman & Kircaal-Iftar 1996, Leonard et al. 1981). Moreover, prosodic analysis deals with natural speech directly. Issues such as intonation contour patterns, prosodic annotation, prosodic phrasing and marking are complex, when spontaneous speech is involved (Pierrehumbert 1980, Levelt & Cutler 1983, Fowler & Housum 1987, Nakatani & Hirschberg 1994).

For a number of research disciplines, the form under investigation commences with connected, natural, spontaneous speech. Conversation analysis is one of these fields (Du Bois et al. 1993, Sacks et al. 1974, Schegloff 1982). Works on disfluency are also based on spontaneous speech data, though the focus may be on speech perception, the patterns in natural speech with regard to speech technology, or the indication to discourse structure (Shriberg 1994, Lickley 1996, Swerts 1998, Tseng 1999). As a whole, discussions of spontaneous speech are not new at all. But it has become more popular to use the term "spontaneous speech" in the discipline of speech technology in the last two decades or so, mainly as a contrast to read speech. Out of this recent development, the main aim of this book is to accommodate different research interests involved in the disciplines of linguistics and speech technology.

Shared data resources are an ideal start for interdisciplinary co-operation. Linguists working on speech analysis usually possess a certain size of speech data to do their research. Engineers working on spontaneous speech, in contrast, have to develop a relatively large amount of language resources to train and test algorithms and systems. For instance, the Corpus of Spontaneous Japanese (CSJ), in total 662 hours of data, is jointly developed by the National Institute for Japanese Language (NIJL), the National Institute for Information and Communications Technology, and Tokyo Institute of Technology. In this book, **Maekawa** shows how consistently constructed spoken language resources of a reasonably large data size contribute to studies of language variations such as vowel devoicing, accented particles, word coalescence, and boundary pitch movement. Similarly, variations of word form can also be thoroughly and consistently examined, which significantly contributes to natural language processing tasks such as word frequency distribution and entropy of word-forms. The CSJ has now been used for training speech processing systems and also for linguistic investigations into patterns of different aspects in spontaneous Japanese.

A different issue directly related to spontaneous speech is that a language may be performed differently in the written and in the spoken form. Of particular interest is the discussion of the match and mismatch of prosody and syntax. However, it would be too simple to argue that there is a written grammar and a spoken grammar for a language. Whether the spoken language has its own sentential types can only be investigated if we analyze text and speech language resources and compare them closely. **Ting** in her chapter manages to utilize written and spoken resources to investigate the use of the frequent grammatical particle *suo* in Mandarin Chinese. The text corpus consists of editorials, magazines and fiction. The three spoken corpora she used are all spontaneous Mandarin conversations, but collected in different scenario settings (a brief introduction to these corpora can be found in §2 in this paper). She finds that it is not the genre which results in the difference of use. The situational characteristic is the main determinant. This conclusion is supported by the written and the spoken data. **Yoon**'s study turns our focus from syntax to phonetics again. He nicely shows that phonetic knowledge not only in theory but in praxis helps the performance of an automatic speech recognition system. Making use of the well-known and also widely-used English spontaneous speech corpus, the SWITCHBOARD, Yoon finds that allophones can be distinguished in terms of their harmonic structure and the autocorrelation ratio of each phone. This has been identified by using twelve hours data out of the 300 hours SWITCHBOARD corpus. Furthermore, a recognition dictionary with allophones accommodated was built and a voice quality dependent speech recognition experiment was run to verify the effect of allophonic variations. The word accuracy rate actually improved. This demonstrates that linguistic features or patterns can be concretely incorporated into systems for developing speech technology.

Another outstanding example of integrating linguistic knowledge into automatic speech recognition systems is shown by **Hasegawa-Johnson et al.** for English. They have nicely integrated prosodic context, disfluency, and voice quality into the feature set of the defined phone inventory. The features are 1) position within intonational phrase, 2) phrase prominence, 3) position within prosodic word, 4) N-phone context, 5) lexical stress, 6) syllable position, 7) fluency, and 8) voicing. In various recognition experiments, the word error rate can be reduced by adopting a prosodic hierarchy as an organizing framework for the speech sources. This shows that prosodic phrasing in spontaneous speech is regarded as a substantial cue for language planning and prosodic organization. In terms of the regional differences found in prosody use, **Leemann & Siebenhaar** examined two dialects spoken in Switzerland, Bernese and Valais. Their study focuses on the use of pauses (filled and empty) and the prosodic feature of the filled pause itself, the initial phrase after the filled pause, and the rest of the phrases. First, a difference is found in the overall use of pauses between these two speaker groups. Second, they found a low-high-low F0 pattern of the filled pause, the following phrase, and the rest of the phrases in both regional dialects. They point out an issue which is relevant for conversation analysis, as this prosodic nuance may influence the perception of the listeners to interpret the discourse meaning of the filled pauses. Also, it is relevant for automatic speech recognition work with regard to how to detect occurrences of disfluency such as filled pauses. To examine prosodic features and disfluency, a more direct way to deal with prosodic phrasing in spontaneous speech is proposed by **Liu & Tseng**. A set of spontaneous Taiwan Mandarin speech data has been annotated in terms of their prosodic boundaries by adopting an IU-like notion of prosodic units. An automatic POS tagging experiment and a cue phrase experiment both show that lexical items which tend to appear at prosodic boundaries are not only relevant from the syntactic and pragmatic points of view, they are in fact also prosodically marked. The prosodic marking may appear in the form of pitch reset, lengthening, pause, or speech rate alternation. Thus, the issue of incorrect word segmentation and POS tagging results caused by disfluency can be dealt with by utilizing prosodic segmentation.

Prolongation is another disfluency phenomenon in spontaneous speech. **Den** investigates prolongation of the clause-initial mono-word phrases (CIMWP) in spontaneous Japanese, also making use of the CSJ data. Syntactic and acoustic features of CIMWP occurrences, as well as the duration feature of the clause-initial conjunction *de* were studied. Particularly interesting was the fact that the duration of the preceding pause and the presence of a succeeding pause have the same influence in both cases. The features found in both analyses do not vary significantly in different genres of CSJ: APS (Academic Presentation Speech) and SPS (Simulated Public Speaking). According to Den’s conclusion, it is unlikely that the cognitive load plays a role in the production of

CIMWP, since the features behave similarly in these two genres, although different degrees of cognitive load were expected. In order to deal with disfluency, duration-related features and pitch-related features both play important roles. However, they are relevant for different types of disfluency. In an experiment on spontaneous Taiwan Mandarin speech, **Lin et al.** used Latent Prosodic Modeling (LPM) to detect the interruption point (IP) within disfluencies. They defined two sets of acoustic parameters: pitch-related and duration-related features. The result shows that pitch-related features influence the IP detection of overt repairs, while direct repetitions are associated more with the duration-related features. This suggests that different disfluency types may be marked by different sorts of prosodic means. Interestingly, **Shriberg et al.** show that prosodic marking in spontaneous speech may also be similar across different speaking styles when identifying/classifying dialog act boundaries. Two corpora of different speaking genres, MRDA (Meeting Recorder Dialog Act) and BN (English Broadcast News) are used to automatically classify a set of defined dialog acts, i.e. whether it is a boundary or a non-boundary. Duration, pitch and energy features at boundary/non-boundary in both corpora are compared based on the automatic classification result. It is found that speakers in meetings use relative lengthening for boundaries as frequently as speakers in news broadcasts (after the duration distribution is normalized). Pitch and energy features behave similarly across speakers in both corpora. This indicates that the same set of classifiers for dialog act boundaries could be developed to process speech corpora of different speaking styles.

Although disfluent pauses may not be consciously intended by the speaker, pragmatic silences, discussed by **Yang**, are usually adopted following a number of principles of how and when silence is necessary and appropriate for certain purposes. She concludes that silence functions as a verbal speech act, where silence can enforce the addressee to do things and can be understood in specific culture and context. In addition, it is not easy to teach systems all the rules used by human beings in real life. So **Takano & Shimazu** defined a modified set of local dialog acts and adopted the decision tree method for their machine learning algorithms. The task is to recognize the dialog structure in Japanese route guidance dialogs. The result shows that a combination of the local dialog structure with the previous dialog act leads to better results than just the previous dialog act. Compared with **Yang**'s study, it may be the case that for conversation and communication, not only linguistic rules and markings are required, but cultural and contextual constraints and understanding are also indispensable. A system may need to learn the linguistics of a language and the culture of the language speakers' community to recognize the dialog structure and meaning (Grosz & Sidner 1986).

Spontaneous speech is the most frequently used speech form. Language experiences, cultural conventions, and also background knowledge about topics and speakers are necessary for human beings to understand and produce spontaneous speech. It seems to be a simple task for human beings; listening, processing, searching, planning, and speaking all happen in an extremely short period of time. But it is difficult to build a system which is capable of doing similar tasks. The contributions of this volume have revealed the importance of shared language resources, as corpus creation and sharing provide system developers with a good basis for their development and testing. But how much data is enough data? It is likely that the number of features human beings employ to analyze speech is finite, so it is not a real problem if we don’t have a billion words of spoken data. Maybe the problem is how we should learn to recognize the knowledge and capabilities we use to solve problems and carry out tasks. Understanding and producing spontaneous speech is only one of the many tasks we perform everyday. The solution to this problem lies not only with engineers or linguists. An interdisciplinary approach is needed. Hopefully, this volume fulfills this purpose.

Linguistic Patterns in Spontaneous Speech, 1-25
2009-1-025-001-000263-1

Can there be Standards for Spontaneous Speech? Towards an Ontology for Speech Resource Exploitation

Dafydd Gibbon
Universität Bielefeld

1. Speech resources, standards and spontaneous speech

By *speech resources* in this contribution means relatively homogeneous audio and visual speech corpora, including recordings, transcriptions, annotations, metadata, and perhaps associated word lists and corpus-based language models. The concept of *spontaneous speech* is in the centre of discussion in the following sections. The strategy taken in this presentation is to embed the characteristics of spoken language into a broader functional and structural linguistic context.

1.1 The problem

At first glance, the terms “standards” and “spontaneous” would seem to contradict each other. But spontaneity is complex and ambiguous, and speech which is spontaneous in one sense (not being read, not being consciously planned or rehearsed) may be quite non-spontaneous in another (in using sociolinguistically restricted codes, clichés, conventionalized phrases and idioms). The high dimensionality of the family of genres which might be called spontaneous speech is gradually becoming clearer as databases of spontaneous speech are being collected, and applications of resources—particularly in speech synthesis—are being made, and emotional speech and multimodal speech are becoming commonplace objects of investigation. And on the other hand, the standards referred to are metatheoretical guidelines for language and speech information interchange, not prescriptive instructions for human behaviour, though of course they may be interpreted as prescriptive specifications for speech technology application development.

In the present contribution, the main focus is to specify linguistic and phonetic background infrastructure for designing and using resources for spontaneous speech. Speech resources are generally purpose-built, whether they are intended for speech technology applications at one end of the scale or for conversation analysis at the other, but it has long been recognized that such data can have applications beyond the original purpose, and therefore mechanisms for sharing the data are required.

One such mechanism is ontology-based search and analysis. Ontologies in this sense are heuristic classification systems, including taxonomies, meronomies and other network structures, and were developed for inference in expert systems in Artificial Intelligence work in the 1980s. Recently, ontologies for language and speech have been developed, such as GOLD (General Ontology for Linguistic Description—Farrar & Langendoen 2003), and have become rather popular. But for the speech domain nothing comparable exists.

Assuming that such an ontology is just as necessary for speech—at the current state of search technology—as in other domains, an initial ontology, and surely a highly controversial one, is presented for discussion, concentrating on prosody (particularly pitch systems and timing) and disfluencies.

But although ontologies may be heuristically motivated and usually task-driven, repeated re-invention of the wheel in the area of speech resources for different task areas can be—again, heuristically speaking—rather a waste of time, energy and funding. So the strategy taken in the present contribution is to take applications-driven constraints into consideration, but to take a step back and look at more generic issues involved in the creation of a re-usable ontology.

1.2 Standards and spontaneity—a contradiction?

On the one hand, resources need to be standardized for information exchange, computational processing, and linguistic and phonetic analysis, otherwise they are of no use. At first glance, the terms “standards” and “spontaneous” would seem to contradict each other. But spontaneity is complex and ambiguous, and speech which is spontaneous in one sense (not being read, not being consciously planned or rehearsed) may be quite non-spontaneous in another (in using sociolinguistically restricted codes, clichés, conventionalized phrases and idioms).

The high dimensionality of the family of genres which might be called spontaneous speech is gradually becoming clearer as databases of spontaneous speech are being collected, and applications of resources—particularly in speech synthesis—are being made, and emotional speech and multimodal speech are becoming commonplace objects of investigation. Standards, on the other hand, reduce basic parameters and values to a manageable set, and are not intended to be completely comprehensive, whether institutional standards such as ISO standards, or *de facto* industry standards such as email formats, popular operating systems, or computer types, or academic standards such as the International Phonetic Alphabet for transcription (see Gibbon et al. 1997, Gibbon et al. 2000 for discussion), or the EUROTYP[1] or WALS[2] category sets for descriptive linguistics (Haspelmath et al. 2005).

The purpose of this paper is not really to provide a definitive overview of the field. This would be a useful task, and coherent consolidation of previous results is very necessary. The purpose is more forward-looking, projecting experience in a number of projects and personal research ventures into a research space for the future. Furthermore, in view of the overall context of this contribution, the aim is to point towards underlying linguistic (descriptive, formal, computational) requirements specifications for a re-usable ontology of spontaneous speech, rather than towards operational requirements for a specific ontology of spoken language systems which are concerned with particular aspects with spontaneous speech or particular speech technology applications in this area. For these two reasons, the references will also be somewhat selective.

1.3 Overview

Section 2 is concerned with the specification phase of developing an ontology for spontaneous speech. Section 3 is concerned with the concept of ontology, and current contributions to discussion in this field. Section 4, starting from the basic premise that phoneticians, linguists, speech technologists all deal with the domain of signs, looks at linguistic and semiotic requirements for such an ontology, with particular attention to the adequacy of basic models of sign structure. In section 4, a brief overview of selected current discussion on the development of ontologies for spoken language (and language in general) is given. Section 5 introduces a generic structure for a re-usable ontology, which takes the content, the structure and the rendering of signs into account. Section 6 examines contributions of speech specific categories.

2. Specifying resources for spontaneous speech

This section is concerned with taking a close look at specifying resources for spontaneous speech. Speech resources are generally purpose-built, whether they are intended for speech technology applications at one end of the scale or for conversation analysis at the other, but it has long been recognized that such data can have applications beyond the original purpose, and therefore mechanisms for sharing the data are required.

[1] http://wwwlot.let.uu.nl/Research/ltrc/eurotyp/index.htm

[2] http://wals.info/

2.1 On defining "spontaneity"

It is a long time since the Text Encoding Initiative produced initial recommendations for the transcription of speech corpora; in fact this was before the blossoming of the linguistic resources and language documentation paradigms, and before the development of computational methods of dialogue modelling (recently adaptations for XML have been made, but without substantive extension). So let us take a step back and ask a basic question: what do we mean by "spontaneous"?

The Merriam-Webster online English dictionary provides the following definitions:

Etymology: Late Latin spontaneus, from Latin *sponte* of one's free will, voluntarily
1: proceeding from natural feeling or native tendency without external constraint
2: arising from a momentary impulse
3: controlled and directed internally: SELF-ACTING <spontaneous movement characteristic of living things>
4: produced without being planted or without human labor: INDIGENOUS
5: developing or occurring without apparent external influence, force, cause, or treatment
6: not apparently contrived or manipulated: NATURAL

The term is evidently highly polysemous. Nevertheless, all these readings point in the same direction as the term *authentic* as it is currently used in foreign language teaching methodology. In this context, authentic texts are simply texts which are not produced for the purpose of language study.

In speech technology, linguistics, phonetics and psycholinguistics, a number of definitions have traditionally been offered, which were summarized in a *Linguist List* discussion over a decade ago (Fagyal 1995), which has lost none of its relevance:

'Spontaneous speech' is a
(1) type or 'mode' of speech production opposed to 'read-aloud' speech;
(2) real-time generated, unplanned and non-rehearsed type of encoding linguistic information;
(3) casual 'way of speaking' or 'style', characterizing informal speech situations;
(4) naturally occurring, non-experimental type of speech event of any kind.

These definitions are clearly narrower. The second and the fourth definitions correspond to the general dictionary definition of *spontaneous*, the third is the one which a linguist concerned with language varieties would choose. The first definition is, however, probably the definition which is most common in the phonetic, psycholinguistic or speech technological laboratory. The first definition is a compromise, and very incomplete. There is, after all, such an activity as spontaneous read-aloud speech, for instance when one reads an extract from some text, whether a newspaper or a menu, to a companion.

2.2 Spontaneity or authenticity?

As already noted, the closest definition to what we need appears to come from foreign language teaching, which of course has centuries of experience with which the few decades during which speech technology has been around cannot compete. So let us take the notion of *authentic text* and its spoken language twin *authentic speech* to be what we are looking for:

> *Authentic speech is speech which is not produced for the purpose of the study of speech.*

This definition should serve us in good stead, as long as we bear in mind that for many purposes laboratory speech is *authentic laboratory speech*...

But now we have a problem: authentic speech is simply everything, and thus needs delimitation. One of the delimitation strategies which have been used during the past few years is to use *emotional speech*. But consider the range of uses of spoken language—from spontaneous discussions among academics to motherese between mothers and children, and from chance conversations between strangers to everyday talk between married couples. All of this can be classified as authentic speech, as spontaneous speech, but the main classification of these speech registers or styles is rarely “emotional vs. unemotional” speech. A further problem here is that so-called “emotional speech” is often, rather, emulated emotional speech based on stage conventions learned by professional actors. So we need to look for a parametric space within which speech instances are located.

3. Linguistic requirements for an ontology of signs

One strategy for restricting the parametric space for domain descriptions is ontology-based search and analysis; this section is concerned with characterizing the term “ontology”, and in exploring linguistic models of language structure in search of an appropriate set of foundational categories.

3.1 On defining "ontology"

The term "ontology" is ambiguous. A general definition of *ontology* in its traditional sense can be found in Floridi (2003:155):

> Ontology as a branch of philosophy is the science of what is, of the kinds and structures of the objects, properties and relations in every area of reality.

And of course as empirical scientists, we find this notion somehow appealing, until we realize that *reality* is itself an elusive concept. A homogeneous definition is perhaps even less easy to find in contemporary computational usage, however. Floridi (2003: 158) provides a very general definition:

> In the field of information processing there arises what we might call the Tower of Babel problem. Different groups of data-gatherers have their own idiosyncratic terms and concepts in terms of which they represent the information they receive. When the attempt is made to put this information together, methods must be found to resolve terminological and conceptual incompatibilities. Initially, such incompatibilities were resolved on a case-by-case basis. Gradually, however, it was realized that the provision, once and for all, of a common backbone taxonomy of relevant entities of an application domain would provide significant advantages over the case-by-case resolution of incompatibilities. This common backbone taxonomy is referred to by information scientists as an 'ontology'.

The term *taxonomy* is not meant here in the restricted sense of lexical semantics: a hierarchy of classes or categories which is defined by relations of implication. It is meant more in the sense of *semantic network*, that is, a system of categories linked by several kinds of relation, including that of implication, but also including part-whole relations, temporal and spatial relations, and other kinds of relation.

Ontologies in this sense are classification systems, including taxonomies and other network structures, and were developed for inference in expert systems in Artificial Intelligence work in the 1980s.

The E-MELD project definition is useful (Anon 2005):

> An ontology here is essentially a machine-readable formal statement of a set of terms and a working model of the relationships holding among the concepts referred to by those terms in some particular domain of knowledge. Its purpose is not to define meaning, but to allow computers to navigate human knowledge

in a way that mimics intelligence.

But let us de-mythologize the term even more thoroughly:

> An ontology is a highly structured terminological dictionary designed to facilitate search for information in some technical domain.

It should be emphasized that in general ontologies are heuristically motivated: they do not impose constraints on new theoretical requirements which may arise, but need to be flexible enough to accommodate such requirements for practical search purposes at some level of conceptual granularity.

3.2 GOLD: General Ontology for Linguistic Description

The most well-known ontology for linguistics is currently the *General Ontology for Linguistic Description* (GOLD[3]), developed by Farrar & Langendoen (2003) as part of the metadata standardization effort of the E-MELD[4] (Electronic Metastructures for Endangered Languages Data) project. This ontology, and its motivation, is a good example of why ontologies are important for scientific activities: the more resources there are, the harder it is to search for and find relevant resources, and the more necessary it is to describe resources in a heuristically useful agreed vocabulary. But for the speech domain nothing comparable exists. The essential features of GOLD are conveyed in this definition from the GOLD website:

> GOLD is an ontology for descriptive linguistics. It gives a formalized account of the most basic categories and relations (the “atoms”) used in the scientific description of human language. First and foremost, GOLD is intended to capture the knowledge of a well-trained linguist, and can thus be viewed as an attempt to codify the general knowledge of the field. GOLD is aimed at facilitating automated reasoning over linguistic data and at establishing the basic concepts through which intelligent search can be carried out.

In its current state, GOLD has a number of weaknesses:

1. The methodology is heavily slanted towards particular structuralist and generative descriptive methodological traditions, and is acknowledged to be in need of

[3] http://emeld.org/gold
[4] http://emeld.org/index.cfm

extension to include functionalist or diachronically oriented categories and relations.

2. In terms of linguistic unit size, GOLD is restricted to the traditional word constituents of phoneme and morpheme, and to sentence constituents. Text and dialogue modelling, which are crucial for an adequate ontology for spoken language, is not included.
3. The domains of inter-modality relations, in particular the role and structure of prosody and conversational gesture, are not included. Indeed, phonology and phonetics in general are also not well represented, though initial proposals have been made by Aristar (2005) for phonetics and Kamholz (2005) for phonology.
4. More fundamentally, GOLD pays little attention to the functionality of structures and their constituents. The functions of spoken language and its compositional or idiomatised parts, are at the core of an integrated theory of language.
5. Finally, as with other current ontologies, GOLD is biassed towards the analysis of text—whether text in the sense of written language, or text in the sense of traditional linear phonetic and phonological transcriptions, whereas spoken language requires compositional operations of overlap (also known as association, alignment, or parallelism) in addition to concatenation.

The last point is worth dwelling on a little. Formally, written texts are analyzable by means of concatenation grammars, at least until the domain of text and *Document Type Description* is reached. Concatenation is also adequate for the description of strings of phonemes, morphemes, words, sentences and so on, and can be used to define syntagmatic hierarchies over sequences of units (i.e. constituent, part-whole hierarchies, as opposed to classificatory taxonomic hierarchies). In the context of texts, concatenation can be interpreted as immediate proximity within a stylized spatial coordinate system. In the context of speech, concatenation can be interpreted as immediate proximity within a stylised temporal coordinate system with a relation of *temporal precedence* (corresponding to concatenation).

But also, in speech, a relation of *temporal overlap* is also needed, in order to express prosodic or “supra-segmental” and gestural facts of many kinds, from vowel harmony through intonation to the co-occurrence of emphatic accentuation and emphatic gestures. The range of overlap functions is broad, and may be based on “infra-segmental” structures (as in assimilation patterns) or on relatively independent “autosegmental” structures (as in tonal and intonational patterns or co-occurrent conversational gesture). In fact, a *spatial overlap* relation also holds for writing systems in the visual domain, in terms of the “prosody” of writing, such as highlighting and layout, though this is rarely dealt with in linguistic studies.

3.3 Linguistic essentials for a spontaneous speech ontology

Not all of the points listed as weaknesses of the current version of the GOLD model can be dealt with in the present contribution. The main points covered are:

1. Formal, structural features of spoken language, with particular reference to rank hierarchical structure and to prosody.
2. A ternary model of the relation between signs and the world, the *Content-Structure-Rendering* (CSR) model, which differs from traditional dualistic sign models in introducing an intermediate component of Structure, and in adopting a dual interpretation of Structure in terms of the world of Content and the world of Realization, following modern linguistic theories.

By Realization is meant the form of signs in different modalities, for example the acoustic realization of signs as speech sounds, and the visual realization of signs as gestures or as text.

In the following subsections, an outline of an initial taxonomy is established, and in the following sections functional aspects and the CSR model are outlined.

3.3.1 Linguistic categories with prosodic relevance

1. Basic category: speech event, a pair of a category and an interval
2. Structural levels of analysis, ranks (units of increasing size)
    1. Phoneme/toneme; syllable
    2. Morpheme/morphophoneme/morphotoneme
    3. Word (simplex, derived, compound)
    4. Phrase, sentence
    5. Text
    6. Dialogue
3. Semantic and pragmatic interpretation (for “concept annotation”)

3.3.2 Linguistic relations

1. Syntagmatic relations in speech:
    1. Sequential (concatenative and hierarchical) relations
    2. Parallel (autosegmental association) relations, including synchronization issues (“absolute slicing”, phonetic operations such as assimilation)
2. Paradigmatic (classificatory) relations of similarity and difference:
    1. Dependent on classification by sequential relations

    2. Dependent on classification by parallel relations
3. Interpretation relations:
    1. Manifestation relations (modality/media oriented):
        1. Acoustic
        2. Visual
    2. Content relations (semantics/meaning/function oriented):
        1. Contrast (phonology, Asian and African tonology)
        2. Morphemic (morpho-syntax, African tonology)
        3. Text, discourse

4. Semiotic bases for a re-usable spoken language ontology

Speech is more than arbitrary patterns of sound; to be speech, the patterns need to be accepted behaviour of a community, and they need to be interpretable in regular ways in terms of the needs of the community for information and action. This section is concerned with the embedding of basic structural concepts into functional frameworks, and outlines a number of traditional approaches for this purpose.

4.1 The Saussurean dualist conceptualist model

The first modern linguistic approach to modelling speech, both from a structural and a contextual perspective, was by the father of modern linguistics, Ferdinand de Saussure. A representation of his model, which is mentalist, in conceiving the sign in terms of mental thoughts and images, and also dualist, in structuring the sign into two parts, the *concept*, concept or thought, which refers to the meaning or *signifié* ("signified") of the sign and the *image acoustique*, acoustic image, which relates to the form or *signifiant* ("signifiant") of the sign. The model is shown in Fig. 1.

A further interesting feature of the model (which is often forgotten in the literature) is that de Saussure's mentalism is of a specific kind: the mind is understood as a "collective subconscious", coordinated by means of a circuit of sign exchange between members of the speech community, and thus introducing the notion of a *channel* or contact between interlocutors; for the significance of the *channel* for the functionality of prosody cf. Gibbon (1976).

The dualist approach provides the minimum number of components for a re-usable ontology which takes the notion of sign seriously.

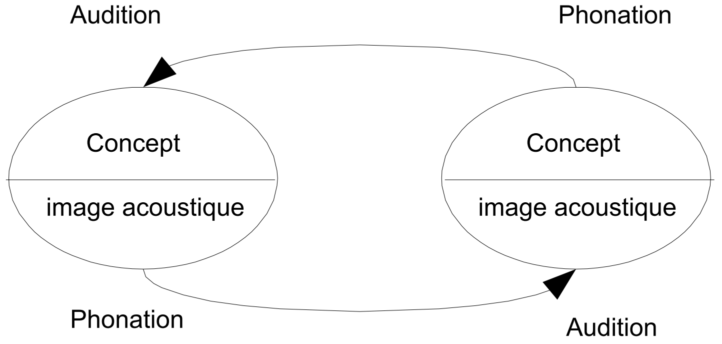


Figure 1: Dualist conceptualist model (de Saussure)

4.2 The Praguean functionalist constitutive factor model

A decade and a half later, in 1934, the psychologist Karl Bühler developed a four-component modification of the Saussurean circuit model in which the *sign* component was abstracted away from the individual interlocutors and presented as an entity which stood in a functional relationship with other components of the speech situation, the transmitter, yielding the expression function (Ausdrucksfunktion), the receiver, yielding the appeal function (Appellfunktion), and the context, yielding the representation function (Darstellungsfunktion) of language, as illustrated in Fig. 1.

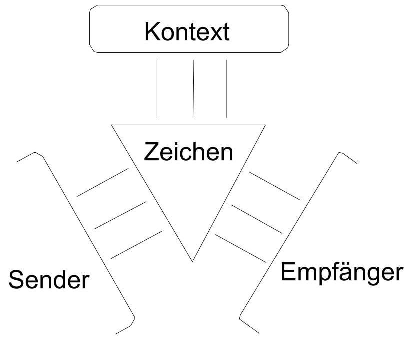


Figure 2: Functionalist instrumental model: Zeichen = sign,
Sender = transmitter, Empfänger = receiver, Kontext = context (Bühler).

Almost 30 years later, in 1960, an Praguean extension of Bühler's model to six components was presented by Roman Jakobson, in which sign was renamed "message", and additionally the channel was made explicit as the "contact" (not only the physical channel, but also the communicative bond between the interlocutors), and the code. Like Bühler, Jakobson related the functions of language to the components of the communication situation, which he termed "constitutive factors": expressive (Sender), conative (Receiver), representational (Context), corresponding to Bühler's components, and metalingual (Code), poetic (Message) and phatic (Contact). The model is shown in Fig. 3. As these models were developed, more situational factors were introduced which are relevant for the description of spoken language, and in particular prosody: the metalingual function subsumes the configurative and delimitative functions of prosody in relation to locutionary structures (cf. accent positioning, contours, and boundary tones), and the phatic function has already been referred to in connection with calling intonations (Gibbon 1976).

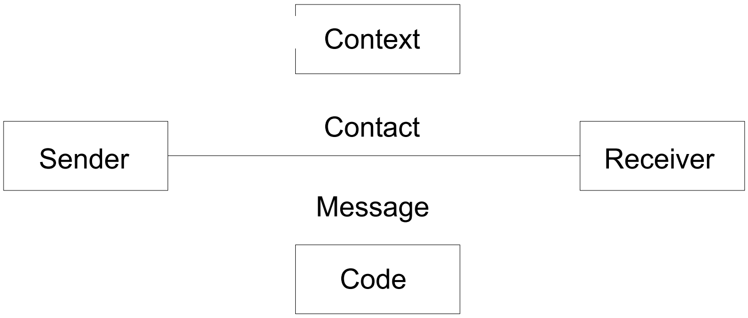


Figure 3: Functionalist constitutive factors model (Jakobson)

4.3 The Carnapian inclusion hierarchy model

A model with a different perspective was introduced by Rudolf Carnap (e.g. Carnap 1958:79), which has remained a kind of standard model in logic and linguistic semantics and pragmatics. The study of the structure of the sign, syntax, is embedded in the study of the meaning of the sign, semantics, which is in turn embedded in the study of the use of the sign by interlocutors, pragmatics. This model is not unrelated to the constitutive factor models, but basically prioritizes the components; it is no accident that Carnap and Bühler were contemporaries in Vienna. The syntax of the sign, i.e. the grammar of the forms and structure of the sign is the formal basis for a full description of the 'semiotic' of the sign, in Carnap's terminology. The context, with the representational function, semantics, is more encompassing, while pragmatics, encompassing the speaker and the hearer, provides the comprehensive environment for the other components.

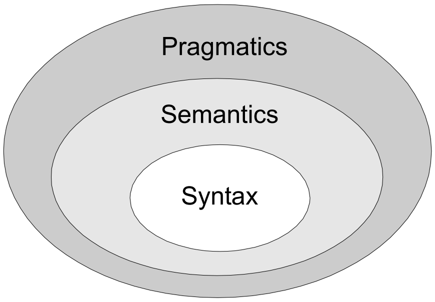


Figure 4: Pragmatic-Semantic-Syntax Inclusion model

It is noteworthy that while the other linguistic models gave much attention to syntax, with the exception of Jakobson, the introduction of syntax into the model, the internal structure of the sign, still does not do justice to essential features of spoken language: pronunciation, i.e. sounds, syllables, and the prosodic hierarchy.

4.4 The system cascade model

The Carnapian inclusion model has been re-interpreted in countless approaches to language modelling in the human language technologies as a cascade: for generation, pragmatics comes first, semantics follows, and syntax is the final stage. Again, there is no place for essential components required in a re-usable ontology for spoken language, from concepts of ritual, routine and idiomaticity, through the lexicon to the written-spoken distinction itself and, with it, phonetic interpretation and the prosodic hierarchy. The cascade model in its simplest form is shown in Fig. 5.

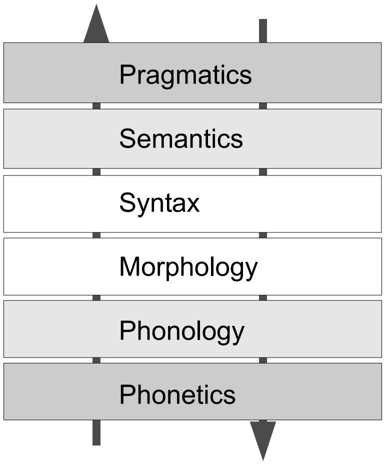


Figure 5: Cascade model of spoken language system architecture

The idea behind this model is that speech production starts with pragmatics, runs through semantics, syntax, morphology and phonology until it reaches the phonetic output. Conversely, speech perception and understanding starts with phonetics and proceeds in the other direction through phonology, morphology, syntax and semantics until the full pragmatic interpretation in context is reached.

This model is very useful for many purposes, and has been used on many occasions, in this or in more elaborated form. But from the point of view of a re-usable ontology it is flawed in more than one respect:

1. The relations between the components are very different: while sentences may be said to be composed of words (syntax), and words of morphemes (morphology), and, though more indirectly, that morphemes are composed of phonemes (phonology), it is not at all obvious that phonemes are composed of phonetic units, such as features or articulation phases in the same way. Neither is it obvious that pragmatic units are composed of semantic units and that semantic units are composed of syntactic units in the same way. Different relations are involved.
2. Syntax and morphology pertain to the structure of the sign, and are very different kinds of entity from semantics and pragmatics, which relate the sign to its real-world environment, and from phonology and phonetics, which give the sign its tangible real-world form.
3. The model is sentence and word oriented, and implies compositionality. There is no room for higher levels of structure such as text and dialogue, which are, in some indeterminate way, consigned to a catch-all pragmatics component.
4. There is no room for a distinction between compositionality of signs and idiomatization or routinisation of signs, which (among other things) are important for the assignment of intonation contour types.
5. There is no room for the incorporation of a prosodic hierarchy, in the sense of a mapping of locutionary units of different sizes into configurational trajectories of phonetic features of different sizes, for instance major and minor intonation phrases, foot units, syllables.

5. The CSR model: content, structure and rendering

The critique of the cascade model in the previous section provides the foundation for a more viable and differentiated model, which is likely to serve better in a re-usable ontology of spoken language. The present section discusses the ternary Content-Structure-Rendering model before the background of *co-interpretation* of signs in terms of the world of content and the world of appearances of signs. An initial discussion of the lexicon in terms of this ternary model can be found in Gibbon (2002).

5.1 Semiotic co-interpretation: the semantics of domains and modalities

A set of core formal metatheoretical concepts will be introduced here. The three part-whole relations of *precedence*, *overlap*, and *hierarchy* constitute what are known as *syntagmatic relations* (structure-building) in linguistics, in contrast to *paradigmatic* or *classificatory* relations (of similarities and differences between categories). Both these types of relation have been formalized in so-called *feature grammars* as attribute value pairs. Attributes are in general taken to represent functional parts of a larger structure which is represented by an attribute-value matrix, while sets of values of attributes represent the partitions and equivalence classes which characterize paradigmatic relations.

It is this complex of relations which determines the *levels of analysis* in linguistics, for example the realization of phonemes as allophones, of morphemes by phonemes, of (some) morphosyntactic categories by tone in tone languages, and of utterance and discourse categories by intonation. Formal accounts of this complex of relation are given in Generative Phonology, in phonological Optimality Theory, Declarative Phonology, Computational Phonology, and Two-Level Morphology, for example. A theoretically founded and operational computational model of part of this complex of relations is to be found in the *Time Map Phonology* of Carson-Berndsen (1998).

The structure of an elementary sign with the co-interpretation architecture is shown in Fig. 6.

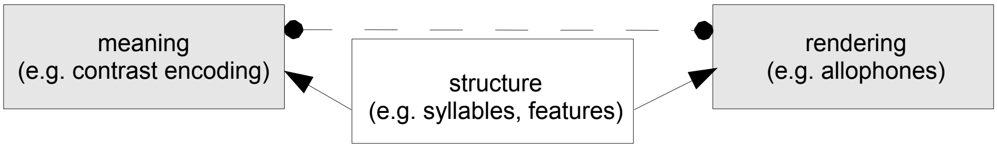


Figure 6: Sign structure with domain and modality co-interpretation

5.2 Co-interpretation in the hierarchical CSR architecture

The architecture of the CSR model is illustrated in Fig. 7. The core of the model is a hierarchical model of *ranks* of signs which not only differ in size but also in functionality. At each level, local kinds of hierarchy are also present: syllable structure, morphological structure, phrase structure, text structure and dialogue structure.

The sign units are represented in the middle column of the figure, and constitute a rank of inclusive units with inclusive functions. At each rank, there are two interpretations, one, domain interpretation, in terms of content or function, and the other, modality interpretation, in terms of rendering or appearance. The sign is therefore an abstract (or mental) unit which relates in these two ways to reality: the content domain and the modality domain. The linguist can discourse about any part of this model, of course: in this metalinguistic discourse, the entire model—content, structure and rendering—then becomes part of the metalinguistic content domain, so in a sense the modality domain is, at a very general level, always potentially a part of the content domain in metalinguistic discourse.

With regard to the content domain, there are many appropriate theories of formal and functional semantics, pragmatics, conversation description and modelling which are easily accessible and which are not of concern here. In the centre of attention is the specific formal structure of spoken language.

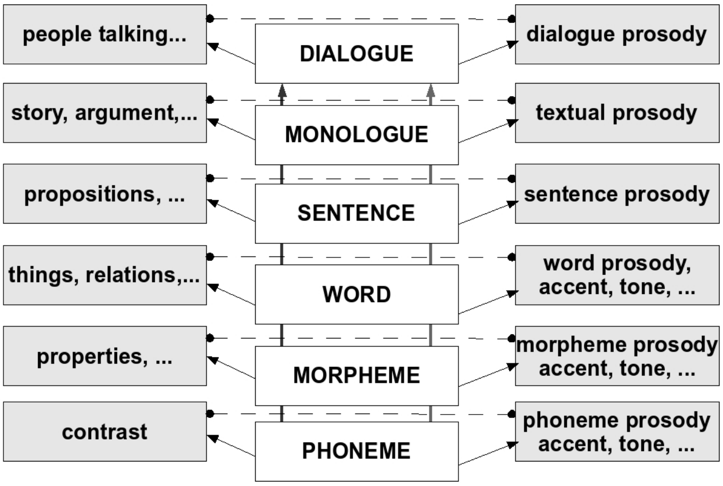


Figure 7: Content-Structure-Rendering (CSR) architecture

5.3 Structure: metasyntactic meronomies and taxonomies

These very basic relations, which are to be found in much introductory linguistic literature, but are rarely made explicit either in theoretical or applications work on linguistic matters, are primary requirements for a linguistic ontology for speech. Summarizing, the basic relations which determine linguistic levels of analysis required for describing speech resources are:

1. Syntagmatic relations (sequential and parallel) determining part-whole relations in complex constructions.
2. Paradigmatic relations (categories, classes) determining similarities and differences.
3. Interpretation (manifestation) relations (the modality and semantic-pragmatic interpretations) determining time types.

Failure to recognize these elementary distinctions has led to much confusion in phonological theory over the years, but in a workable linguistic ontology the distinctions are essential at all levels of description.

5.3.1 Syntagmatic (meronomic) relations

1. In syntax, sequential syntagmatic relations are expressed by labels such as *SUBJECT*, which represent sequential parts within a larger attribute-value matrix; overlap relations are generally ignored, but would be needed in accounts of intonation, accentual focussing, and morphosyntactic tone.
2. In phonology and prosody, sequential syntagmatic relations are expressed by labels such as *ONSET* (of syllables), *NUCLEUS* (ambiguous—of syllable or of intonation group), *ACCENT* (associated with syllable nuclei), which represent either sequential parts (syllable *ONSET* or *NUCLEUS*, intonation *NUCLEUS* and *ACCENT* with respect to other parts in intonation structure) or overlapping parts such as *ACCENT* with regard to syllable, word or sentence association.

5.3.2 Paradigmatic (taxonomic) relations

1. In syntax, paradigmatic relations hold over sets of contrasting items which enter the same syntagmatic relation, such as *pronoun*, *noun phrase*.
2. In phonology and prosody, paradigmatic relations hold over sets of contrasting items which enter the same syntagmatic relation, such as contrasting consonants and consonant clusters in *ONSET* position, or contrasting L* or H* accents in *ACCENT* position, or rising LH and falling HL pitch contours in *NUCLEUS* position.

5.4 Rendering interpretation: modality semantics

But this is not all. There is one other important complex of relations, namely the relation of linguistic structures to produceable and perceivable motor-sensory modalities, which is variously referred to in different linguistic frameworks as *expression*, *realization*, *manifestation*, *interpretation* (e.g. *phonetic interpretation*). Generally this relation pertains to the acoustic modality, but if conversational gestures are included, as is becoming more and more common in many branches of linguistics and the human language technologies (Trippel et al. 2003), then the visual modality also has to be included here. A serious theory of writing (which does not yet exist) will also need to include the visual modality into its interpretation model.

Formally speaking, the *modality interpretation* relation is based on a *modality model*, which consists of the following two components (Carson-Berndsen 1998, Gibbon 1992, 2006):

1. a *domain* of phonetic categories and relations, including the *time type*—categorial, relational or absolute time types—and the *time map* between types as well as *precedence/overlap relations* which constitute type structure, and
2. a *function* which maps linguistic, phonological and prosodic units into the domain.

Nor is this all: different syntagmatic relations and the paradigmatic choices associated with them have to be interpretable in terms of their meaning or function in discourse, as well as in terms of the modalities of expression. In generative linguistics, as well as in formal logic, this dimension is known as *semantic interpretation* and is associated with a *semantic model*. For simplicity, but also because the traditional terms overlap considerably, I use "semantic" here to cover both conventional *semantics* and conventional *pragmatics*: consider the meanings of deictic categories, of speech act verbs, of the scope of conjunction or negation as marked by intonation and accentuation, in which semantics (briefly, concerned with *truth*) and pragmatics (briefly, concerned with *use*), as conventionally defined, overlap.

Formally speaking, the *semantic-pragmatic interpretation* relation is based on a *semantic-pragmatic model*, which consists of the following two components:

1. a *domain* of semantic and pragmatic categories and relations,
2. a *function* which maps linguistic units into the domain.

Formally, therefore, modality interpretation and semantic-pragmatic interpretation are very similar; the domains differ (though the modality domain is clearly one which we can also speak about, and therefore, strictly speaking, is also a subdomain of the semantic domain). Indeed, the fundamental sound-meaning relation is neatly explicated as *the pair of modality and semantic interpretations*.

6. The Time Type model of speech modality rendering

6.1 Grounded formal categories for transcription and annotation

The secondary formal categories are primarily concerned with the instantiation of the primary categories in terms of types of representation and interfaces (mapping functions) between them. Much could be written about this area, and the transcription and signal annotation systems which implement it, but attention will be restricted here to those relations which are essentially concerned with speech resource creation. More work is needed on these relations, so initially some of the basic concepts will simply be listed, before being explicated in more detail below:

- Transcription is the assignment of a segment of a speech event to a symbol (orthography, IPA, iconic).
- Annotation is a pair of a transcription label and a time-stamp (in the simplest, single-track case; cf. also multi-track annotation, hierarchical (tree-bank) annotation, and multi-stream annotation as with audio-visual recordings using appropriate alignment software).

Underlying these ideas is the insight that communication takes place in simultaneous channels and that these channels may be in different modalities:

1. *Vocal-acoustic*: speech
2. *Vocal-visual*: lip-reading
3. *Gesture-visual*: gesture
4. *Gesture-acoustic*: clapping, snapping, stamping

Functionally determined submodalities of the same modality also need to be defined, for instance in the vocal-acoustic modality the following:

1. *Locution*
2. *Prosody*
3. *Paralinguistics*

6.2 Time types as determinants of levels of analysis

The following time-oriented ontology outline for formal speech-specific categories and relations is modified slightly from Gibbon (2006), and is based on Gibbon (1992). First, three Time Types are needed as a basis for prosodic event alignment in the present analysis:

Absolute Time relates to signal-oriented phonetics, that is, to time points and intervals determined by calibrated physical measurement. For example, standard digital signal sampling techniques generate Absolute Time structures. In the Absolute Time domain, the quantitatively measured lengths of phones, syllables, etc., are important. Impressionistic phonetic judgements on length and tempo, as practised in phonostylistics and discourse analysis, may be seen as coarse-grained and uncalibrated quantitative measures.

Relative Time relates to 'interpretative phonetics', i.e. phonology and prosody, and defines intervals and other relations between points in time with no explicit assignment to Absolute Time. Relative Time characterizes the prosodic phonologies; the key relations are sequence, overlap and hierarchy, which are interpretable in terms of the Absolute Time domain.

Categorial Time relates to underlying lexical and grammatical levels, in particular to categories linked by algebraic operations such as concatenation. In the Categorial Time domain, there is only a notion of temporally uninterpreted structure; to include a notion of time, phonetic interpretations into the Relative Time and Absolute Time domains are required.

The three-level distinction between Time Types is supported by work in formal linguistic theory, in particular in Event Phonology, in Time Type Theory and in the Time Map Phonology approach to alignment theory (for further references, cf. Gibbon 2006).

6.3 Speech mining procedures

Key data-mining procedures in the exploitation of temporal annotations for speech resource creation can now be formulated in terms of the Time Types:

1. Analog-digital transformation in the signal sampling process, between two subdomains of Absolute Time.
2. Annotation as a mapping of the quasi-continuous digital domain of speech signals into the discrete Absolute Time domain of annotation intervals.
3. Induction of temporal structures from the discrete Absolute Time subdomain of annotation intervals to linear and hierarchical Relative Time structures.
4. Mapping of Relative Time structures to Categorial Time grammatical and discourse patterns.

6.4 Event alignment: streams, tracks, tiers

Each of the three Time Types is associated with its own specific range of sequential and partially aligned parallel structures at different theoretical and heuristic levels of description. The relevant levels for the present study are distinguished as follows:

1. a set of parallel signal *streams* (time functions describing continuous or discrete sampled speech signals),
2. partially aligned with a set of parallel annotation *tracks* (time functions describing discrete, categorial sequences of events, as in a speech editor, for example, with sampled speech signal and parallel annotation tracks),
3. which are often derived from specific phonological *tiers* (linguistic constructs defining partially aligned trajectories through a feature space in Relative Time, as in autosegmental and other prosodic phonologies).

The "stream-track-tier" terminology is intended to keep apart clean ontological levels which are often indiscriminately labeled with terms like 'tier', 'track', 'level', 'layer', 'stratum', 'stream'.

The following more detailed terminological overview is based largely on the related models of Event Phonology, Time-Map Phonology, and Annotation Graph theory (cf. Gibbon 2006 for further details, including references).

6.5 Speech events and their representation

Speech event: A pair of an Absolute Time or Relative Time interval and a trajectory or pattern of values in some phonetic dimension, parameter or feature.

Examples:

- an interval of 120 ms and a phone segment, as a static or a dynamic time function in Absolute Time on an annotation track;
- an interval of 10 ms and a pitch value in an Absolute Time F0 stream;
- an interval of 0.0208333 ms (corresponding to 48 kHz sampling rate) paired with an amplitude value in an Absolute Time signal stream;
- a pair of a phonological segment and its phonemic or feature-based properties in Relative Time.
- a signal annotation <<x_{min}, x_{max}>, transcription>, where x_{min} and x_{max} range over points, *transcription* ranges over textual symbols, <x_{min}, x_{max}> ranges over intervals ($x_{max} – x_{min}$ ranges over durations); cf. the following (Brazilian Portuguese) syllable annotation extracted from a Praat annotation file:
  xmin = 0.48473069812858305
  xmax = 0.6301876830002222
  text = "koN"

Transcription: The name of the pattern of an event.

Annotation: The name of an Absolute Time event, consisting of a set of pairs of transcriptions and either interval time-stamp pairs or point time-stamps.

Point: The undefined primitive for defining intervals as a pair of points (whether abstract points as in Relative Time, or clock time points as in Absolute Time), ignoring for present purposes the traditional discussion on whether points or intervals are primitives.

Time-stamp: The name of a point or a pair of points (an interval) in Absolute Time, i.e. a calibrated quantitative designation of a relative to some pre-defined initial point (the term 'tick' is used in digital music and virtual machine technology).
Examples:

- Mon Mar 28 13:32:30 BST 2005
- 321.5 ms

Absolute interval: The difference between two time points.
Examples:

- the subtraction of two time-stamps
- the time elapsed between two metronome beats.

Relative interval: A segment at an abstract phonological level, related to other intervals by relations of precedence and overlap. A relative interval has no absolute duration unless explicitly mapped into an absolute interval.
Example:

- The epenthetic [t] in English [prints] "prince" arises when the end of the nasal event interval of [n] precedes the end of the occlusive event interval of [n].

Time-Map: A function within one Time Type or between Time Types, mapping one temporal representation into another.
Examples:

- speech signal digitization (analogue signal sampling),
- annotation (aligns digital speech signal with event label sequence),
- phonetic interpretation (mapping of lexico-syntactic representation of speech forms into a phonetic representation.

This list is not complete, but is intended to serve as an initial orientation point for formal elements of a speech ontology.

6.6 Substantive categories

To discuss a detailed ontology for the substantive categories of speech would go much too far in the present context. For segmental phonology, fundamental issues are outlined by Aristar (2005) and Kamholz (2005) in the context of E-MELD project discussions of the GOLD ontology.

Aristar details the following questions with respect to the GOLD ontology and his own proposal for segmental phonetic categories (minimally edited here):

- What is missing in the ontology?
- What is present that should not be there?
- How do we handle binary features such as coronal and dorsal, tense and high for consonants? Are these phonetic, or something else? Do we need a separate feature node for these?
- Have we made a proper distinction between phonetics and phonology? E.g. does the notion syllabic/non-syllabic properly belong here? What a syllable can be is reasonably decided by the phonology of a language, except in cases like semivowels (e.g. [w], [j]) which automatically become a vowel if they are syllabic, no matter what the language. So syllabicity is partly phonological, and partly phonetic.
- The ontology defines only membership of classes. For example, the IPA symbol [u] is a member of the classes Vowel, Back, Round, High, Vowel_Symbol. But there are facts about language which cannot be described in terms of class-membership. For example, can a voiceless sound ever be laryngealized? Can a voiced segment be ejective? These are essentially constraints between sounds. What constraints does a phonetic ontology need?
- There are sounds written as clusters in standard transcriptions, e.g. nasal clicks, multiple articulations like the labial-velar [gb]. Phonologically and phonetically, these seem to function as single sounds. But transcriptionally they function as clusters. What is the best way to handle these?

At a metatheoretical level, these issues are addressed in the present contribution; however, detailed solutions are still to be worked out.

One subset of substantive categories comprises the alphabets used for labeling segmental and prosodic categories:

1. *Segmental*: Kamholz (2005) essentially follows the International Phonetic Alphabet (but also points out the need for a notion of gradual or scalar feature values). The IPA is certainly the major candidate for phoneme-sized segmental labeling. It is essential in discussing the IPA for ontological purposes to make the following distinctions:

   1. The IPA as a set of phonetic categories, defined by a set of feature matrices.
   2. The IPA as a set of glyphs (font shapes) representing these categories, for which there are many (mutually largely incompatible) mappings to numerical codes and implementations, such as:
      1. Unicode (output oriented for publishing; problems with manual character input),
      2. Truetype fonts (widespread, highly inconsistent amongst each other),
      3. Metafont tools for LaTeX (easily modifiable),
      4. SAMPA representation in ASCII (the most useful for research-oriented computation).
2. *Prosodic*: no proposals have been made so far for a prosodic ontology, and there are several proposals, each emphasizing different category and relation sets, but evaluations of these with regard to the requirements of speech resources are still not available. The main candidates are:
   1. The IPA prosodic categories (with special glyphs).
   2. Hirst's IntSint relational categories of Hirst (Hirst & Di Cristo 1998), with special glyphs and ASCII representations.
   3. Gibbon's SAMPROSA compendium of symbols (cf. Gibbon et al. 1997, with ASCII representations).
   4. The ToBI symbol set (cf. Silverman et al. 1992, with ASCII representations).
   5. Other representations:
      1. The "tadpole" iconic representation of tonetic language teaching materials.
      2. Numerical representations of stress and pitch heights.

7. Conclusion

This contribution has outlined some of the essential features of an ontology for speech resource administration and search, including formal ontological categories, in some detail, and has pointed out strategies for developing substantive ontological categories.

Many open questions remain, including the following:

1. Content:
    1. Which prosodic category systems—one, more than one, all?—are appropriate for including in a *General Heuristic Ontology for Speech Technology* (let's call it *GHOST*).
    2. Important areas of prosody are not included, such as speech timing, including rhythm (cf. Gibbon 2006).
    3. The area of conversational gesture (Trippel et al. 2003) was mentioned, but not discussed in detail.
    4. A whole area has not been covered by the present discussion, namely performance problems of *disfluency* (cf. Tseng 1999) and *discourse particles* (cf. Fischer 2000).
2. Implementation:
    1. What is the most useful mapping of the GHOST categories and relations discussed in the present contribution into markup languages such as XML?
    2. How can a comprehensive GHOST system be incorporated into current annotation software, and into a semantic web oriented search system such as OLAC?

These questions will, for the moment, be left open for further discussion.

Dafydd Gibbon
Fakultät für Linguistik und Literaturwissenschaft
Universität Bielefeld
Postfach 100 131, D-33501 Bielefeld
Germany
gibbon@uni-bielefeld.de

Linguistic Patterns in Spontaneous Speech, 27-50
2009-1-025-002-000264-1

Analysis of Language Variation Using a Large-Scale Corpus of Spontaneous Speech[*]

Kikuo Maekawa
The National Institute for Japanese Language

Large-scale corpus of spontaneous speech can be a powerful tool for the study of language variation. Moreover, given that the corpus is publicly available, corpus-based analysis could open up the possibility of follow-up analysis in this area of linguistic study. Generally speaking, follow-up study is highly desirable in sciences but so far it has been virtually impossible in the area of socio-linguistics due to the lack of shared corpus. In this paper, I will present some results of the analyses of the *Corpus of Spontaneous Japanese* (CSJ) that we developed in the years 1999-2003.

CSJ is a large, richly annotated corpus of spontaneous speech of present-day Japanese (http://www2.kokken.go.jp/~csj/public/index.html), containing more than 660 hours of speech uttered by more than 1400 speakers. This corpus was designed primarily for statistical machine learning of acoustic- and language-models for automatic spontaneous speech recognition, but it was also designed for the study of language variation.

So far, we have analyzed variations at different levels of language structures including, vowel devoicing, pitch-accent location in adjectives, coalescence of particle succession, moraic nasalization of particles, diffusion of the new potential verb forms, choice of phrase-final boundary pitch movements (BPM), and strength of the prosodic boundary preceding accented particle. In addition to these, analysis of word-form variation was conducted. The last analysis was concerned not only with individual lexical items, but also with the lexicon as a whole.

Key words: spontaneous speech, corpus, variation, intonation, Japanese

1. Introduction

Linguistic variation is a statistical phenomenon. Prediction of the occurrence of a

[*] I am very grateful to my colleagues and former colleagues at the NIJL who participated in the CSJ project and analysis of language variation; Hanae Koiso, Hideki Ogura, Masaya Yamaguchi, Hideaki Kikuchi, Takayuki Kagomiya, Wataru Tsukahara, Kiyoko Yoneyama, Masako Fujimoto, Kenya Nishikawa, Yoko Mabuchi, Yohichi Maki, Kenji Yamazumi, Takehiko Maruyama and Yosuke Igarashi.

given variant is possible only on probabilistic ground even when the context in which the variant is located is given. The natural consequence is that the study of linguistic variation tends to require large amount of data. This is particularly true when one wants to study variants whose occurrences are influenced by many factors encompassing both linguistic and social aspects of the target language.

Accordingly, those linguists who study variation are among the people who most enthusiastically welcome the availability of large-scale corpora and the rise of a new discipline called corpus linguistics. Today, as a matter of fact, substantial parts of introductory textbooks of corpus linguistics are devoted to the study of variation.

While this may be true for written language, the situation is drastically different for spoken language. It is also true with spoken language that the study of variation requires large amounts of data. What is drastically different from the case of written language is that there isn't any corpus of spoken language that could be used for the study of language variation.

This may appear odd given the facts that 50% of the data collected by the *Survey of English Usage* project (known today as the London-Lund Corpus), and 10% of the *British National Corpus* are devoted for spoken language. In the case of the SEU, the total amount of data, 500 thousand words, is too small to conduct complex analyses, and in the case of BNC, the transcription is too broad to get fine information about the phonetic details. In addition, probably the most important drawback is that both corpora do not provide speech sound files (In the case of BNC, we can listen to the speech materials in the British Library, but the materials are not publicly available).

In the field of speech engineering, on the other hand, statistical approach, hence corpus-based approach, has been the main-stream for at least 20 years. Spoken language corpora have been widely used for the purpose of automatic learning of language- and/or acoustic-models for automatic speech recognition (ASR) system. Later on, in the 1990s, corpus-based speech synthesis was also developed. It turned out that it was possible to synthesize naturally-sounding speech just by making optimal concatenation of labeled speech sounds in the corpora.

Corpora compiled for speech processing purposes are, however, of little use for the study of language variation, because the speech material is hardly spontaneous. Typically, the materials in such corpora are spoken version of written texts like newspaper articles or so-called phonemically balanced sentences. These materials were pronounced, typically, by professional narrators to have as small a number of fluctuations as possible.

Accordingly, scholars of language variations had to compile databases, or corpora, of their own each time they started examining a new variable. Needless to say, it is a time-consuming effort. Moreover, those corpora constructed for personal use become rarely available for other researchers.

I didn't hesitate much, in February 1998, when the director general Seiichi Yamamoto of the ATR spoken language translation laboratory (currently professor of Doshisha University) called me and asked if I was willing to be one of two sub-leaders of a new speech processing project in which I was expected to design and compile a large corpus of spontaneous Japanese under the supervision of professor Sadaoki Furui of the Tokyo Institute of Technology. I could almost intuitively understand that it was a good occasion for the discipline of the study of language variation.

Almost a year later, we submitted a research proposal to the former Science and Technology Agency (currently a part of the Ministry of Education), and the proposal was accepted without much ado. This is how the *Spontaneous Speech: Corpus and Processing Technology* project got started in the spring of 1999. This was a five-year (1999-2003) joint project of the National Institute for Japanese Language (NIJL), the National Institute for Information and Communications Technology (aka Communications Research Laboratory till 2001), and the Tokyo Institute of Technology.

Although the goal of the project was to develop a prototype system for the next generation ASR system that could recognize spontaneous speech, there was a clear consensus among the project members that development of large-scale spontaneous speech corpus was the key issue. In the first six months or so of the project, my colleagues and I concentrated our efforts in designing a corpus that could capture as much information as possible about the variability—both physical and linguistic—of spontaneous speech, based upon the belief that an optimal corpus of spontaneous speech designed for ASR system could be an excellent resource for the study of language variation as well.

2. Outline of the Corpus of Spontaneous Japanese

The spontaneous speech developed by the abovementioned project is known as the *Corpus of Spontaneous Japanese*, or CSJ. It was released in June 2004, and more than 300 copies have been purchased by researchers in various research institutions including universities, national laboratories, and companies.

2.1 The size

Table 1 shows the whole size of CSJ with respect to the numbers of words, speakers, talks, and total hour of recorded speech. The number of speakers is smaller than the number of talks because there were many speakers who provided more than one talk.

Table 1: Size of CSJ

N of running words	7,525,125
N of different speakers	1,417
N of talks	3,302
Total hour of speech	662

Table 2 shows the type of talks recorded in the CSJ. Parenthesized numbers of speakers were counted more than twice. As shown in the 'MODE' column of the table most of speech materials are devoted to monologues, but at the same time, they cover wide range of speaking styles ranging from read speech to free conversation. 'Reread speech' is the reading aloud of the transcription of spontaneous speech previously uttered by the same speakers.

Difference of talk types is an important factor of data analyses when we conduct linguistic analysis of language variation. Fig. 1 shows the ratio (%) of word-form variation, i.e. the total number of non-standard variants divided by the number of total occurrence of the word in question multiplied by 100, as a function of the type of talks. There is a clear correlation between the ratio of word-form variation and the expected ranking of speaking style. See also §4.2.

Table 2: Type of talks in CSJ

TYPE OF TALKS	MODE	N FILE	N SPKER	HOUR
Academic Presentation Speech (APS)	Monologue	987	819	274.4
Simulated Public Speaking (SPS)	Monologue	1,715	594	329.9
Public Lectures (PL)	Monologue	19	16	24.1
Interview on APS	Dialogue	10	(10)	2.1
Interview on SPS	Dialogue	16	(16)	3.4
Task-oriented dialogue	Dialogue	16	(16)	3.1
Free dialogue	Dialogue	16	(16)	3.6
Reread speech	Monologue	16	(16)	5.5
Read speech	Monologue	507	(248)	15.5

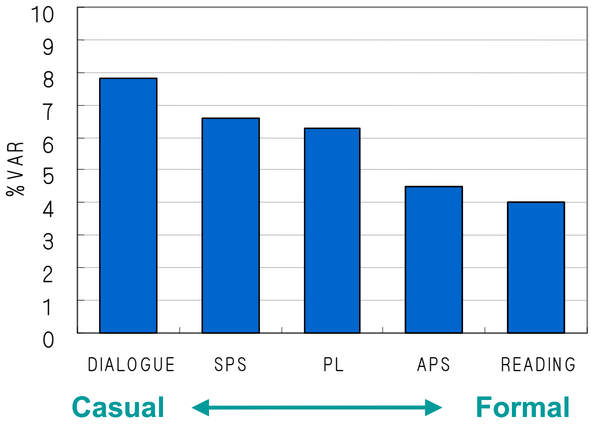


Figure 1: Correlation between the type of talks and the ratio of word-form variation

2.2 Speakers

Another important social factor of language variation is the age of speakers. Fig. 2 shows the distribution of the CSJ speakers with respect to their birth years that were sectionalized for every decade. There is clear difference between the speakers of APS and those of SPS.

APS speakers are heavily concentrated in their twenties, because most of the APS speakers were graduate students. SPS speakers, on the other hand, shows less skewed distribution compared to APS. This is because SPS speakers were recruited so that their age and sex show distribution as uniform as possible. Note, in passing, that the number of speakers shown in Fig. 2 is the total (cumulative) number of speakers (i.e. one and the same speaker may be counted more than twice when he/she gave more than two talks). If we count the number of different speakers as in Fig. 3, distribution of the SPS speakers is not as uniform as in Fig. 2, but it is still much more uniform compared to the distribution of the different APS speakers.

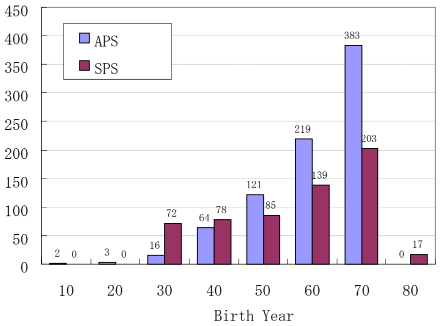


Figure 2: Number of total speakers as a function of their birth year

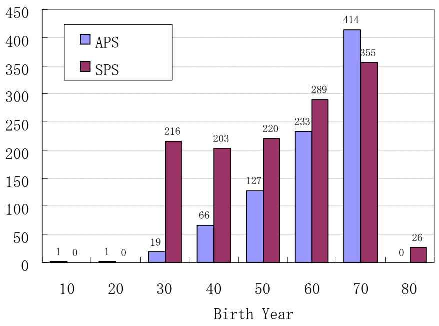


Figure 3: Number of different speakers as a function of their birth year

Table 3: Distribution of the cumulative number of speakers

SEX	APS	SPS	PL	READ	REREAD	INTERVIEW	Total
FEMALE	173	910	9	252	8	29	1,381
MALE	814	805	10	255	8	29	1,921
Total	987	1715	19	507	16	58	3,302

Table 4: Distribution of the number of different speakers

SEX	APS	SPS	PL	READ	REREAD & INTERVIEW	Total
FEMALE	138	331	6	(122)	(8)	470
MALE	681	263	10	(124)	(8)	947
Total	819	594	16	(246)	(16)	1,417

Tables 3 and 4 show the distribution of speakers' sex as a function of talk types. Parenthesized numbers in the latter table indicate that the speakers were already counted as the speakers in other types of talks. All speakers of read speech, reread speech, and interview speech were counted more than twice. Note that the speakers of reread speech and interviewees of interview speech belong to the same group of speakers.

2.3 Annotations

As shown in Fig. 4, CSJ consists of several layers differing in the richness of annotation. This multi-layer structure was introduced into the corpus to satisfy incompatible needs of the corpus: richness of annotation and the size of corpus.

The Core of the CSJ includes half a million words and is the part of the corpus to which the cost of annotation was concentrated, the most crucial difference being the application of segmental and intonation labeling. Moreover, there are two more layers inside the Core that differ in the richness of annotation. The richest part in the Core are annotated with respect to segmental label, intonation label, dependency structure label, impression rating, and, topic structure label, in addition to the following annotations that are provided for all speech files, i.e., two-way transcription, two-way POS information, clause boundary information, impression rating, and, information about the speaker and the talk per se.

The rest of this section is devoted to a brief introduction to some of the CSJ annotations that will be referred to in the sections on language variation.

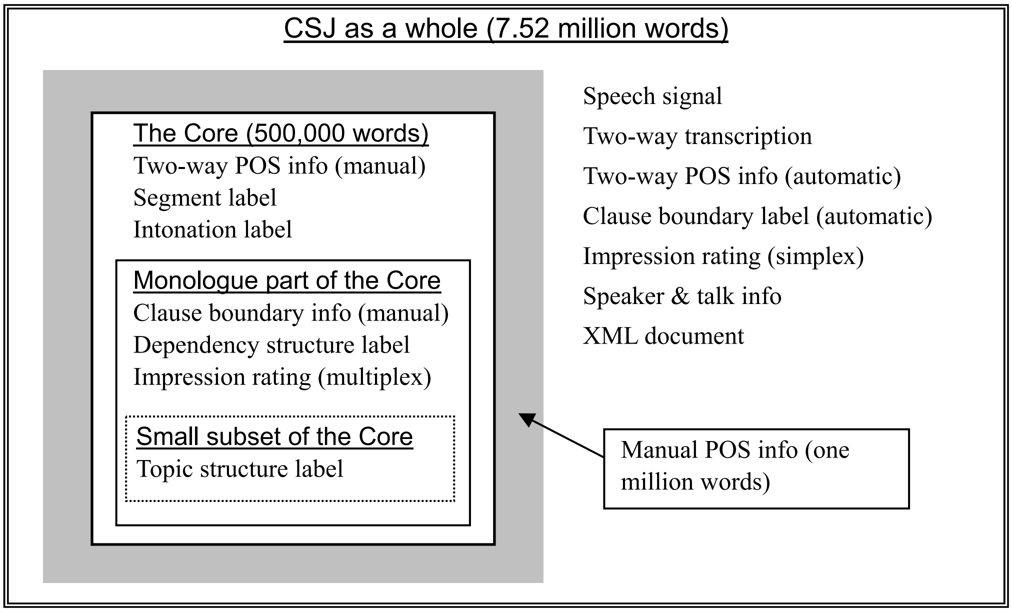

CSJ as a whole (7.52 million words)
The Core (500,000 words)
Two-way POS info (manual)
Segment label
Intonation label
Monologue part of the Core
Clause boundary info (manual)
Dependency structure label
Impression rating (multiplex)
Small subset of the Core
Topic structure label
Speech signal
Two-way transcription
Two-way POS info (automatic)
Clause boundary label (automatic)
Impression rating (simplex)
Speaker & talk info
XML document
Manual POS info (one million words)

Figure 4: Layered structure of annotation in the CSJ

2.3.1 Two-way transcription

Transcription of Japanese speech requires special treatment, because the language's orthography has a very high degree of freedom. There are, almost always, more than two ways of writing down the same linguistic message. For example, there are at least four common ways of writing a compound verb /hanasi-au/ ("discuss"), viz., 話し合う, 話合う, 話しあう, and はなしあう. This flexibility in orthography could be a strong obstacle for the corpus search, needless to say.

CSJ overcame this problem by providing two independent transcriptions called orthographic and phonetic transcriptions. In the orthographic transcription, utterances were transcribed by Kanji (Chinese logographs) and Kana (Japanese syllabary) characters following the rules of orthography that we established for CSJ. The new orthography was designed so that there is no degree of freedom.

Phonetic transcription, on the other hand, uses Kana exclusively to transcribe the phonetic details of the utterances as exactly as possible within the limitation of a syllabary.

The combination of orthographic and phonetic transcriptions provides powerful tool for the search of word-form variations. Fig. 5 shows schematically how the transcriptions could be used for a search. In this case, word-form variation of adverb morpheme {yahari} ("after all") was examined. The left hand string of Kanji and Kana (矢張り) is the orthographic transcription of (the dictionary form of) the morpheme, while the right hand strings of /yahari/, /yaQpari/, /yaQpa/, /yapa/, and /yaQpasi/ are the variants of the adverb as they are represented in the phonetic transcription, and they are written exclusively in Kana. By making comparison of the two transcriptions in this way, it is possible to extract useful information about word-form variations.

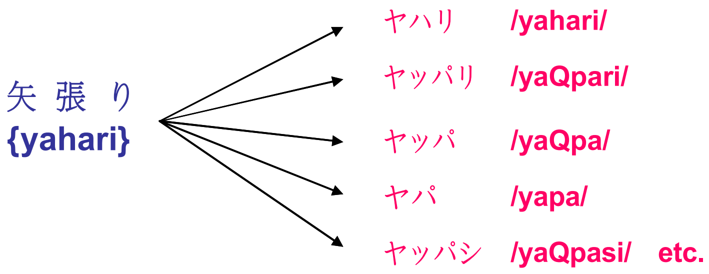


Figure 5: Orthographic and phonetic transcriptions

2.3.2 Two-way POS information

CSJ provides two-way POS information based upon the SUW (short unit word) and LUW (long unit word). The two-way POS analysis is required because Japanese is a highly agglutinative language, and, unlike languages like English or Chinese, the definition of "word" is heavily theory dependent.

Table 5 shows an example of two-way POS analysis. The example utterance is taken from an APS of CSJ and means "Information obtained by means of binaural perception includes power spectrum information and binaural phase difference." The second column of the table represents Romanized phonetic transcription corresponding to each of the SUW shown in the third and fourth columns. "Dictionary form" in the third and fifth columns means the standardized orthographic representation of LEMMA corresponding to each SUW or LUW. Because dictionary form is concerned with LEMMA rather than word form, it represents the infinitive or ending form of a conjugation words (verbs and an auxiliary verb in the Table 5).

If we compare the results of SUW analysis shown in the third and fourth columns and that of LUW analysis shown in the fifth and sixth, it turns out that successive occurrence of SUW nouns is interpreted as a single LUW noun. This happens three times in the example in “binaural perception”, “power spectrum information”, and “binaural phase difference”. As long as these examples are concerned, the LUW corresponds to a compound and the SUWs are its components. In the same table, however, we see more complex example, where a single LUW case particle /niyoQte/ (“by means of”) corresponds to 3 elements in the SUW consisting of a case particle (/ni/) followed by a verb (/yoQ/), which is, in turn, followed by another case particle (/te/).

Table 5: Example of two-way POS analysis

Gloss	Phonetic Transcription	Short Unit Word (SUW)		Long Unit Word (LUW)	
		Dictionary Form	POS	Dictionary Form	POS
binaural	ryoHzi	両耳	N	両耳受聴	N
perception	zyuchoH	受聴	N		
PLACE	ni	に	Ptcl-Case	によって	Ptcl-Case
be based upon	yoQ	拠る	V		
CONJUNCTION	te	て	Ptcl-Conj		
obtain	eru	得る	V	得る	V
information	zyoHhoH	情報	N	情報	N
PLACE	ni	に	Ptcl-Case	に	Ptcl-Case
TOPIC	wa	は	Ptcl-Topic	は	Ptcl-Topic
power	pawaH	パワー	N	パワースペ クトル情報	N
spectre	supekutoru	スペクトル	N		
information	zyoHhoH	情報	N		
and	to	と	Ptcl-Case	と	Ptcl-Case
binaural	ryoHzi	両耳	N	両耳間位相差	N
between	kaN	間	Suffix		
phase	isoH	位相	N		
difference	sa	差	N		
NOMINATIVE	ga	が	Ptcl-Case	が	Ptcl-Case
exist	ari	有る	V	有る	V
POLITE	masu	ます	Auxiliary	ます	Auxiliary

2.3.3 Intonation labels

X_JToBI (Maekawa et al. 2002, NIJL 2006), an extended version of the J_ToBI intonation labeling scheme (Venditti 1997, 2005) was developed for the labeling of spontaneous speech. The new scheme is capable of expressing the details of the prosodic characteristics of spontaneous speech including, among other thing, boundary pitch movements (i.e., the characteristic movement of pitch at the phrase boundary) and their variation.

2.3.4 Impression rating

Impression rating is the subjective rating of various impressions that listeners perceive from spontaneous talks. CSJ provides two kinds of impression rating data.

Each monologue talk (APS, SPS, and PL) was evaluated at the time of recording by a rater about speaking rate, speaking style, spontaneity of the talk, etc. Although different talks were evaluated by different raters (i.e., the rater was not uniform for all talks), the resulting impression rating data turned out to be very useful for the analysis of language variations. See §§3.3 and 3.4 below. This data is called simplex impression rating data.

There is another type of impression rating data called multiplex data. Monologue talks in the Core were evaluated by 10 raters using psychological scales developed specially for CSJ (Yamazumi et al. 2005).

2.3.5 Clause boundary label

It is often very difficult to label the sentence boundary of spontaneous speech. It is in most cases possible, however, to label the boundary of syntactic *clauses*, i.e., the syntactic unit consisting of predicates and their complements. All transcription files of the CSJ were automatically classified with respect to the morphological characteristics of the predicates using the result of POS analysis (mostly SUW information). For all talks included in the Core, the results of the automatic classification were checked and, if necessary, corrected by human labelers.

3. Analysis of some selected language variations

In this section, results of some pilot studies about language variation will be presented. Examples are selected so that they cover as wide range of the levels of linguistic structure as possible. All examples use CSJ as the source, needless to say. Note, however, that some of the studies reported below were conducted while the compilation of the CSJ was underway in order to evaluate the usefulness of the corpus (Maekawa 2004, Maekawa et al. 2003 for example). As the result, some of the results reported below did not use the current version of CSJ as its source, but I believe that there would not be much difference even if we conduct reanalysis using the publicly available version of the CSJ.

3.1 Vowel devoicing

It is well known that in Japanese close vowels, —/i/ and /u/—, are devoiced when they are preceded and followed by voiceless consonants. This is the typical environment of vowel devoicing in Japanese, but this is by no means the only environment of vowel devoicing, as aptly summarized in Vance (1987). Maekawa & Kikuchi (2005) analyzed a subset of CSJ-Core containing 427,973 vowels and reported several interesting findings.

Table 6 compares the devoicing rate of five Japanese vowels under the four phonological environments defined in terms of the voicing of adjacent consonants. It shows that close vowels are not completely devoiced even under the typical environment of devoicing (i.e., C1 and C2 are both Co); it also shows that there isn't any environment where devoicing is completely avoided.

Table 6: Rate of vowel devoicing as a function of the voicing of adjacent consonants
C1=Preceding consonant, C2=Following consonant, Co=Voiceless consonant, Cv=Voiced consonant.

VOWEL	C1	C2	VOICED	DEVOICED	%DEVOICED
a	Co	Co	12,214	262	2.10
	Co	Cv	18,570	92	0.49
	Cv	Co	24,943	481	1.89
	Cv	Cv	19,867	29	0.15
e	Co	Co	5,550	190	3.31
	Co	Cv	10,890	116	1.05
	Cv	Co	11,552	323	2.72
	Cv	Cv	11,388	29	0.25
i	Co	Co	1,475	12,124	89.15
	Co	Cv	10,556	2,219	17.37
	Cv	Co	9,200	126	1.35
	Cv	Cv	12,072	133	1.09
o	Co	Co	12,247	437	3.45
	Co	Cv	19,752	365	1.81
	Cv	Co	14,650	13	0.09
	Cv	Cv	16,802	14	0.08

VOWEL	C1	C2	VOICED	DEVOICED	%DEVOICED
u	Co	Co	1,732	9,267	84.25
	Co	Cv	11,851	3,133	20.91
	Cv	Co	5,562	127	2.23
	Cv	Cv	7,748	61	0.78

It has been pointed out by phoneticians that close vowels tend not to be devoiced in the environment where more than two morae (syllables) could be sequentially devoiced. Words like /susi/ (“sushi”), /kucusita/ (“sox”), /kikuci/ (Japanese surname) and /fukusiki/ (“duplex”) contain environments of sequential devoicing.

This tendency has been acknowledged by many phoneticians, but the mechanism of the avoidance, i.e., the mechanism that determines which vowel is to be devoiced and which is not, was not clearly recognized. Analysis of spontaneous speech revealed interesting tendency with respect to sequential devoicing.

Fig. 6 shows the devoicing rate of two adjacent close vowels in the environment of sequential devoicing as a function of the combination of the manners of articulation of the mora-initial consonants. The notation like ‘F/A’ means that the consonant of the first close vowel (‘V1’) is fricative and that of the second close vowel (‘V2’) is affricate. There is a clear trading relationship between the devoicing rates of V1 and V2, with the sole exception of ‘S/S’.

Speakers tended to avoid devoicing of V1 especially in the environment of ‘F/F’, ‘S/S’, and ‘A/F’. In these environments, devoicing of V1 gives rise to two consecutive fricatives (including the last half of affricates) or consecutive stops. The consonant sequences like [kk] (as in /kikuci/), [kts] (as in the first half of /kucusita/, where /c/ is affricate) are phonetically realized as the consecutive occurrence of two sound spikes on the time dimension, and is often difficult to be perceived. Similarly, fricative sequences like [sʃ] (as in /susi/) or [tsʃ] (as in /kucusita/) can be difficult to be perceived. On the other hand, environments like ‘F/A’ and ‘F/S’ are easy to be perceived even when V1 is devoiced, because the consonant sequences are clearly punctuated by the presence of stops (including the first half of affricates).

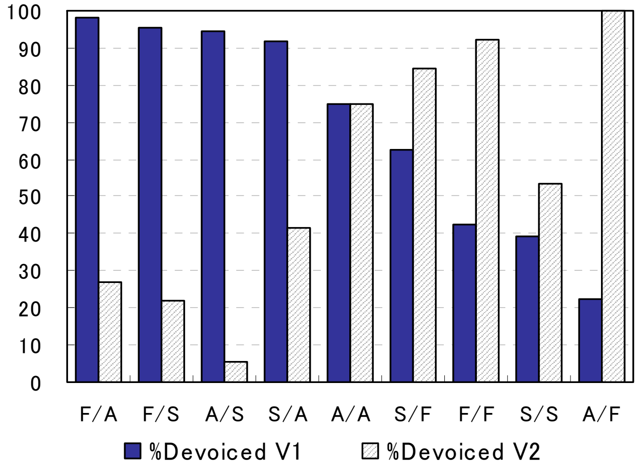


Figure 6: Devoicing rate of adjacent close vowels in the environment of sequential devoicing. V1: First vowel, V2: Second Vowel, F: Fricative, A: Affricate, S: Stop.

3.2 Phrasing of accented particles

Most dialects of Japanese have lexically specified pitch accent. Tokyo Japanese that is the main target of the CSJ is no exception. Lexical items in these dialects are specified for the presence and absence of lexical accent, and in the case of presence, the location of pitch accent as well.

When lexical items are joined together to form an utterance, however, not all accents are realized as they are specified in the lexicon. There are rules of compound word accentuation, and there are also rules of accentual phrasing.

Accentual phrase (AP) is the most important unit of Japanese prosody in which at most one lexical accent can be specified (i.e., an AP is either accented or unaccented). Therefore, rules, or principles, of accentual phrasing has to determine which accent is to be deleted when there are more than two accents in the string of lexical items that are to be integrated into a single AP.

Most of the existing literatures on AP in Tokyo Japanese says that accent in the accented particles will be lost when they follow accented nouns, or verb sometimes, to form an AP. For example, particle /ma*de/ (“to”), where asterisk is used to denote lexical accent, will become unaccented when it follows accented nouns like /kyo*Hto+made/ (“to Kyoto”) or /yo*ru+made/ (“to the night”), while it retains its accent when it follows unaccented nouns like /yokohama+ma*de/ (“to Yokohama”) or /yuHgata+ma*de/ (“to the evening”).

This description is widely acknowledged. But astute observers of spontaneous Japanese are aware that this is not always the case. Fig. 7 is taken from Maekawa & Igarashi (2006) that examined the behavior of two-mora accented particles that formed an AP with the immediately preceding accented lexical items in the CSJ.

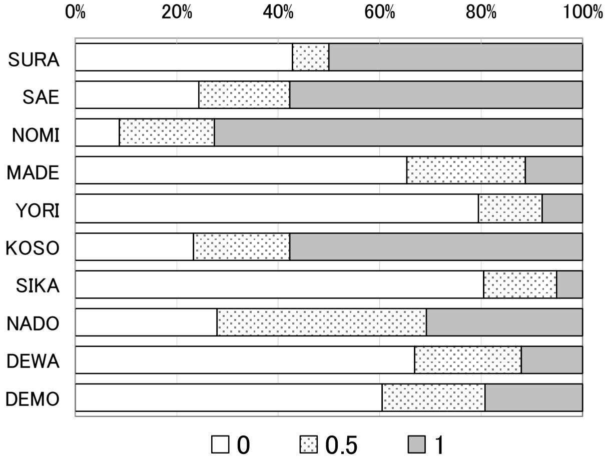


Figure 7: Prosodic independence of two-mora accented particle
0: particle accent is deleted, 1: accent not deleted,
0.5: two labelers did not agree with respect to particle accent.

In this figure, accentedness of 10 two-mora accented particles was compared. The shaded bar (shown as "1" in the legend) represents the percentage of cases where particle accent was not deleted, and, the dotted bar represents the case where two raters gave different judgment about the accentedness of particle. The open bar, accordingly, represents the cases where accent in the particles were deleted. This figure suggests strongly that the rule about AP formation of accented particles is virtually an optional rule, or there might be many hitherto unknown factors that prevent the rule from being applied.

Maekawa & Igarashi (2006) examined the effects of various linguistic and extra-linguistic factors on the phrasing and concluded that the most influential factor was the semantic property of particles. Particles whose semantic function is emphasis and/or limitation tend to constitute an AP of their own.

3.3 Word coalescence

Under some circumstances, function words like particles and auxiliary verbs can be merged with their adjacent words. This phenomenon is called word coalescence. Among the most frequent word coalescences of Tokyo Japanese, coalescence of /de/ and /wa/ into /zya/ was analyzed. This coalescence can be found in two word sequences that are completely distinct from a linguistic point of view; case particle /de/ followed by topic particle /wa/ on the one hand, and, auxiliary verb /da/ in its adverbal form (i.e., /de/) followed by the topic particle, on the other.

Fig. 8 is the result of a pilot decision-tree analysis of the coalescence that I recently conducted. As shown in the top box, the overall coalescence rate is 22.0%. But we can predict the occurrence of coalescence more accurately if we know the POS of /de/; the rate becomes 1.7% and 42.6% when /de/ is particle and auxiliary verb respectively. In the save vein, but to a much lesser extent, factors like type of talks (APS or SPS), speaking style (formality), and spontaneity (spontaneous or prepared) could be useful for the prediction of coalescence. Note the last two factors mentioned above are part of simplex impression rating data. Note also that all these factors, both linguistic and extra-linguistic are provided in the CSJ.

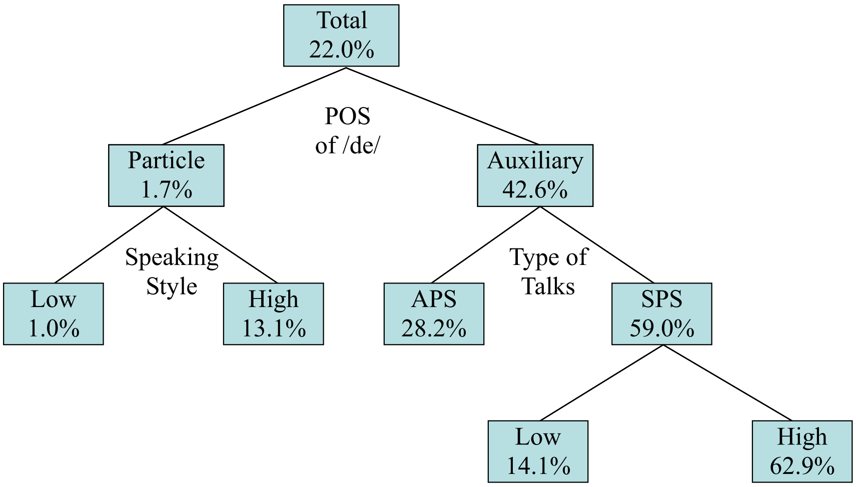


Figure 8: Decision-tree of word coalescence /de/+/wa/ => /zya/
Digits show the rate of coalescence.

3.4 Boundary pitch movements

Boundary pitch movement (BPM) is those characteristic intonations that mark the end of accentual phrases. This is one of the areas of Japanese intonation study that is very interesting but underdeveloped.

In the X-JToBI labeling, the normal falling tune was marked by the label 'L%' and all BPM were labeled as one of the followings: 'L%H%' (rise), 'L%LH%' (another type of rise, called "insisting rise"), 'L%HL%' (rising-falling tune), 'L%HLH%' (rising-falling-rising tune). In addition to these basic categories, variations of intonation were represented by additional labels like 'FR' (standing for "floating rise", a variant of L%H% and L%HL%) and 'PNLP' ("penult non-lexical prominence", a temporal variant of L%HL%).

Fig. 9 shows the occurrence rates of L%H% and L%HL% as a function of the impression rating of speaking style and spontaneity (Maekawa et al. 2003). It is interesting to see that the behaviors of the two BPM are complementary. The rate of rising tune (L%H%) correlates positively and negatively with speaking style and spontaneity respectively, while the rate of rising-falling tune (L%HL%) correlates negatively and positively with speaking style and spontaneity. Note that higher number in speaking style means that the speaking style is more formal.

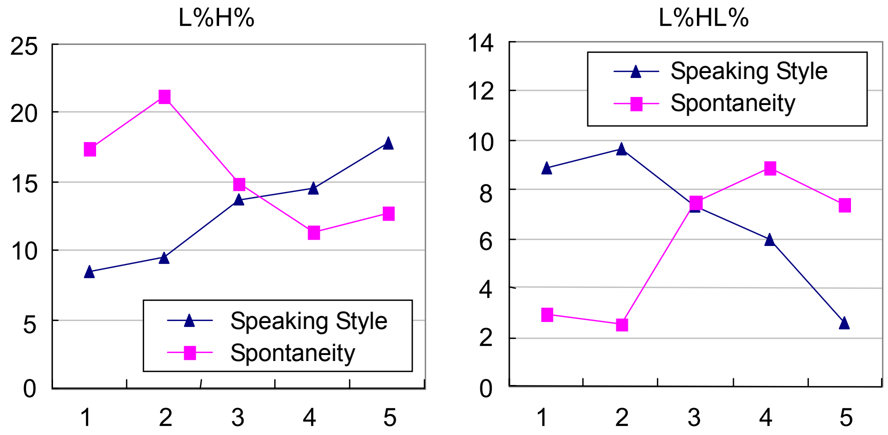


Figure 9: Relationship between the occurrence rates of BPMs [%] and the impression rating of speaking style and spontaneity (abscissa).

3.5 Potential verb (introspection and behavior)

Variation of potential verbs is one of the most well-known variations in the verb-morphology of the present-day Japanese. Traditionally, potential forms of vowel-ending verbs like {miru} ('see'), and {taberu} ('eat') are derived by inserting a potential suffix {rare} between their roots and suffix (i.e., /ru/), the resulting forms being

/mi-rare-ru/ and /tabe-rare-ru/. During the past hundred years or so, however, new potential suffix /re/ has been emerging steadily. This is per se an interesting morphological variation. But the analysis of potential verbs provides very interesting finding about the survey methodology in the study of variation (Maekawa 2005b).

Fig. 10 is the result of questionnaire survey about the potential form of {kuru} ('come') done by Japanese Government's Agency of Cultural Affairs in 2001. In this survey, the subjects were shown the list of traditional /ko-rare-ru/ and innovative /ko-re-ru/ (both mean 'able to come'), and asked which one they used. In this figure, the innovative form overtook the traditional form in the group of subject born in the years 1971-80.

On the other hand, Fig. 11 is the result obtained by analyzing CSJ. In this figure the traditional form was overtaken by the innovative form as early as in the group of subjects born in 1940-49. So, there is at least about 30 year difference between the two surveys with respect to the timing of the innovative form's take-over.

The most straightforward interpretation of this discrepancy would be that most subjects of the questionnaire survey were influenced by their norm of writing, probably without knowing it. Use of innovative forms in writings is still exceptional even among the subjects who use innovative forms constantly in their speech. Needless to say, the data in CSJ is the 'real' recording of the subjects' speech behavior without being biased by speakers' incorrect introspection on their own speech behavior. Data of CSJ can be used to check the validity of questionnaire survey in this way.

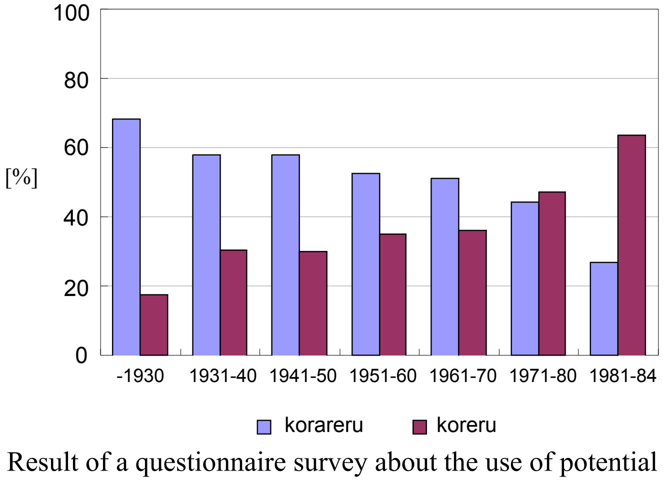


Figure 10: Result of a questionnaire survey about the use of potential form of {kuru} as a function of speakers' birth year.

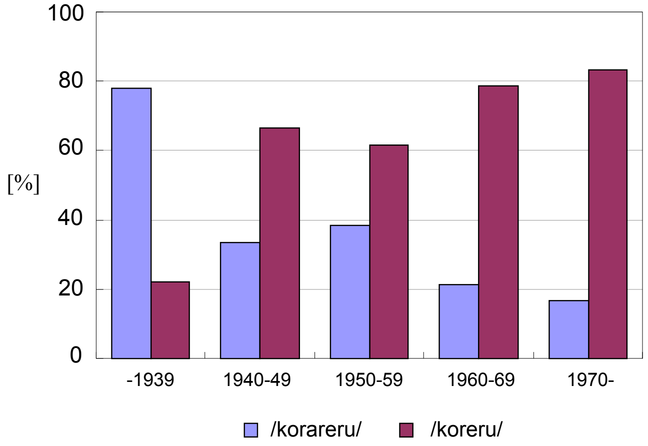


Figure 11: CSJ data about the use of potential form of {kuru} as a function of speakers' birth year.

4. Analysis of word-form variation

So far, we have seen results of pilot surveys about language variations in the levels of phonology, morphology, and prosody. In the rest of this paper, I will present the result of ongoing study concerning the variation of word-forms. This is not the analysis of particular lexical items, but the overall survey of the lexicon as a whole (Maekawa 2005a, 2005b).

4.1 Data

In this study, two types of data about word-from variation were extracted from the CSJ: they are tentatively called phonetic and morphological variations.

Phonetic variation, or P-variation, was recorded in the phonetic transcription by using the tag (W) as in the following examples; (W kokoH; kokoro), (W deHtabeHsu; deHtaHbeHsu), and (W kakemaHru; kakemawaru). In these examples, observed word-forms were recorded as the first element of the tag. The second element of the tag, separated from the first one by a semicolon, is the 'standard' word-from.

The first example deals with the case where consonant /r/ is dropped and replaced by a long vowel (/H/ represents the second mora of a long vowel). The second example is concerned with shortening of lexically specified long vowel. And, the last one is concerned with the drop of /w/ and replacement by a long vowel. All these examples are concerned with articulatory weakening.

The examples shown above occurred very frequently in spontaneous Japanese, but the tag is also applied for sporadic variations. For example, among the 8240 occurrences of lexeme {niQpon}, the country name of Japan, /nihoN/ and /niQpoN/ occurred 7977 and 195 times respectively. In addition to these two, there were other sporadic variants like /nioN/ (39 times), /nihoHN/ (16), /zihoN/ (2), /ioN/ (2) etc.

Morphological variation, or M-variation, is the word-form variation that is not labeled by the tag (W). For example, none of the two variants of the country name of Japan, /nihoN/ and /niQpoN/ are marked by (W). Similarly, variants of the verb meaning to 'say', —/iu/ and /yuH/—, and variants of the first-person singular pronoun, — /watashi/ and /atashi/— are not marked by (W) altogether. The tag (W) is not applied to these variants for two reasons. For one, it is practically impossible to determine which variant is the 'standard' one. For another, some of these variants are not phonetically motivated hence inappropriate to be marked by (W).

Put differently, it was our principle to apply the (W) tag to the variations that are either sporadic or caused by articulatory weakening, or both. On the other hand, the tag is not applied to the cases where most native speakers are aware of the existence of the variation. In fact, the examples of M-variants shown above are usually found among the direction words in ordinary Japanese dictionaries. This is the direct consequence of speakers' awareness about the variation and variants.

Due to the limitation of pages, detailed explanation about how M-variations were extracted has to be omitted with the exception of the following two important points. First, when we talk about 'word-form' rather than 'word', every conjugation form of a conjugational word will be counted as different word-forms. Ending-form, adnominal-form, hypothetical form etc. of a verb, for example, will be recognized as separate word-forms. Second, we hypothesized that every word-form has only one 'standard' form, which is called dictionary form or DF. As will be discussed in §4.5, this is clearly too strong a hypothesis, but this is required for the automatic extraction of M-variations. The process of M-variation extraction is described in Maekawa (2005a, b).

The data that will be analyzed below contains 302,019 M-variations. Because there were 130,951 P-variations in the corpus, the total number of variations was 432,970. Needless to say, these numbers represent the total (or 'running') number of word-forms. The number of different word-forms was 11,379 including both P- and M-variations.

4.2 Correlation with the talk types

There is a correlation between the total occurrence rate encompassing P- and M-variations and talk types. We have already seen this in Fig. 1 earlier. The general tendency is that the variation rate becomes higher in talks with lower formality and less spontaneity. It is important to note, however, that even in the least spontaneous talk type of ‘READING’, about 4% of words showed word-form variation. This fact suggests strongly that the presence of word-form variation is virtually inseparable from our speech behavior.

4.3 Word-forms with high frequency of variation

Table 7 lists 20 word-forms that showed the highest frequency of non-DF variants. The fourth column is the total frequency of the word-forms in question. The fifth column is the frequency of variants other than the ‘standard’ form (DF). The sixth column is the ratio of fifth column over the fourth. And, the last column is the number of speakers who uttered the word-form at least once.

The most important fact concerning this table is that the sum of the frequency of non-DF variants (the 5th column) reached as many as 325,639 and covers about 75% of the total number of non-DF variants in the corpus.

The rate of variation shown in the sixth column of Table 7 is not necessarily the ratio of a single variant. Rather, it was usually the case that multiple variants were observed for a single lexeme. Table 8 is prepared to examine this problem. The second column is the number of different variants observed more than twice in the current data. As can be seen from the table, some word-forms have more than 50 different variants. The third column of the table shows the coverage by the most frequent variant, i.e., the frequency of the top variant divided by the number shown in the fifth column of Table 7. Similarly, the fourth column shows the cumulative coverage by the top 3 variants. In 14 word-forms out of 20, top 3 variants cover more than 99% of the variants, and, there are only two items whose cumulative coverage does not reach 95%, {yahari} and {sore}. This table shows convincingly that it is not necessary to make a long list of non-DF variant to cover the majority of total variation.

4.4 Word-forms with high rate of non-DF variation

It is important to note that Tables 7 and 8 are concerned with the absolute frequency of non-DF variants, and not with the rate of variation. Consequently, there are word-forms whose variation rate is not so high but listed in Table 7 because its occurrence frequency is quite large. Particle /ni/ and copula /desu/ are good example.

Table 9 shows the top 10 lexemes of the highest occurrence rate of non-DF variants. There are 3 items—{niQpon}, {iu}, and {yoi}—shared by Tables 8 and 9. Note, in passing, word-forms whose frequencies were fewer than 10 were removed from the computation for this table.

Table 7: Twenty word-forms that showed the highest frequency of variations

LEXEME	GLOSS	POS (CF)	N	Freq.	%	N of
				Non-DF	Non-DF	Speaker
{iu}	‘say’	Verb (adnominal form)	132,818	132,332	99.6	1,411
{no}	‘of ’	Adnominal particle	153,521	79,829	52.0	1,326
{keredo}	‘but’	Conjunction particle	47,032	26,534	56.4	1,092
{nani}	‘what’	Pronoun	23,067	17,140	74.3	1,054
{iu}	‘say’	Verb (ending form)	9,155	7,991	87.3	1,031
{Qte}	---	Adverbial particle	50,704	7,834	15.5	956
{niQpoN}	‘Japan’	Noun	8,242	8,045	97.6	849
{kurai}	‘even’	Adverbial particle	8,947	7,758	86.7	951
{ni}	‘at’	Case particle	206,614	7,568	3.7	1,097
{yahari}	‘after all’	Adverb	11,746	7,022	59.8	706
{sore}	‘that’	Pronoun	44,000	6,016	13.7	767
{yoi}	‘good’	Adjective (adnominal form)	5,950	5,177	87.0	934
{yoi}	‘good’	Adjective (ending form)	4,446	4,026	90.6	866
{moH}	‘anymore’	Adverb	18,501	3,669	19.8	674
{desu}	Copula	Aux. verb (ending form)	141,084	3,431	2.4	624
{de}	‘and then’	Conjunction	55,717	3,290	5.9	756
{mina}	‘everyone’	Noun	4,309	2,634	61.1	593
{mono}	‘thing’	Noun	31,794	2,373	7.5	593
{watasi}	‘I’	Pronoun	15,749	2,367	15.1	395
{soH}	‘so’	Adverb	29,698	2,327	7.8	619

Table 8: Coverage of non-DF variants by top variants (Same order of row as in Table 7)

LEXEME (CF)	N of Different Variants	Coverage by the Top Variant (%)	Coverage by the Top 3 Variants (%)	Top 3 Variants (From left to right)
{iu} (adnom.)	31	90.3	99.6	/yuH/, /yu/, /yuu/
{no}	15	52.2	99.7	/N/, /no/, /do/
{keredo}	53	53.2	98.5	/kedo/, /keredo/, /keHdo/
{nani}	25	73.9	97.4	/naN/, /nani/, /naNni/
{iu} (ending)	11	90.3	99.0	/yuH/, /yu/, /yuu/
{Qte}	22	82.6	99.2	/Qte/, /te/, /Qti/
{niQpoN}	6	96.8	99.6	/nihoN/, /niQpoN/, /nion/
{kurai}	6	88.7	99.7	/gurai/, /kurai/, /gura/
{ni}	33	96.3	99.8	/ni/, /N/, /i/
{yahari}	56	49.3	91.9	/yaQpari/, /yahari/, /yaQpa/
{sore}	98	85.8	93.8	/sore/, /soe/, /soi/
{yoi} (adnom.)	5	86.0	99.7	/iH/, /yoi/, /i/
{yoi} (ending)	7	91.1	99.4	/iH/, /yoi/, /i/

LEXEME (CF)	N of Different Variants	Coverage by the Top Variant (%)	Coverage by the Top 3 Variants (%)	Top 3 Variants (From left to right)
{moH}	20	80.1	99.3	/moH/, /mo/, /mu/
{desu} (ending)	60	97.4	99.2	/desu/, /esu/, /su/
{de}	33	91.6	98.9	/de/, /Nde/, /te/
{mina}	6	63.3	99.3	/miNna/, /mina/, /miNHna/
{mono}	25	92.3	99.4	/mono/, /moN/, /moH/
{watasi}	34	83.5	98.0	/watasi/, /atasi/, /tasi/
{soH}	28	92.0	99.0	/soH/, /so/, /soQ/

Table 9: Word-forms of the highest occurrence rates of non-DF variants

LEXEME	POS (CF)	N (including DF)	N of Different Variants	Freq. Non-DF	% Non-DF
{iu}	Verb (adnominal form)	132,818	31	132,322	99.6
{meHN}	Noun	162	2	157	98.1
{niQpoN}	Noun	8,242	6	8,045	97.6
{kaNzuru}	Verb (adnominal form)	274	2	266	97.0
{simyureHsyoN}	Noun	227	5	226	96.9
{enueicikeH}	Noun	183	7	176	96.2
{taiiku}	Noun	151	3	145	96.0
{syoHzuru}	Verb (adnominal form)	116	2	106	94.0
{poi}	Suffix (adnominal form)	145	2	136	93.8
{yoi}	Adjective (ending form)	4,446	7	4,026	90.6

4.5 Entropy of word-forms

In the computation of variation rates presented above, we hypothesized that there is one and only one ‘standard’ DF for a given word-form, but this is a problematic hypothesis. There are many word-forms that have more than two ‘standard’ forms. For example, lexeme {niQpon} has two frequent word-forms /nihoN/ and /niQpoN/ both of which are registered in dictionaries. Similar examples include {yoi} (“good”, frequent DF being /yoi/ and /iH/), {iu} (“say”, /iu/ and /yuH/), {iku} (“to go”, /iku/ and /yuku/), {watasi} (1st person pronoun, /watasi/ and /atasi/), {mina} (“everybody”, /mina/ and /miNna/), and so forth.

In these word-forms, the rate of variation could change drastically depending on the choice of DF. In the case of {niQpoN} for example, the current rate of 97.6% (see Table 7) becomes 2.4% or even less, if we adopt /nihoN/ as the DF. Note that this is not at all a strange choice for the native speakers of Japanese. Clearly, an index of variability that does not make reference to 'standard' word-form is needed to avoid this kind of indeterminism in the quantification of word-form variation.

Entropy (in the sense of information sciences) is one such index. Entropy H of a probabilistic event E is the index of the predictability of E and is defined as

$$H=\Sigma P_i\, I(P_i)$$

where p is the probability distribution of the event E and I(pi) is defined as

$$I(P_i)= -\log_2 P_i$$

and is called the information (or information quantity) of the event.

If H=1 (unit of entropy is BIT), the event is as predictable as the result of coin tossing; the entropy of dice is about 2.585, showing that the prediction of dice is much more difficult than coin tossing.

Table 10 shows the entropy of word-forms previously shown in Tables 7 and 8. As predicted, entropy of {iu} in its adnominal form, or, {niQpoN} is low because most of the occurrences is occupied by a single word-form which happened not to be identified as the DF. On the other hand, entropy of {no}, {nani}, and {mina} are about 1.0 because in these items two equally frequent word-forms are observed. And, lastly, entropy of {yahari} is higher than 2.0 because there are many word-forms that are used more or less frequently: for the total occurrence of 11,746, /yaQpari/ (N=5793), /yahari/ (3999), /yaQpa/ (998), /yaQpai/ (256), /yaQpasi/ (112), /pari/ (112) and so forth.

Table 10: Entropy (H) of word-forms shown in Tables 7 and 8

LEXEME	GLOSS	POS (CF)	N	Freq. Non-DF	% Non-DF	H
{iu}	'say'	Verb (adnominal form)	132,818	132,332	99.6	0.587
{no}	'of '	Adnominal particle	153,521	79,829	52.0	1.033
{keredo}	'but'	Conjunction particle	47,032	26,534	56.4	1.477
{nani}	'what'	Pronoun	23,067	17,140	74.3	1.106
{iu}	'say'	Verb (ending form)	9,155	7,991	87.3	0.628
{Qte}	---	Adverbial particle	50,704	7,834	15.5	0.899
{niQpoN}	'Japan'	Noun	8,242	8,045	97.6	0.251
{kurai}	'even'	Adverbial particle	8,947	7,758	86.7	0.546
{ni}	'at'	Case particle	206,614	7,568	3.7	0.262
{yahari}	'after all'	Adverb	11,746	7,022	59.8	2.010
{sore}	'that'	Pronoun	44,000	6,016	13.7	1.146
{yoi}	'good'	Adjective (adnominal form)	5,950	5,177	87.0	0.687

LEXEME	GLOSS	POS (CF)	N	Freq. Non-DF	% Non-DF	H
{yoi}	‘good’	Adjective (ending form)	4,446	4,026	90.6	0.515
{moH}	‘anymore’	Adverb	18,501	3,669	19.8	0.799
{desu}	Copula	Aux. verb (ending form)	141,084	3,431	2.4	0.251
{de}	‘and then’	Conjunction	55,717	3,290	5.9	0.566
{mina}	‘everyone’	Noun	4,309	2,634	61.1	1.076
{mono}	‘thing’	Noun	31,794	2,373	7.5	0.464
{watasi}	‘I’	Pronoun	15,749	2,367	15.1	1.650
{soH}	‘so’	Adverb	29,698	2,327	7.8	0.516

5. Conclusion

As shown by the examples shown above, CSJ is an invaluable resource for the analysis of language variation. It is the current author’s wish to conduct these kinds of surveys more systematically to grasp the entire picture of language variation in spoken Japanese. And, ultimately, the result obtained from the CSJ should be compared to the variation of written language to get the full picture of the variation in the Japanese language.

Kikuo Maekawa
The National Institute for Japanese Language
3591-2 Midori-cho, Tachikawa-shi
Tokyo 190-8561, Japan
kikuo@kokken.go.jp

Linguistic Patterns in Spontaneous Speech, 51-75
2009-1-025-003-000265-1

Situational Characteristics and Register Variation: A Case Study of the Particle *suo* in Mandarin Chinese[*]

Jen Ting
National Taiwan Normal University

This article explores register variation by investigating the linguistic function of the particle *suo* in Mandarin Chinese in different registers. The data for analysis included corpora collected from editorials, magazines, fiction and speeches as well as oral corpora constructed by Tseng (2004). The results of our research show that *suo* serves ideational, (non-)contextual, personal and esthetic functions in the communicative situation. It is claimed that a dichotomy between written and spoken registers cannot fully account for the distribution of *suo* across registers. Rather, it is the situational characteristics of a register, written or spoken, that determine the appropriateness of *suo*'s occurrence in a register. The findings support the view that textual relations are defined by the situational characteristics shared among written and spoken registers.

Key words: register variation, particle *suo*, linguistic function, situational characteristic

1. Introduction

It has long been noticed that written language and spoken language exhibit distinct linguistic features. For example, to capture the linguistic differences between the two modes, Chafe (1982, 1985, Chafe & Danielewicz 1987) proposes four functional notions, namely, integration vs. fragmentation, and detachment vs. involvement. Under this approach, written language tends to show integrative devices such as nominalization,

[*] I would like to express my sincere gratitude to Dr. Shu-Chuan Tseng at Academia Sinica for generously providing the oral corpora she constructed, sharing her precious time discussing with me various issues in this article and offering valuable comments. Results of earlier stages of this research were presented at NACCL-18, West Washington University, 2006 and OCLF-3, Xuzhou Normal University, 2007. Thanks go to the audiences, in particular Zhiming Bao, Shengli Feng, Miao-Ling Hsieh, Mei-chun Liu, Kuo-ming Sung, Kathleen Tomlonovic and Yiming Yang, whose questions and comments have greatly improved the version of this paper. Finally, I would like to thank Prof. Rong-kui He for kindly helping with the statistics applied in this study. Mistakes are exclusively my own. Research for this article was supported by the NSC grant #94-2411-H-003-023.

whereas spoken language tends to exhibit fragmentation such as sentence-initial conjunctions. Similarly, written language is characterized by devices such as passive voice for distancing the writer from the audience, whereas spoken language is characterized by devices showing the speaker's involvement with the audience such as first person references. As a result, as summarized by Biber (1988:47), "in general, writing is claimed to be more structurally complex and elaborate than speech, … more explicit than speech, in that it has complete idea units with all assumptions and logical relations encoded in the text… more decontextualized or autonomous, than speech… less personally involved than speech and more detached and abstract than speech…".

These seemingly categorical statements about written and spoken language, however, apply to extremes on a continuum (Chafe 1982:49). If we take spontaneous conversational language and formal academic prose as two extremes, there are other styles of speaking which are more like writing, and other styles of writing which are more like speech. Furthermore, in addition to the differences that exist due to differences in the speaking and writing processes themselves, there are other differences that have arisen because of the varied contexts, purposes, and subject matters of both spoken and written language (Chafe & Danielewicz 1987:87). Therefore, when we identify functions of a particular linguistic feature, it is inadequate to characterize its distribution across communicative situations by mere reference to a single dimension (such as casual/formal; written/spoken; or attention paid to speech) (Biber & Finegan 1994:326). In this paper, we will support this view on register variation by investigating the use of *suo* in modern Chinese, a particle often claimed to be a remnant from Classical Chinese (e.g. by Chu 1987, Chiu 1995). The results of this study show that as it serves ideational, (non-)contextual, personal and aesthetic functions, *suo* cannot be characterized as being associated with written or formal registers, but rather that it is the situational characteristics of a register, written or spoken, that determine the appropriateness of *suo*'s occurrence in the register. A written/spoken dichotomy thus cannot adequately capture textual relations. We conclude that textual relations are defined by the situational characteristics shared among written and spoken registers (see Biber 1986, 1988 and subsequent works).

This paper is organized as follows: Section 2 presents basic facts of *suo* in modern and Classical Chinese and reviews previous proposals of the function of *suo* in modern Chinese. Section 3 describes the methodology of conducting this research, including description of the database and procedures. Results of this study are provided in §4. Based on these results, in §5, functions of *suo* are proposed and their implications for textual relations are discussed. Section 6 concludes this article by restating the thesis and pointing out issues for further studies.

2. Basic facts and previous analyses of the function of *suo*

The particle *suo* in modern Chinese most often occurs in relative clauses, as in (1a-b) and occasionally in passives, as in (1c).

(1) a. 李四所愛的人
Lisi suo ai de ren
Lisi SUO love DE person
'the person that Lisi loves'
b. 小偷所沒有偷走的那些首飾
xiaotou suo meiyou tou zou de naxie shoushi
thief SUO not-have steal away DE those jewel
'the jewelry that the thief didn't steal'
c. 兩千三百萬人卻總覺得自己的命運在被一個人所決定。(Editorial)
liang qian san bai wan ren que zong
two thousand three hundred ten:thousand person conversely always
juede ziji de mingyun zai bei yige ren suo jueding
feel self DE fate at BEI one person SUO decide
'Twenty three million people, on the contrary, always feel that their fates are being decided by one person.'

Such uses of *suo* are often said to be a remnant from Classical Chinese (e.g. Chu 1987, Chiu 1995) possibly because of the original and also most common occurrence of this particle in the relative-clause-like construction in Classical Chinese as indicated by the underlined sequences in (2) and its later occurrence in the passive construction in Classical Chinese in (3) (see Ting 2005, 2008 for discussion).

(2) a. 仲子所居之室（《孟子・滕文公下》）
Zhongzi suo ju zhi shi (Mengzi: Tengwengongxia)
Zhongzi SUO live ZHI room
'the room that Zhongzi lives'
b. 而語及所匿之事（《韓非子・說難》）
er yu ji suo ni zhi shi (Hanfeizi: Shuonan)
ER speak reach SUO hide ZHI thing
'But we might mention what was hidden (by him).'
c. 民所食之粟（adapted from《孟子・滕文公》）
min suo shi zhi su (adapted from Menzi: Tengwengong)
people SUO eat ZHI barley
'the barley that people eat'

(3) 常被元帝所使。(《顏氏家訓‧雜藝》)
chang bei Yuan di suo shi (Yanshi Jiaxun: Zayi)
often BEI Yuan emperor SUO order
'He was often ordered around by Emperor Yuan.'

Recently, studies of the *suo* construction in modern Chinese have achieved good results in the implications for the phrase structure, Case assignment, chain relations in Chinese syntax (e.g. Chiu 1995, Ting 2003). To illustrate, Ting (2003) argues that *suo* behaves on a par with pronominal clitics in Romance in many respects (cf. Chiu 1995) and proposes to analyze the licensing of *suo* along the line of Kayne's (1989, 1991) theory. Crucially, contra the claim of Chiu (1995), the licensing of *suo* is argued not to involve any functional projection associated with accusative Case licensing. Empirical arguments come from instances where *suo* is licensed by elements not receiving accusative Case (4) and where *suo* may occur in either the embedded or matrix clause (5). We will assume with Ting (2003, 2005, 2006) that a pronominal clitic analysis of *suo* in Mandarin Chinese is on the right track.

(4) a. [李四所服務/工作]的機構/地方
[Lisi suo fuwu/gongzuo] de jigou/difang
Lisi SUO serve/work DE organization/place
'the organization/place that Lisi serves/works in'

b. [那條小溪中所飄過]的枯葉
[natiao xiaoxi zhong suo piaoguo] de kuye
that small-stream middle SUO float:past DE withered-leaf
'the withered leaves that floated in the river'

(5) a. [我逼迫張三所購買]的書
[wo rang/bipo Zhangsan suo goumai] de shu
I make/force Zhangsan SUO buy DE book
'the book that I forced Zhangsan to buy'

b. ?[我所逼迫張三購買]的書
?[wo suo rang/bipo Zhangsan goumai] de shu
I SUO make/force Zhangsan buy DE book

In contrast to these results obtained from the study of syntactic properties of the particle *suo* in modern Chinese, there has not been much investigation of its function, namely, why *suo* is used. To the best of our knowledge, the scarce studies of the function of *suo* in the literature fall to four types of claim. First, *suo* is claimed not to have any effect on the clause and is thus optional (Zhang 1981). This conclusion is reached probably because most relatives containing *suo* appear to have their non-*suo* counterparts. This observation cannot be used to argue against the role played by *suo* in the clause. A similar case can be found with the optionality of the complementizer *that* in English. As pointed out by Biber (1988), the deletion of this complementizer occurs rarely in edited writing, which may be due to "the concern for elaborated and explicit expression in typical edited writing" (p.244). In other words, the use of the complementizer *that* is associated with the written register and gives formal flavor to the clause. Under this reasoning, then apparent optionality of a lexical item cannot be taken as the sole evidence for its lack of any function in the communicative situation. On the other hand, several studies have suggested a function served by *suo*. To begin with, Chao (1968) regards *suo* as an adverb added for emphasis; thus, (6) can mean 'the words he actually said' or 'all the words he said.'[1]

(6) 他所說的話
ta suo shuo de hua
he SUO say DE word
'words that he said'

Another claim that has been made in the literature is that *suo* is limited to the formal or written register (e.g. Chu 1987, Lu 1999, cf. Chiu 1995). Chu (1987:53), for example, points out "the presence of *suo* renders the whole utterance more formal… depending on the style of speech". Lu (1999:256) also observes that *suo* is mainly used in writing and rarely in speech. Assuming that typical written language is more formal than typical spoken language (Tannen 1982, Chafe 1982) if formality is defined as attention paid to language (cf. Labov 1972, Trudgill 1974), Lu's and Chu's claim can be subsumed under the same approach.

Lastly, *suo* has been observed to fulfill syllabicity requirement in the clause (Lu 1999, cf. Ting 2003). According to Lu (1999), the monosyllabic words that follow *suo* in (7) cannot stand alone; their disyllabic counterparts *dedao* 'gain' 得到, *fuyu* 'bestow' 賦予 and *zaoshou* 'undergo' 遭受 have to be used instead.

[1] Wang (1958) points out that *suo* may emphasize not only the agent but also what comes before *suo*. For example, in (i), *suo* does not 'specify' the agent *wo* 'I' but the temporal adverb *zuotian* 'yesterday'.

(i) 我昨天所買的書
wo zuotian suo mai de shu
I yesterday SUO buy de book
'the book I bought yesterday'

(7) a. 我個人在這個月中*(所)得的只是一點點微小的收穫了。
wo geren zai zhege yue zhong *(suo) de de
I individual at this month middle SUO gain DE
zhi shi yidiandian weixiaode shouhuo le.
only be a:bit meager harvest LE
'What I got during this month was just some meager results.'

b. 他的慈悲性情是上天*(所)賦的。
tade cibei xingqing shi shangtian *(suo) fu de
his mercy disposition be heaven SUO bestow DE
'His kind disposition is bestowed by God.'

c. 但是雖都叫著“提案”，因內容的不同，*(所)遭的命運也有著很大的差異。
danshi sui dou jiao-zhe "ti'an", yin neirong de butong,
but although all call ZHE proposal because content DE difference
*(suo) zao de mingyun ye you-zhe hen da de chayi
SUO undergo DE fate also have-ZHE very big DE difference
'Although all were called a proposal, due to differences in the content, they underwent big differences in their fates.'

Lu (1999) thus concludes that *suo* is used for the purpose of syllabicity in modern Chinese. In order to investigate whether the use of *suo* shows any variation across registers, we examine the distribution of *suo* in written and spoken corpora. Meanwhile, this examination is also intended to verify the validity of the various claims of the function of *suo*.

3. Methodology

3.1 Database

Corpora of both written and spoken language are included in this study. The database for written language includes the corpora of the register of editorials, magazine articles and fiction. A text corpus for each register was compiled. The corpus of editorials comprises 34 editorial samples containing the particle *suo* we collected from the major newspapers in Taiwan, including China Times, United Daily, Freedom Times and Commercial Times, etc. The corpus of magazines comprises 25 text samples containing the particle *suo* we collected from the major magazines in Taiwan, including Business Weekly, China Times Weekly, China Times Weekly, Marie Claire, etc. The corpus of fiction comprises 10 text samples containing the particle *suo* we collected from novels or novel excerpts written by Jin Ba (巴金), Ailing Zhang (張愛玲), Long Gu (古龍), Yutang Lin (林語堂), Kuang Ni (倪匡), Yao Qiong (瓊瑤), Guangzhong Yu (余光中),

etc. From each of the ten novels approximately 2,000 words were excerpted.

The texts were then segmented by the automatic segmentation system of Academia Sinica. The results of segmentation show that the total corpus for editorials contains approximately 26,229 words of running text; the total corpus for magazines contains approximately 22,900 words; and the total corpus for fiction contains approximately 21,063 words.

Table 1: Information of the written corpora

	Editorials	Magazines	Fiction
Text	34	25	10
Sources	China Times, United Daily, Freedom Times, Commercial Times, etc.	Business Weekly, China Times Weekly, TVBS Weekly, Marie Claire, etc.	novels or novel excerpts by Jin Ba (巴金), Ailing Zhang (張愛玲), Long Gu (古龍), Yutang Lin (林語堂), Kuang Ni (倪匡), Yao Qiong (瓊瑤), Guangzhong Yu (余光中), etc.
Words	26,229	22,900	21,063

The database for spoken language comprises 3 corpora of dialogues collected by Tseng (2004)[2] and one corpus of transcripts of speeches we collected on line. Description of the three dialogue corpora, Mandarin Conversational Dialogue Corpus (MCDC), Mandarin Map Task Corpus (MMTC) and Mandarin Topic-Oriented Conversation Corpus (MTCC), is summarized in Table 2.

Table 2: Information of the spontaneous spoken corpora

	MCDC	MMTC	MTCC
# of dialogues	8	26	29
# of hours transcribed	6.5 hrs	5 hrs	11 hrs
Topic	Free	Map task	News event
# of words	84,165	30,390	91,408

MCDC consists of 8 transcribed dialogues on free topics between two strangers, totaling 6.5 hours. The transcription comprises 84,165 words after segmentation. MMTC consists of 26 task-oriented dialogues produced by two participants who know each other well, totaling 5 hours. These dialogues are task-oriented because one participant with a detailed map needs to explain to the other with a simplified map how to get to a

[2] The databases are downloadable on the website http://mmc.sinica.edu.tw/.

particular destination. The transcription comprises 30,390 words after segmentation. MTCC consists of 29 topic-oriented dialogues between two participants who were familiar with each other, 11 hours recording in total. These dialogues are topic-oriented because each pair was asked to choose one topic related to an event having taken place in 2001 and to talk about it. The transcription comprises 91,408 words after segmentation by the segmentation system at Academia Sinica.

In addition to the data on informal spoken language, it is also important to consider an oral register which is more in the direction of writing. For this purpose, we collected five transcriptions of speeches on the Internet. Judging from these transcriptions, they are prepared speeches, planned but without orally reading a written text. From each of the transcriptions approximately 3,000 words were excerpted and then segmented by the segmentation system of Academia Sinica.

Table 3: Information of the corpora of speeches

	Speeches
Text	5
Sources	Talks by Han-Ding Hong, Ming-Hui Wang, Yong-Bao Hu, Ao Li and Xiang-Fa Yang
Words	14,904

3.2 Procedures

Under our analysis, clauses containing *suo* are divided into four types. In type I data,[3] omission of *suo* apparently does not yield ungrammaticality. *Suo* of type II, in contrast, is required in the clause; in other words, omission of it would render the sequences unacceptable. Type III data involve occurrence of *suo* in fixed expressions. By fixed expressions, we mean that the *suo* sequence is used as an idiomatic expression and may not have an internal structure for the speaker. E.g. *qian suo wei you de* ‘unprecedented’ 前所未有的 and *zhong suo zhou zhi* ‘well-known’ 眾所周知 may simply be stored as idioms in the speaker’s mental lexicon, equivalent to *unprecedented* and *as well-known* in English. Finally, *suo* of type IV is licensed by passivization.[4] Examples illustrating the four types are given in (8) to (11) respectively.[5]

[3] With thanks to Miao-Ling Hsieh for pointing out such labels to me.

[4] In most cases, *suo* of this type is droppable without yielding any ungrammaticality. In comparison with *suo* licensed by relativization, *suo* licensed by passivization is relatively rare.

[5] Due to limitation on space, the readers are referred to Ting (2006b) for more examples illustrating various points made in this paper.

(8) Type I

a. 「湯」所扮演的角色看似微不足道。(Magazine)
'tang' suo banyan de jiaose kan si wei bu zu dao
soup SUO play DE role look like insignificant not deserve say
'The role played by "soup" looks very insignificant.'

b. 我想這是一般人所無法擁有的高檔待遇。(Magazine)
wo xiang zhe shi yiban ren suo wu fa yongyou de
I think this be ordinary person SUO no way have DE
gaodang daiyu
high:class treatment
'I think this is the luxurious treatment that ordinary people cannot have.'

(9) Type II

a. 也高於國務院總理溫家寶在十屆全國人大會議中所宣稱。(Editorial)
ye gaoyu guowuyuan zongli Wenjiabao zai shijie quanguo
also higher State:Council Premier Wen:Jia-bao at tenth national
renda huiyi zhong suo xuancheng
People's:Congress meeting middle SUO claim
'(It is) also higher than what the Premier Wen, Jia-bao claimed at the tenth National People's Congress.'

b. 打破海上由美日安保所形成封鎖之勢 (Magazine)
dapuo hai shang you Mei Ri an bao suo xingcheng
break sea on by USA Japan security protection SUO form
fengsuo zhi shi
blockade ZHI situation
'[It] broke the blockade situation that was formed by mutual cooperation and security between the United States and Japan'

c. 從未有人能像李安這般成爲奧斯卡揭曉前後佳評所集的中心。(Editorial)
cong wei you ren neng xiang Li An zhe ban chengwei Aosika
ever not have person can like Li An this way become Oscar
jiexiao qian hou jia ping suo ji de zhongxin
announce before after good comment SUO accumulate DE center
'There has been no one who becomes the center of good comments around the time the results of Oscar awards are announced.'

(10) Type III

a. 品牌與產品背後的藝術家、設計師及創意人員，都受到前所未有的重視。(Magazine)
pinpai yu chanpin bei hou de yishujia, shejishi ji
brand and product back behind DE artist designer and
chuangyi renyuan, dou shoudao qian suo wei you de zhongshi
creativity staff all receive before SUO not have DE attention
'Artists, designer and creativity staff behind the brands and products all receive unprecedented attention.'

b. 眾所周知，奧斯卡是由美國影藝學院的五千名成員及受邀影人共同投票。(Editorial)
zhong suo zhou zhi, Aosika shi you Meiguo ying yi
people SUO around know Oscar be by USA movie art
xueyuan de wuqianming chengyuan ji shou yao
academy DE five:thousand member and receive invite
ying ren gongtong tou piao
movie people together cast vote
'As well-known, Oscar Awards are voted by the five thousand members of AMPAS together with the invited workers in the movie industry.'

(11) Type IV
香奈兒的創意再度爲主流所崇拜。(Magazine)
Xiangnaier de chuangyi zaidu wei zhuliu suo chongbai
Chanel DE creativity again WEI mainstream SUO worship
'Creativity of Chanel is once again worshiped by the mainstream.'

There are some expressions containing *suo* which are not considered in this study including: *suo* occurring in the conjunction item *suoyi* 'consequently' 所以, *suo* used as part of a nominal such as *yanjiu suo* 'graduate school' 研究所, *paichusuo* 'precint police station' 派出所, *cesuo* 'toilet' 廁所, *suode* 'income' 所得, and *suoyouquan* 'right of possession' 所有權, and also *suo* occurring in the expression *suowei(de)* 'so-called' 所謂(的) and *suoyou(de)* 'all' 所有(的).

Both quantitative and qualitative study was then conducted. Quantitatively, the total frequency of *suo* in each register was counted. Furthermore, tokens of *suo* of each type (I, II, III or IV) were counted and percentage of the types in each corpus was calculated for both the written and spoken database. Furthermore, the frequency counts of *suo*'s occurrence were normalized to a text length of 1000 words. Such normalization, according to Biber (1988), is necessary for conducting a cross-register comparison.

Qualitatively, we examined the patterns of Type II *suo*'s occurrence and also investigated whether *suo* co-occurs with emphasis expressions. Since other types of *suo* may involve some other factors such as syllabicity, only Type I *suo* is considered regarding its co-occurrence with emphasis expressions. Emphasis expressions were classified into two types: lexical expressions such as emphatics and amplifiers and syntactic devices such as pseudo-clefts. According to Biber (1988:241), emphatics "simply mark the presence (versus absence) of certainty while amplifiers indicate the degree of certainty towards a proposition;" amplifiers, on the other hand, have the effect of boosting the force of the verb. Examples of these expressions in English are given in (12).

(12) a. Emphatics: for sure, a lot, such a, really, so, just, most, more.
b. Amplifiers: absolutely, altogether, completely, enormously, entirely, extremely, fully, greatly, highly, intensely, perfectly, strongly, thoroughly, totally, utterly, very.

An important syntactic device of expressing emphasis is the pseudo-cleft construction. In English the class of this construction is argued by Collins (1991) to comprise three subclasses: *wh*-clefts, *th*-clefts headed by lexically empty 'pro-nouns' such as *thing*, *one*, *place*, *kind*, etc, and *all*-clefts illustrated in (13a-c) respectively.

(13) a. What the car needs is a new battery.
b. The thing the car needs is a new battery.
c. All the car needs is a new battery.

Similarly, Chinese also has pseudo-clefts as discussed by Tang (1980) and Huang (1988). The examples in (14) are taken from Tang (1980:252).

(14) a. 湯先生十五年前在美國學的是語言學。
Tang xiansheng shiwu nian qian zai Meiguo xue de shi
Tang Mr. fifteen year before at USA learn DE be
yuyanxue
linguistics
'What Mr. Tang studied in the USA 15 years ago was linguistics.'
b. 十五年前在美國學語言學的是湯先生。
shiwu nian qian zai Meiguo xue yuyanxue de shi
fifteen year before at USA learn linguistics DE be
Tang xiansheng
Tang Mr.
'The one who studied linguistics in the USA fifteen years ago was Mr. Tang.'

4. Results

The overall distribution of *suo* in the three written corpora and in the four spoken corpora are shown in Table 4 and Table 5 respectively.

Table 4: Distribution of *suo* in the written corpora

	Editorial	Magazine	Fiction
Type I	38 (52.7%)	39 (65%)	5 (38.4%)
Type II	25 (34.7%)	8 (13.3%)	1 (7.6%)
Type III	6 (8.3%)	9 (15%)	6 (46.1%)
Type IV	3 (4.1%)	4 (6.6%)	1 (7.6%)
Total	72 tokens (100%)	60 tokens (100%)	13 tokens (100%)
Mean frequency	2.74/1000 words	2.62/1000 words	0.61/1000 words

Table 5: Distributon of *suo* in the spoken corpora

	MCDC	MMTC	MTCC	Speech
Type I	18 (54.5%)	0	22 (62.9%)	10 (33.3%)
Type II	6 (18.1%)	0	6 (17.1%)	12 (40.0%)
Type III	8 (24.2%)	0	7 (20.0%)	3 (10.0%)
Type IV	1 (3.0 %)	0	0 (0 %)	5 (16.6%)
Total	33 (100 %)	0	35 (100 %)	30 (100 %)
Mean frequency	0.39/1000 words	0/1000	0.37/1000 words	2.01/1000words

The statistics result showing significant differences among the seven registers is given in Table 6.[6]

Table 6: Significant differences among the seven registers

	Editorial 2.74/1000 72/26229	Magazines 2.62/1000 60/22900	Speech 2.01/1000 30/14904	Fiction 0.61/1000 13/21063	MCDC 0.39/1000 33/84165	MTCC 0.37/1000 35/94108	MMTC 0/30390
Editorial		0.27	1.44	5.43**	10.79**	11.40**	9.14**
Magazines			1.18	5.15**	10.15**	10.71**	8.93**
Speech				7.62**	7.23**	7.62**	7.82**
Fiction					1.40	1.58	4.33**
MCDC						0.22	3.45**
MTCC							3.36**
MMTC							

Note. ***p*<.01

[6] With thanks to Prof. Rong-kui He from the Graduate Institute of Information and Computer Education at NTNU for helping me with statistics involving significant differences.

4.1 Mean frequency and significantly different distribution of *suo* across the registers

We shall first consider the mean frequency of *suo* in the written corpora. As shown in Table 4, the frequency of *suo* in editorials is slightly higher than in magazine articles (2.74/1000 in editorials vs. 2.62/1000 in magazines) but both are much higher than the mean frequency in fiction (i.e. 0.61/1000). Statistics indicate that in terms of *suo*'s frequency, editorials and magazines show no significant difference but either of them shows significant differences with fiction. This indicates that the register of editorials and of magazines have similar characteristics in licensing the occurrence of *suo* while fiction, though a written register, should be distinguished from both of them.

When we examine the result of oral corpora shown in Table 5, the two corpora MCDC and MTCC exhibit almost the same mean frequency of *suo*. Statistics also show that these two oral corpora do not have a significant difference in terms of *suo*'s frequency, indicating that they share similar characteristics in the licensing of *suo*'s occurrence. Another corpus MMTC, however, presents a striking contrast in not including any token of *suo* out of a total of 30,390 words. This absence of *suo* in MMTC has a statistically significant difference with the other two spontaneous dialogue corpora that contain some tokens of *suo*. If we compare the two spontaneous dialogue corpora with another oral register, speeches, we find a statistically significant difference with respect to their frequency of *suo*. In other words, the oral corpora, in terms of the frequency of *suo*, now fall into three groups: Speeches have the highest frequency of *suo*, MCDC/MTCC have rare occurrences of *suo*, and MMTC none. This finding indicates inadequacy of simply characterizing *suo* as rarely occurring in spoken registers since the distribution contrast of *suo* among the three groups of oral corpora requires an explanation, to which we shall return in §5.

When we do comparison across the written and oral corpora, there are some interesting results that emerge. Notice that as shown in Table 6, in terms of the frequency of *suo*, the oral corpus speech is statistically non-distinct from the two written register editorial and magazine whereas the written corpus fiction is statistically non-distinct from the two spontaneous dialogue corpora. Given these findings, all the corpora now fall into three groups in terms of frequency of *suo* as shown in (15).

(15) editorial/magazine/speech > fiction/MCDC/MTCC > MMTC

That is, editorials/magazines/speeches have the highest frequency of *suo*, fiction/MCDC/MTCC has low frequency of *suo*, and MMTC has none. All the three groups show statistically significant differences among one another. Once again, a dichotomy between written and spoken registers is shown to be inadequate in characterizing use of the particle *suo*.

4.2 Distribution of types

We shall now consider the distribution of types across and within each register. Comparing the percentage of types across the registers, we found that there is a higher percentage of Type III *suo* than Type I and Type II *suo* in the group of registers consisting of MCDC, MTCC and fiction than in the group consisting of editorials, magazines and speeches (Fiction: 46.1%, MCDC: 24.2%, MTCC: 20.0%, vs. Editorial: 8.3%, Magazine 15%, Speech 10.0%). We speculate that this may have to do with the former group using relatively less Type I and II *suo* than the latter group, a fact that may be due to the ideational and (non-)contextual function of *suo* (to be discussed in §5), thus making the former group exhibit a relatively higher percentage of Type III *suo* than the latter group. Comparing the percentage of types within the registers, as shown in Table 4 and Table 5, Type IV *suo*, namely *suo* associated with the passives, has the lowest frequency counts in comparison with other types in each corpus containing *suo*. Similarly, *suo* of Type III, namely those associated with fixed expressions, also has relatively low frequency counts in comparison with other types in each of the corpora containing *suo*. These facts suggest that in modern Chinese use of *suo* is mainly associated with relative clauses.

Furthermore, within each corpus containing *suo*, percentage of *suo* of Type I is much higher than that of *suo* of Type II, showing that the use of *suo* must have some functional purpose because *suo*'s occurrence is apparently not required in many cases. In addition, the two highest percentages of Type II *suo* in comparison with other types in each of the corpora containing *suo* are revealed in the register of speeches and of editorials. This may have to do with the attitude or stance generally encoded in these registers and with the characteristics of Type II *suo*. Before the discussion of the reason of the frequent use of Type II *suo* in speeches and editorials is presented in §5, we shall turn to the characteristic of Type II *suo* in the next sub-section.

4.3 *Suo* of Type II

It is found that occurrences of Type II *suo* fall into two types: those involving imitation of Classical Chinese style as shown in (9a-b) and (16) and others involving prosodic requirement in modern Chinese as shown in (9c) and (17).

(16) 我 (unrecognizable_speech_sound) 我所知道最大的業務 (MCDC)
wo (unrecognizable_speech_sound) wo suo zhidao zui da de yewu
I I SUO know most big DE transaction
'the biggest transaction that I know of'

(17) a. 是在告訴我們說 E 人生所屬的四大象限 (inhale) 就是說…(MTCC)
shi zai gaosu women shuo E rensheng suo shu de
be at tell we say E life SUO belong DE
si da xiangxian (inhale) jiushi shuo
four big quadrant just say
'(It) tells us that the four quadrants that life belongs to, in other words…'

b. 所以可以花時間鑿出所要的精鋼筆尖。(Magazine)
suoyi keyi hua shijian zao chu suo yao de
so can spend time chisel out SUO want DE
jing gang bi jian
refined steel pen tip
'So one can spend time chiseling a refined steel tip of a pen that one wants.'

Some properties of *suo* in Classical Chinese are in order; first, it is not allowed to be optional as shown in (18a). The *suo* construction, furthermore, does not have to include an overt head noun as shown in (18b) (see Ting 2005 and references cited there). In addition, a linker *zhi* between the relative clause containing *suo* and the head noun is not required as shown in (18c) (see Ting 2008 for the syntactic differences between *zhi* and its modern Chinese counterpart *de*).

(18) a. 民所食者
min suo shi zhe
people SUO eat ZHE
'what people eat'

b. 行法志堅，好修正其所聞，以矯飾其情性。(《荀子・儒效》)
xing fa zhi jian, hao xiuzheng qi suo wen,
behavior legitimate will determined like correct he SUO hear
yi jiaoshi qi qingxing (Xunzi: Ruxiao)
YI modify his nature
'His behavior is reasonable and his will is determined; he likes to correct what he hears in order to modify his nature.'

c. 和氏璧，天下所共傳寶也。（《史記・廉頗藺相如列傳》）
He shi bi, tianxia suo gong chuan
HE surname jade world SUO together recognize
bao ye (Shiji: Lianpo Linxiangru Liezhuan)
treasure YE
'The jade Heshi *(is) the treasure that is unanimously recognized by the world.'

Now we see that the examples (9a-b) and (16) with obligatory occurrence of *suo* exactly reflect these characteristics of the *suo* construction in Classical Chinese. In (9a), there is no head noun in the relative clause; in (9b), a linker *zhi* used in Classical Chinese between the relative clause and the head noun is present; in (16), there is no such linker *zhi*. It is therefore not surprising that *suo* is obligatory in these modern Chinese examples as in Classical Chinese.

The other environment for the obligatory occurrence of *suo* is when some prosodic constraint of modern Chinese is at work. The other subtype of Type II *suo* involves a monosyllabic verbal bound morpheme such as *ji* and *shu* in (9c) and (17). Such morphemes may have been free morphemes in earlier stages of Chinese but in modern Chinese they cannot stand alone and therefore must form a phonological word with *suo* in these examples. On the other hand, the other subtype, though rare in number, reveals a different prosodic constraint in modern Chinese. Though monosyllabic, *yao* in (17b) is not a verbal bound morpheme in modern Chinese but its occurrence without *suo* in the clause would yield unacceptability.

4.4 Co-occurrence with emphatic expressions

It is also observed that *suo* quite often co-occurs with emphatic adverbs and amplifiers. Illustrated by the italicized parts in (19) are emphatic adverbs meaning *best*, *most* and *more* and illustrated by the italicized parts in (20) and (21) are amplifiers meaning *perfect, enormous, fully*, *specially*.

(19) a. 《亞元》雜誌所評選的年度該國*最佳* 經營管理公司，Infosys 連續七年奪冠。(Magazine)
<Yayuan> zazhi suo pingxuan de niandu
Yayuan magazine SUO appraise:select DE annual
gai guo *zui jia* jingying guanli gongsi,
above-mentioned country most good operate manage company
Infosys lianxu qi nian duo guan
Infosys consecutively seven year take champion
‘In terms of the annual best company of operation and management, Infosys won this honor in consecutively seven years.’

b. 它所被要求的是能夠伴隨著用筆者一生，書寫生命中*最* 重要的時刻，書寫給生命中*最* 重要的人。(Magazine)
ta suo bei yaoqiu de shi nenggou bansui-zhe yongbizhe
ta SUO BEI require DE be can accompany-ZHE pen-user

yisheng, shuxie shengming zhong *zui* zhongyao de shike,
one:life write life middle most important DE moment
shuxie gei shengming zhong *zui* zhongyao de ren
write give life middle most important DE person
'What is required on it is to accompany the pen-user all his/her life, for him/her to write about the most important moment in life and to write to the most important people in life.'

c. 我們早期所學的東西都*比較*(short_break)(inhale) 窄。(MCDC)
women zaoqi suo xue de dongxi
we early:stage SUO learn DE thing
dou *bijiao* (short_break)(inhale) zhai
all more narrow
'What we learned in the early stage was rather narrow.'

(20) a. 六角星*完美的*四十三個切割面所散發出的亮度,讓人驚豔。(Magazine)
liujiaoxing *wanmeide* sishisange qiegemian suo sanfa chu de
Star:of:David perfect forty-three facet SUO emit out DE
liangdu, rang ren jingyan
brightness make person impressed
'The brightness that the perfect 43 facets emitted impressed people.'

b. 從進入飯店開始,即能看見鄉野風格的裝潢,融入各種*巨大的*機芯圖所傳達的機械結構之美。(Magazine)
cong jinru fandian kaishi, ji neng kanjian xiangye
from enter hotel begin immediately can see county
fengge de zhuanghuang, rongru gezhong *judade* jirui tu
style DE décor blend various giant movements picture
suo chuanda de jixie jiegou zhi mei
SUO convey DE machinery structure ZHI beauty
'Ever since entering the hotel, one can immediately see the décor of county style blended with the beauty of machinery structure conveyed by various gigantic pictures of movements.'

(21) a. 2 樓是*專* 爲時裝秀以及特別的時尚活動所做的空間設計。(Magazine)
erlou shi *zhuan* wei shizhuang xiu yiji tebiede
second:floor be specially for fashion show and special
shishang huodong suo zuo de kongjian sheji
fashion activity SUO do DE space design
'The second floor is a spatial design specially made for fashion shows and special fashion activities.'

b. 因此湯汁不必加糖也*滿*是海鮮肉質所散發出來的鮮甜。(Magazine)
yinci tangzhi bu bi jia tang ye *man* shi haixian
consequently soup not need add sugar also fully be seafood
rouzhi suo sanfa chulai de xiantian
meat:texture SUO emit out DE freshness
'Consequently, the soup is full of freshness emitted by the texture of seafood without adding sugar.'

Other lexical emphasis expressions that are observed to co-occur with *suo* are expressions with intensifying force, for example, the universal quantifier *jie* in (22) and *meiyige* in (23). Note that the latter also illustrates an environment of a contrastive focus with the syntactic pattern *bushi*… *ershi* 'not … but …'.

(22) 至於湯頭所使用的複方花茶原料，*皆*採自法國及德國山區。(Magazine)
zhiyu tangtou suo shiyong de fufang huacha yuanliao,
regarding soup SUO use DE compound herb:tea raw:material
jie cai zi Faguo ji Deguo shanqu
all pick from France and Germany mountain:area
'Regarding the raw materials of the compound herb tea which were used in the soup, they were all picked in the mountain areas of France and Germany.'

(23) 品牌區隔*不*單指在產品廣告，*而是*員工所做的*每一件*事。(Magazine)
pinpai quge *bu* dan zhi zai chanpin guanggao, *er shi*
brand segmentation not simply refer at product advertisement but be
yuangong suo zuo de *meiyijian* shi
employee SUO do DE every thing
'Brand segmentation refers to not only product advertisements, but also everything that the employees do.'

In addition to co-occurrence with these emphasis-associated lexical expressions, *suo* is also found to appear in the pattern of pseudo-clefts, as indicated by the italicized parts in (24) taken from the written corpora (24) and in (25) taken from the oral corpora.

(24) Written corpora
a. 「潮流」這件事一向都不*是他所在意的*。(Magazine)
"chaoliu" zhejian shi yixiang dou bu *shi ta suo zaiyi de*
trend this thing consistently all not be he SUO mind DE
'The issue "trend" has always not been what he minds.'

b. 這意味著兩岸將重新陷入新一波的「不確定」，這豈*是美國所樂見的*。(Editorial)
zhe yiwei-zhe liang an jiang chongxin xian ru xin yi po
this mean-ZHE two coast future again fall into new one wave
de “bu queding”, zhe qi *shi Meiguo suo le jian de*
DE not certain this QI be USA SUO happy see DE
‘This means that two sides across the strait will fall into another wave of “uncertainty”; is this what the USA is happy to see?’

(25) Spoken corpora

a. 像 (inhale) 很多東西*是你以前所沒有學到的*。(MCDC)
xiang(inhale)henduo dongxi *shi ni yiqian suo meiyou xuedao de*
like many thing be you before SUO not learn DE
‘[It is] like many things you didn’t learn before.’

b. 成果到最後 (pause) 不*是他[6iaN2]所想像的*。(MTCC)
chengguo dao zuihou (pause) bu *shi ta [6iaN2] suo xiangxiang de*
result arrive final not be he SUO imagine DE
‘The ultimate result is not what he can imagine.’

c. *我所要與各位同學分享的是* 從空間性角度分析原住民族運動。(Speech)
wo suo yao yu gewei tongxue fenxiang de shi cong
I SUO want with every classmate share DE be from
kongjianxing jiaodu fenxi yuanzhumin yundong
spatial angle analyze aboriginal movement
‘What I’d like to share with you is analyzing aboriginal movements from a spatial perspective.’

Since Type I *suo* has the most frequency counts in almost all of the corpora containing *suo*, we counted the occurrences of such *suo*, whose environments include some emphasis/focus expressions or patterns and obtained the results as summarized in Table 7.

Table 7: Frequency counts of optional occurrences of *suo* involving emphasis

	Editorial	Magazine	Fiction	Speech	MCDC	MTCC
Type I *suo* involving emphasis	20	22	5	10	10	13
Total occurrences of Type I *suo*	38	39	3	6	18	22
Percentage	52.63%	56.41%	60%	60%	55.55%	59.09%

These results indicate that an important function of *suo* is to give intensifying force in the clause. In other words, by examining the expressions that co-occur with *suo*, we have confirmed Chao's claim that *suo* may be added for emphasis.[7]

Summarizing, in this section we have presented the results obtained from investigating the mean frequency of *suo* across the registers, distribution of types of *suo* within and across each corpus containing *suo*, characteristics of Type II *suo* and the relatively high co-occurrence with emphatic expressions of *suo* (of Type I).

5. Discussion

In this section, we discuss how these characteristics of *suo*'s distribution may be analyzed. Under a multi-dimensional approach, we claim that *suo* serves ideational, (non-)contextual, personal and aesthetic functions and propose that these functions of *suo* may be closely associated with its explicit form.

The multi-dimensional approach to register variation is advocated by Biber (1988), Biber & Finegan (1994, 2001), among others. According to them (Biber & Finegan 1994: 320), there are two competing forces in communication: the 'be quick and easy' mandate and the 'be clear' mandate and they can be identified by the use of economy and elaboration features respectively. "Many variables and optional expressions can be regarded as more or less elaborated (alternatively, more or less compressed or economical)." (ibid.) Illustrating examples include omission/retention of the marker *that* from a complement clause and absence/presence of a prepositional phrase for a noun/verb phrase. "Because of their differing communicative demands, different registers have a functional preference for the clarity mandate or for the ease mandate." (Biber & Finegan 1994:321) The clarity mandate is favored by stereotypically literate varieties such as academic prose while the ease mandate is favored by stereotypically oral varieties such as conversation. Regarding other registers, frequencies of particular features fall between the two extreme mandates and reflect the situational characteristics of the registers. An important claim in this approach is thus that "the distribution of linguistic features across communicative situations cannot be adequately characterized by reference to a single dimension (such as casual/formal; written/spoken; or attention paid to speech); rather, a multidimensional framework is needed (see Hymes 1974, Biber 1988)." (Biber & Finegan 1994:326)

Given this characterization of register variation, the fact that *suo* tends to appear in stereotypically literate registers naturally follows. In contrast to Classical Chinese, "*suo* V" and "V" may alternate with each other in most cases of modern Chinese as shown in Table 4. Thus, as an apparently optional expression, *suo* can be viewed as an elaboration

[7] But this is certainly not the only function of *suo*. See more discussion in §5.

feature. Elaboration features have an ideational function, presenting informational rather than interactive communicative purposes. Elaboration and explicitness of expressions are necessary in stereotypically literate registers that arise in circumstances characterized by careful production, informational purposes and relatively little shared context between interlocutors. If *suo* is an elaboration feature, then its tendency to occur in literate registers but not oral registers is not a surprise. This is supported by the quantitative results of the frequency counts of *suo* in the registers we have seen. In editorials and magazines, which are stereotypically literate varieties, the frequency counts of *suo* are the highest and those in the two spontaneous dialogue corpora, which are stereotypically oral varieties, are the lowest.

Worth noting is the distribution of *suo* in speech and fiction, the two registers falling between the stereotypically literate and oral extreme. Despite being an oral register, speeches show a mean frequency statistically non-distinct from those of the two written registers, editorials and magazines; fiction, though a written register, shows a mean frequency statistically non-distinct from those of the two spontaneous dialogue corpora MCDC and MTCC.[8] These findings indicate that the distribution of *suo* cannot be adequately characterized as relating to a single parameter such as written/spoken as in Lu (1999:256). The written/spoken dichotomy between registers fails to account for why speeches as a spoken register are grouped as editorials and magazines and why fiction as a written register is grouped as the two spontaneous dialogue corpora in terms of licensing *suo*'s occurrence. On the other hand, the distribution of *suo* also cannot be captured by a single parameter such as formal/informal as in Chu (1987:53). Recall that in the task-oriented corpus MMTC, there is not a single token of *suo* found, which presents a statistically significant difference with the distribution of *suo* in the other two dialogue corpora. Given the difficulty of characterizing the register variation between MMTC and the other two dialogue corpora in terms of formality, absence of *suo* in corpora like MMTC shows that some other factor than formality determines *suo*'s distribution.[9] The results that we obtained thus do not support a single dimension approach to register

[8] Some may attribute the low frequency of *suo* shared by fiction and the two spontaneous corpora to the assumption that fiction includes lots of dialogues. This line of reasoning, however, may not hold. As argued later in the text, fiction is a register intermediate between a stereotypically literate and a stereotypically oral one. Then it is expected that the frequency of *suo* in fiction should be intermediate between the stereotypically written and oral extreme, contrary to fact. Furthermore, although fiction is characterized as having features of stereotypically written and oral registers, i.e. both informational and involved (Biber 1988), it is a written register and most often does not contain as many dialogues as spontaneous dialogues do, which is the case for the novels excerpted in this study.

[9] With thanks to Shengli Feng for sharing with me the opinion that *suo*'s occurrence is licensed by formality.

variation but rather a multi-dimension one as proposed by Biber (1988) and Biber & Finegan (1994, 2001), among others.

The fact that speeches show a mean frequency of *suo* as high as that of editorials and of magazines is not surprising if *suo* is analyzed as an elaboration feature because speeches, according to Biber (1988:154), are highly informational. However, the elaboration feature account of *suo* cannot fully explain its relatively low distribution in the fiction as in the two spontaneous dialogue corpora. According to Biber (1988:167), fiction is a register characterized by both informational and involved production. If *suo* exclusively serves as an elaboration feature, its distribution in fiction is expected to be intermediate between the stereotypically written and oral extreme, contrary to fact. Because its distribution in the fiction is grouped as those in the two spontaneous dialogue corpora, we propose that in addition to being an elaboration feature with an ideational function, *suo* also serves a (non-)contextual function, conveying highly explicit, context-independent, endophoric references. Registers such as editorials and magazine writing require highly explicit, text-internal reference, while registers such as conversation permit extensive reference to the physical and temporal situation of discourse, showing a high dependence on the context and thus conveying exophoric references. Like conversations, fiction also makes exophoric reference in the sense that "there is a fictional situation that is referred to directly in the text … the context of discourse production is not the same as the context of events" and that "the reader understands this reference in terms of the internal physical and temporal situation developed in the text rather than any actually existing external context" (Biber 1988:148). The similar distribution of *suo* in fiction and in the two spontaneous dialogue corpora receives a reasonable explanation if *suo* not only serves as an elaboration feature but also conveys endophoric references.

This interpretation of the function of *suo* as conveying endophoric references is further supported by the absence of *suo* in the corpus MMTC. As described in §3, this corpus consists of 26 task-oriented dialogues, each of which was produced by two participants who knew each other well. In contrast to the other two spontaneous dialogue corpora MCDC and MTCC, which contain dialogues more resembling regular conversations, the dialogues in MMTC are task-oriented because one participant with a detailed map had to explain to the other with a simplified map how to get to a particular destination. In such a task, extensive reference to the spatial situation described in the map makes the dialogue even more exophoric than a regular conversation. If a function of *suo* is to convey highly explicit, context-independent, endophoric references, then it is not surprising that there is not a single token of *suo* found in MMTC, a fact that presents a sharp contrast with the distribution of *suo* in the other two spontaneous dialogue corpora.

Still another function of *suo*, we claim, is a personal one. According to Biber (1988: 34), "personal functions include markers of group membership, personal style and attitudes towards the communicative event or towards the content of the message." This explains the occasional occurrence of *suo* in the two spontaneous dialogue corpora. Spontaneous dialogues are stereotypically oral registers, which according to Biber & Finegan (1994, 2001) are subject to the 'ease' and not the 'clarity' mandate. If *suo* is only used for the purpose of elaboration or conveying endophoric references, it would be expected to be absent in those two corpora, contrary to fact. Then in these registers conforming to the 'ease' mandate, why do the interlocutors bother to use *suo*? Given the high co-occurrence of *suo* with emphasis expressions we have seen, we suggest that the use of *suo* in the two spontaneous dialogue corpora mainly conveys emphasis, showing the speaker's feelings, judgments or attitudinal 'stance' towards the content.

Lastly, we shall point out the aesthetic function of *suo*. According to Biber (1988: 36), "aesthetic functions are those relating to personal or cultural attitudes about the preferred forms of language". For example, although contraction is usually attributed to be a consequence of fast and easy production, Biber (1988:243), based on the findings of Biber (1986) and Chafe & Danielewicz (1987), suggests that "the use of contractions seems to be tied to appropriateness considerations as much as to the differing production circumstances of speech and writing". *Suo*, a widespread particle in Classical Chinese, is used by Li & Thompson (1982) as one of the examples illustrating the significant role played by Classical Chinese in accounting for the gulf between spoken and written Chinese. They show that the phrase *shao you suo wu* "gain a little bit awareness" occurring in contemporary Chinese writing is in fact in the mold of Classical Chinese. According to them (ibid.: 87), most of the educated Chinese "share the feeling that the succinctness of the classical style carries with it an elegance and pithiness not found in the colloquial style, and inevitably slip into the classical tradition in their writing". We suggest that it is such "elegance" associated with *suo* that gives it an aesthetic function. We shall illustrate by comparing the use of *suo* and its alternative forms. Consider the sequence *suo shu* 'SUO belong' in (17b). In the first place the speaker does not have to select the monosyllabic verb *shu* 'belong', which requires the accompanying use of *suo* because instead of *suo shu* 'SUO belong', a disyllabic *shuyu* 'belong' could be used. But a clear difference between these two forms is that *suo shu* 'SUO belong' carries a Classical Chinese flavor, which gives "elegance" to the writing. Likewise, the speaker does not have to imitate classical Chinese grammar by using *suo* and may instead use modern Chinese grammar in many instances. The choice of *suo* over its alternative forms thus supports our claim for its aesthetic function. As we have seen in Table 4 and Table 5, Type II *suo*, namely the type of *suo* often associated with Classical Chinese grammar, has a relatively high frequency in editorials and speeches. These two registers, according to Biber (1988:159), often present information "in relation to the attitudes, opinions, or statements of specific individuals." We suggest that the aesthetic function of *suo* is employed to reinforce the speaker's or writer's affective tones.

Summarizing, *suo* serves ideational, (non-)contextual, personal and aesthetic functions. These functions could be multiple roles simultaneously played by *suo* and interacting with one another. Thus, it is sometimes not easy to tease out which function of *suo* is at play in a particular case. We tend to think that all the four functions of *suo* are closely related to its explicit form. Explicitness of the form makes it an elaboration feature, presenting informational purposes and also helps convey highly explicit, context-independent, endophoric references. Explicitness of the form, in addition, adds emphasis in the clause, thus achieving personal functions. Due to close association with Classical Chinese style, this explicit form also gives 'elegance' to the language, serving an aesthetic function. Although this remark is fairly tentative, we believe that this line of reasoning examining a close relation between a linguistic form and its communicative function is on the right tract and worth further pursuing.

6. Concluding remarks

A major goal of this paper is to argue against a single dimension view and support a multi-dimensional approach to register variation. By examining *suo*'s distribution in several written and spoken registers, we show the inadequacy of previous analyses characterizing *suo* solely as conveying emphasis or as associated with written or formal registers. Rather, it is the situational characteristics of a register, written or spoken, that determine the appropriateness of *suo*'s occurrence in the register. To illustrate, fiction, though a written register, is not quite compatible with *suo* because *suo*'s use is highly independent from the context but fiction is a register showing a high dependence on the context, thus resulting in the low frequency of *suo* in fiction. This gains support from the fact that spontaneous dialogues, sharing the situational characteristics of being exophoric with fiction, also exhibit low frequency of *suo*. Therefore, in terms of conveying exophoric sense, fiction and spontaneous dialogues are grouped together and need to be distinguished from other registers not conveying exophoric sense. In this approach, textual relations are defined by the situational characteristics shared among written and spoken registers (see Biber 1986, 1988 and subsequent works).

In addition to the ideational, (non-)contextual, personal and aesthetic functions of *suo* identified in §5, processing functions, one of the seven types of functions of linguistic features classified by Biber (1988), is worth further exploring. As pointed out by Hsu (2006), the reaction times are numerically higher in clauses not containing *suo* than in their counterparts with *suo*. This indicates that the presence of *suo* provides information that helps with the parsing of relative clauses to avoid potential garden path effects. In addition to issues regarding such processing functions, equally interesting are issues concerning the use of *suo* for fulfilling prosodic requirements in instances such as (7), (9c) and (17a) and the use of *suo* from social-linguistic perspectives such as social status, age and gender of the speakers. All these issues are beyond the scope of this paper and will be left to future studies.

Jen Ting
Department of English
National Taiwan Normal University
162, Sec. 1, Heping E. Road
Taipei 106, Taiwan
ting@ntnu.edu.tw

Linguistic Patterns in Spontaneous Speech, 77-100
2009-1-025-004-000266-1

Voice Quality Dependent Speech Recognition[*]

Tae-Jin Yoon[1], Xiaodan Zhuang[2], Jennifer Cole[2], and
Mark Hasegawa-Johnson[2]

University of Victoria[1]
University of Illinois at Urbana-Champaign[2]

Voice quality conveys both linguistic and paralinguistic information, and can be distinguished by acoustic source characteristics. We label objective voice quality categories based on the spectral and temporal structure of speech sounds, specifically the harmonic structure (H1-H2) and the mean autocorrelation ratio of each phone. Results from a classification experiment using a Support Vector Machine (SVM) classifier show that allophones that differ from each other regarding voice quality can be classified as distinct using input features in speech recognition. Among different possible ways to incorporate voice quality information in speech recognition, we demonstrate that by explicitly modeling voice quality variance in the acoustic phone models using hidden Markov modeling, we can improve word recognition accuracy.

Key words: ASR, voice quality, H1-H2, autocorrelation ratio, SVM, HMM

1. Introduction

The acoustic source of speech sounds, especially the source of voiced speech sounds, is defined as the airflow through the glottis. Quasi-periodic vibration of the vocal folds results in a volume velocity waveform. The source signal is modulated in the vocal tract, which functions as a resonator or a filter (Fant 1960). The term "voice quality" refers to the quality of sound produced with a particular setting of the vocal folds, and includes breathy, creaky and modal voices. Voice quality provides information at multiple levels of linguistic organization, and manifests itself through acoustic cues including F0, and information in spectral and temporal structures. If we can reliably extract acoustic features that differentiate phones on the basis of voice quality, then voice quality differences can be modeled in an Automatic Speech Recognition system (ASR), improving recognition performance.

[*] This work was supported by NSF (IIS-0414117). Statements in this paper reflect the opinion an d conclusions of the authors, and are not necessarily endorsed by the NSF.

Fundamental frequency (F0) and harmonic structure are acoustic parameters that signal voice quality. Particularly, they are shown to be important factors in encoding lexical contrast and allophonic variation related to laryngeal features (Maddieson & Hess 1987, Gordon & Ladefoged 2001). For example, Maddieson & Hess (1987) observe significantly higher F0 for tense vowels in languages that distinguish three phonation types (tense, lax, and modal) with varying voice quality (Jingpho, Lahu and Yi). However, F0 is not always a reliable indicator of voice quality. Studies of English have failed to show a strong correlation between any glottal parameters and F0 (Epstein 2002). On the other hand, information obtained from harmonic structure has been shown to be more reliable for the discrimination of non-modal from modal phonation. For example, Gordon & Ladefoged (2001) describe the characteristics of creaky phonation as producing non-periodic glottal pulses, lower power, lower spectral slope, and low F0. Among these acoustic features, they report that spectral slope is the most important feature for discrimination among different phonation types.

The observation of voice quality differences, even non-phonemic differences as in English, raises a research question: Will the incorporation of voice quality into a speech recognition system result in improved performance? We hypothesize that the spectral characteristics of phones produced with creaky voice are so different from those produced with modal voice that direct modeling of voice quality will result in improved word recognition accuracy. We test that hypothesis in the present study by labeling the voice quality of spontaneous connected speech using both harmonic structure (a spectral measure) and mean autocorrelation ratio (a temporal measure), which have been identified to be reliable indicators of voice quality.

Speech is usually parameterized as perceptual linear prediction (PLP) coefficients in speech recognition systems, to reflect human auditory characteristics. An important question is whether these parameters used in ASR also reflect voice quality variation. We answer this question by showing that the phone-level voice quality labels automatically generated according to a spectral measure taken from harmonic structure and a temporal measure of mean autocorrelation ratio are predictive of their PLP coefficients. We further show that a PLP-coefficients-based automatic speech recognizer that incorporates voice quality information in the acoustic models performs better than a complexity-matched baseline system that does not consider the voice quality distinction.

The paper is organized as follows. Section 2 illustrates linguistic and paralinguistic functions of voice quality (§2.1) and presents acoustic cues for the voice quality identification (§2.2). Section 3 introduces our method of voice quality decision on the corpus of telephone conversation speech. Section 4 reports a classification result that shows the voice quality distinctions are reflected in PLP coefficients. Section 5 presents an HMM-based speech recognition system that incorporates voice quality knowledge.

Section 6 compares the performance of the voice quality dependent recognizer against a baseline system that doesn’t distinguish different voice qualities. Section 7 concludes the paper with discussion of the source of the ASR improvement in the increased precision of the phone models that are specified for different voice qualities.

2. Voice quality

Among numerous types of voice quality (e.g., see Gerratt & Kreiman 2001), the most frequently utilized cross-linguistically are modal, creaky, and breathy voices. In this section, we briefly illustrate the characteristic of voice qualities, and present uses and functions of voice quality (§2.1) and acoustic correlates of the types of voice quality (§2.2).

Ladefoged (1971) suggests that types of voice quality, or phonation types, be defined in terms of the aperture between the arytenoid cartilages in the larynx. The arytenoid cartilages are a pair of small three-sided components in the larynx. The vocal folds are attached to these cartilages. The degree of aperture between the arytenoid cartilages, hence between the vocal folds, plays a role in producing voice qualities such as modal, breathy, and creaky voices. Modal voice, as is illustrated in Fig. (1a), refers to the phonation of speech sounds produced with regular vibrations of the vocal folds. The modal voice has relatively well-defined pitch pulses. In Fig. (1a), relatively well defined striations in the formants are visible in the region where the vowel [oi] in the word ‘voice’ is uttered. Breathy phonation, as is shown in Fig. (1b), is characterized by vocal cords that are fairly abducted (relative to modal and creaky voice) and have little longitudinal tension. The abduction and lesser tension allow some turbulence of airflow to flow through the glottis. In Fig. (1b), turbulent noise is present across the frequency range. Creaky phonation, as in Fig. (1c), is typically associated with vocal folds that are tightly adducted but open enough along a portion of their length to allow for voicing. Due to the tight adduction, the creaky voice typically reveals slow and irregular vocal pulses in the spectrogram, as in Fig. (1c), where the vocal pulses are farther apart from each other compared to those of modal and breathy voices in Figs. (1a-b).[1]

[1] The sound files are taken from http://www.ims.uni-tuttgart.de/phonetik/EGG/

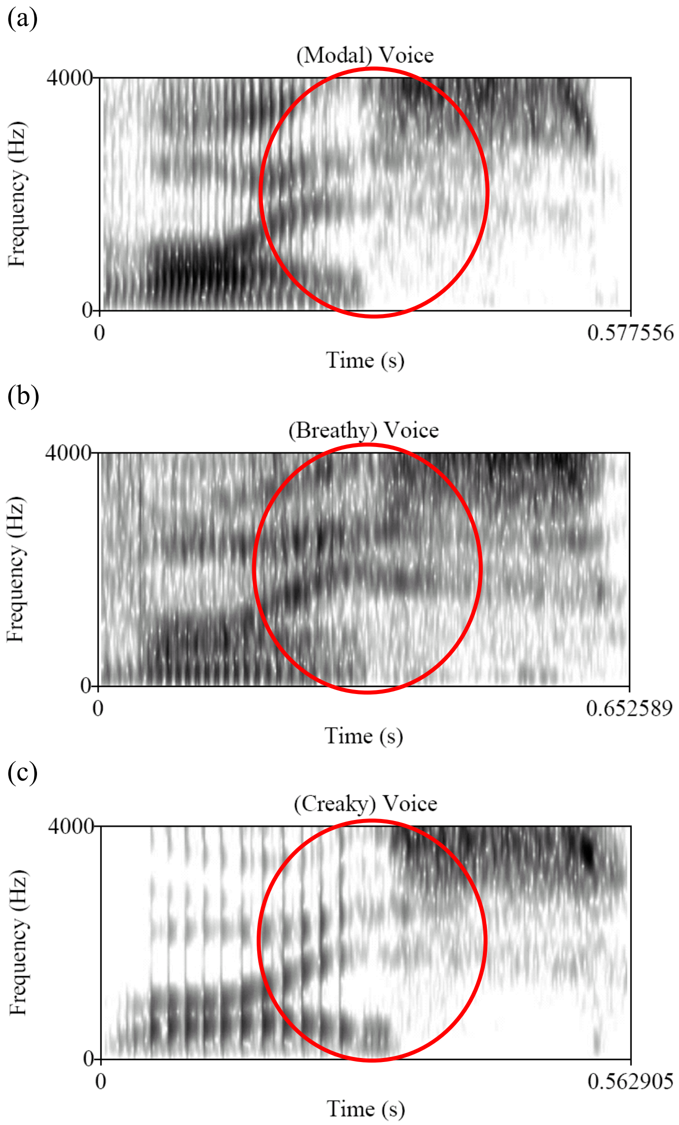


Figure 1: Spectrograms of the same word "voice" that are produced with different phonation qualities. From top to bottom, the word "voice" is produced with (a) modal voice, (b) breathy voice, and (c) creaky voice, respectively. The circles in the above figures indicate regions where different types of voice quality are observed.

2.1 Functions of voice quality

Functions of voice quality include the encoding of lexical contrasts, encoding of allophonic variation, signaling of speaker's emotional or attitudinal status, and socio-linguistic or extra-linguistic indices. The utilization of the voice quality function is language-dependent.

The use of voice quality to encode lexical contrasts is fairly common in Southeast Asian, South African and Native American Languages. For example, the presence or absence of creakiness on the vowel *a* in *já* signals difference in meaning in Jalapa Mazatec such that *já̰* produced with creakiness means "he carries" whereas *já* produced without creakiness means "tree" (Ladefoged & Maddieson 1997, Gordon & Ladefoged 2001). Gujarati speakers need breathy voice or murmured voice to distinguish the word /ba̤r/ produced with murmured voice "outside" from the word /bar/ "twelve" (Fischer-Jørgensen 1967, Bickley 1982, Gordon & Ladefoged 2001).[2]

Voice quality is also commonly used to encode allophonic variation in certain contexts. That is, many languages use non-modal phonation in creaky or breathy voice as variants of modal voice in certain contexts. For example, voiceless stop /t/ in American English is often realized as glottal stop [ʔ]. The spectrogram in Fig. 2 illustrates that the final /t/ in the word "cat" is produced with glottal stop [ʔ], with anticipatory non-modal phonation on the preceding vowel.

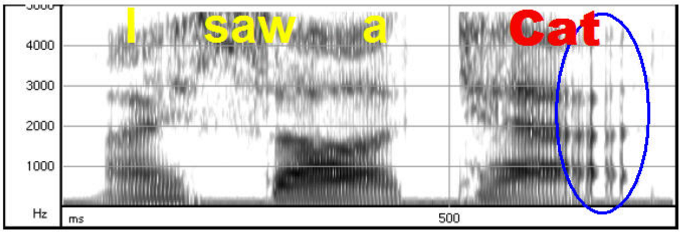


Figure 2: An allophonic realization of the voiceless stop /t/ as a glottal stop [ʔ] (Figure taken from Epstein 2002:2).

A particular voice quality is more likely to be associated with specific tones in tonal languages. Huffman (1987) observes that one of the seven tones in Hmong (a Sino-Tibetan language) is more likely to occur with a breathy voice quality. Cao & Maddieson (1989) describe that the yang tone in the Wu dialect of Chinese differs from the yin tone in that the yang tone is associated with the breathy voice.

[2] For names of languages with different types of phonation contrasts, see Gordon & Ladefoged (2001).

Voice quality can function as a marker for juncture. For example, creaky voice can be used to mark syllable, word, phrase, and utterance boundaries. Kushan & Slifka (2006) report that 5% of their 1331 hand-labeled irregular tokens in a subset of TIMIT database occur at syllable boundaries, and 78% of the tokens at word boundaries. For example, creakiness is observed at the end of a word boundary in Fig. 3.

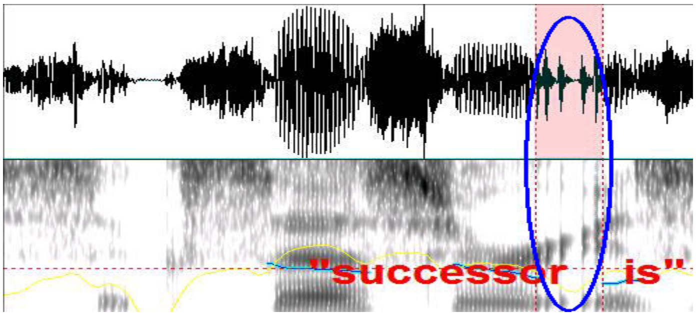


Figure 3: An example of the occurrence of creakiness at a word boundary. Creakiness is used in the realization of the rhotic "r" at the end of the word "successor."

Fant & Kruckenberg (1989) demonstrate that creaky voice is used as a phrase boundary marker for speakers of Swedish. Laver (1980) states that creaky voice with a concomitant low falling intonation may used by speakers of English as a marker for turn taking. Dilley et al. (1996) show, through the analysis of a prosodically labeled speech corpus of American English, that phrasal boundaries of intermediate and intonational phrases influence glottalization of word-initial vowels. Redi & Shattuck-Hufnagel (2001) further demonstrate that glottalization is more likely to be observed on words at the ends of utterances than on words at the ends of utterance-medial intonational phrases, and that the glottalization is more likely to be observed on boundaries of full intonational phrases than on boundaries of intermediate phrases.

In addition to the linguistically determined variation discussed above, there are paralinguistic functions in the use of voice qualities. Modulation of voice quality can be used to convey the speaker's emotion and attitude to the listener. For example, creaky voice signals tiredness or boredom, at least in American English. It should be noted that the use of voice quality and its relation to emotional or attitudinal aspects do not seem to be universal. For many speakers of Swedish, creaky voice is an affectively unmarked quality, whereas the same voice quality is used in Tzeltal (a Mayan language) to express commiseration or complaint (Gobl 2003) and it is use in Slovene to express indecisiveness or uncertainty.[3] In addition, breathy voice is associated with intimacy in

[3] http://www2.ku.edu/~slavic/sj-sls/jurgec_eng.pdf

many languages. The affect of intimacy is typically regarded to be a marker for female speakers rather than a marker for male speakers. For example, Gobl (2003) states that "gender-dependent differences, particularly increased breathiness for female speakers, have been observed in languages," including English.

Finally, it has been observed that voice quality may also have a sociolinguistic dimension serving to differentiate among social groups. Within a particular dialect, voice quality features may signal social subgroups. Esling (1978, quoted in Gobl 2003) states that "in Edinburgh English, a greater incidence of creaky voice is associated with a higher social status, whereas whispery and harsh qualities are linked to a lower social status."

Among the categories of voice quality, creaky voice has been recurrently reported to play a role in American English in signaling linguistic information, even though the function of creakiness in American English is not phonemic. Creakiness in American English is related to prosodic structure as a frequent correlate of word, syntactic, or prosodic boundaries (Kushan & Slifka 2006, Dilley et al. 1996, Redi & Shattuck-Hufnagel 2001, Epstein 2002). Given the linguistic function of creakiness in American English, it is possible to use voice quality to facilitate automatic speech recognition. Information about voice quality can be used to decide between candidate analyses of an utterance by favoring analyses in which the syntactic and higher-level structures are consistent with the observed voice quality of a target word. In this way, voice quality constitutes a new channel of information to guide phrase-level analysis. An even more basic benefit of voice quality information is also possible: Voice quality effects condition substantial variation in the acoustic realization of a word or phone. Modeling that variation offers the possibility of improved accuracy in word or phone recognition. The next section details a method for reliably detecting creaky voice quality based on acoustic cues, independent of higher-level linguistic context, for the purpose of modeling creaky voice for speech recognition.

2.2 Acoustic correlates of voice quality

Acoustic cues obtained from voice source analysis have been identified to be more reliable for voice quality identification than F0 or intensity alone. But analytic studies have largely focused on the more measurable parameters of F0 and intensity (cf. Gordon & Ladefoged 2001, Gobl 2003). This can be attributed to the methodological difficulties in voice source analysis with features other than F0 or intensity. For example, segments with both breathy and creaky voices have been shown to have reduced intensity characteristics. In certain languages such as Chong (Thongkum 1987) and Hupa (Gordon 1996), it has been observed that phones produced with creakiness trigger a reduction in intensity relative to the intensity observed in phones produced with modal phonation. However, the intensity measurement is subject to many external factors such as location of the microphone and background noise, and internal factors such as the speaker's loudness level. Slow and irregular vibration of the vocal folds characterizes creaky voice, resulting in low F0. However, F0 is not always a reliable indicator of voice quality. Studies of English have failed to show a strong correlation between any glottal parameters and F0 (Epstein 2002).

Information obtained from spectral structure is more reliable for the voice quality identification. Ní Chasaide & Gobl (1997) characterize creaky phonation as having slow and irregular glottal pulses in addition to low F0. Specifically, they state that significant spectral cues to creaky phonation are (i) A1 (i.e., amplitude of the strongest harmonic of the first formant) much higher than H1 (i.e., amplitude of the first harmonic),[4] and (ii) H2 (i.e., amplitude of the second harmonic) higher than H1.[5] (See Fig. 4 for an illustration.) Fischer-Jørgensen (1967) conducted a discrimination experiment between modal vowels and breathy vowels with Gujarati listeners using naturally produced Gujarati stimuli. The listeners were able to distinguish breathy vowels from modal ones in cases where the amplitude of the first harmonic dominates the spectral envelope. She observed that other cues such as F0 and duration had little importance in the task. Pierrehumbert (1989) investigated the interaction of prosodically prominent events such as pitch accents and voice source variables. In general, the glottal pulse for high toned pitch accents has a greater open quotient than for low toned pitch accents. The open quotient (OQ) is defined as the ratio of the time in which the vocal folds are open to the total length of the glottal cycle. But it is also occasionally observed that while higher voice level as measured by intensity results in a higher F0, the higher voice level corresponds to a reduced OQ. This implies again that the F0 and cues from voice source are largely independent of each other, and the open quotient, which is related to the harmonic structures of H1 and H2, provides a more reliable cue for the identification of non-modal phonation.

[4] Relative contribution of the A1 to the creaky voice is related to the increased bandwidth of the first formant. Hanson et al. (2001) states that "if the first-formant bandwidth (B1) increases, the amplitude A1 of the first-formant peak in the spectrum is expected to decrease. Therefore, the relative amplitude of the first harmonic and the first-formant peak (H1-A1, in dB) is selected as an indicator of B1." Thus, the relative difference between H1-A1 is relevant to the discrimination between creaky voice and modal voice.

[5] Some researchers use H1 and H2 to refer to individual harmonics, not to the amplitudes of thereof. In this paper, H1 and H2 refer to the amplitude of each harmonic, i.e., first and second harmonics, respectively.

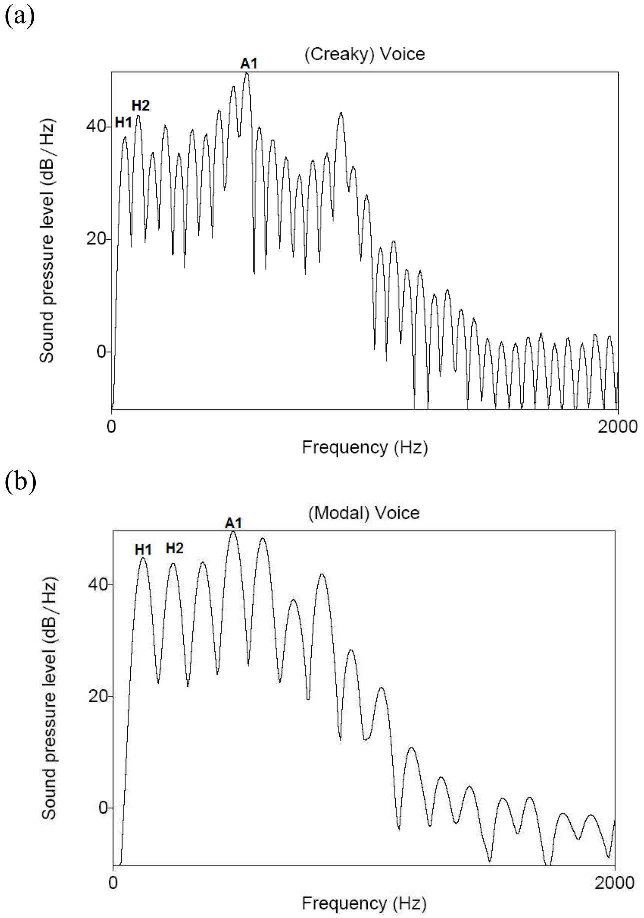


Figure 4: Spectral slices taken from the vowel ‘oi’ in the word “voice” (a) when the vowel is produced with creaky voice, and (b) when the vowel is produced with modal voice. In (a), both H2 and A1 are relatively higher than H1. In (b), H2 is approximately the same as H1, and A1 is relatively higher than H1.

H1 and H2 are related to the open quotient (OQ) (Fant 1997, Hanson & Chuang 1999, Hanson et al. 2001). The numerical relationship between H1-H2 and OQ is reported in Fant (1997) as in (1):[6]

[6] In the literature, H1*-H2* is sometimes used instead of H1-H2. H1*-H2* is a modification of H1-H2 proposed by Hanson (1997), and denotes the measure H1-H2 is corrected for the effects of the first formant (F1). See Hanson (1997) and Hanson & Chuang (1999) for the rationale and procedure of obtaining H1*-H2*.

(1) $H1–H2=–6+0.27 \exp (5.5\times OQ)$

In creaky voicing, the vocal folds are held tightly together (though often with low internal tension), resulting in a low OQ. That is, the more the amplitude of the second harmonic relative to that of the first harmonic, the lesser is OQ. In breathy voicing, the vocal folds vibrate without much contact, thus the glottis is open for a relatively longer portion of each glottal cycle, resulting in a high OQ. In modal voicing, the vocal folds are open during part of each glottal cycle, resulting in the OQ between those for the creaky voicing and for the breathy voicing.

Other relevant cues for the identification of voice quality, especially creaky voice, include aperiodicity, due to the slow and aperiodic glottal pulses in creaky phonation. A couple of measures can be used to quantify the degree of aperiodicity in the glottal source. One is “jitter”, which quantifies the variation in the duration of successive fundamental frequency cycles. Jitter values are higher during creaky phonation than other phonation types. The other is mean autocorrelation ratio. Mean autocorrelation ratio is a temporal measure that quantifies the periodicity of the glottal pulses, which is used in our experiment, as will be detailed in §3.2.

3. Voice quality decision

3.1 Corpus

Switchboard is a corpus of orthographically transcribed spontaneous telephone conversations between strangers (Godfrey et al. 1992). The corpus is designed mainly to be used in developing robust Automatic Speech Recognition. The corpus consists of more than 300 hours of recorded speech spoken by more than 500 speakers of both genders over the phone. Our analysis is based on a subset of the Switchboard files (12 hours) containing one or more utterance units (10-50 words) from each talker in the corpus. Phone transcriptions are obtained by forced alignment using the word transcription and dictionary. In general, the quality of the recorded speech, which is sampled at 8kHz, is much inferior to speech samples recorded in the phonetics laboratory. Although ITU (International Telecommunication Union) standards only require the telephone network to reproduce speech faithfully between 300Hz and 3500Hz (e.g., ITU Standard 1993), our observations indicate that most signals in Switchboard reproduce harmonics of the fundamental frequency faithfully at frequencies as low as 120Hz. This conclusion is supported by the results of Yoon et al. (2005), who demonstrated that measures of H1-H2 acquired from telephone-band speech are predictive of subjective voice quality measures at a significance level of $p<0.001$. Post-hoc analysis of Yoon et al.’s results suggests that H1-H2 is an accurate measure of glottalization for female talkers in

Switchboard, but is less accurate for male talkers, who often produce speech with F0 < 120Hz. The low quality of telephone-band speech is also known to affect pitch tracking; as noted in Taylor (2000), pitch tracking algorithms known to be reliable for laboratory-recorded speech often fail to extract an F0 during regions perceived as voiced from the Switchboard corpus.

3.2 Feature extraction and voice quality decision

As mentioned above, the Switchboard corpus has the drawback that the recordings are band-limited signals. The voice quality of creakiness is correlated with low F0, which hinders accurate extraction of harmonic structure if the F0 falls below 120Hz. This is because harmonics are any whole-number multiple of F0. To enable a voice quality decision for signals with F0 below 120Hz, we use a combination of two measures: H1-H2 (a spectral measure, occasionally corrupted by the telephone channel) and mean autocorrelation ratio (a temporal measure, relatively uncorrupted by the telephone channel) in the decision algorithm for voice quality.

We use Praat (Boersma & Weenink 2005) to extract the spectral and temporal features that serve as cues to voice quality. First, intensity normalization is applied to each wave file. Following intensity normalization, inverse LPC filtering (Markel 1972) is applied to remove effects of the vocal tract on source spectrum and waveform.

From the intensity-normalized, inverse-filtered signal, minimum F0, mean F0, and maximum F0 are derived over each file. These three values are used to set ceiling and floor thresholds for short-term autocorrelation F0 extraction, and to set a window that is dynamically sized to contain at least four glottal pulses. F0 and mean autocorrelation ratio are calculated on the intensity-normalized, inverse-filtered signal, using the autocorrelation method developed by Boersma (1993). The unbiased autocorrelation function $r_x(\tau)$ of a speech signal $x(t)$ over a window $w(t)$ is defined as in (2):

$$(2)\quad r_x(\tau) \approx \frac{\int x(t)x(t+\tau)dt}{\int w(t)w(t+\tau)dt}$$

where τ is a time lag. The mean autocorrelation ratio is obtained by the following formula (3):

$$(3)\quad \overline{r}_x = \left\langle \max_{\tau} \frac{r_x(\tau)}{r_x(0)} \right\rangle$$

where the angle brackets indicate averaging over all windowed segments, which are extracted at a timestep of 10ms. The range of the mean autocorrelation ratio is from 0 to

1, where 1 indicates a perfect match, and 0 indicates no match of the windowed signal and any shifted version. Harmonic structure is determined through spectral analysis using FFT and long term average spectrum (LTAS) analyses applied to the intensity-normalized, inverse filtered signal.

H1 and H2 are estimated by taking the maximum amplitudes of the spectrum within 60 Hz windows centered at F0 and 2×F0, respectively, as in (4):[7]

$$(4)\quad H1 - H2 = \max_{-60<\delta_1<60} 20\log_{10}|X(F_0+\delta_1)| - \max_{-60<\delta_2<60} 20\log_{10}|X(2F_0+\delta_2)|$$

where *X(f)* is the FFT spectrum at frequency *f.*

Yoon et al. (2005) previously used spectral features including H1-H2 to classify subjective voice quality with 75% accuracy. Subjective voice quality labels used in that experiment are not available for the research reported in this paper. In the current work, interactively-determined thresholds are used to divide the two-dimensional feature space [*rx*; *H*1-*H*2] into a set of voice-quality-related objective categories, as follows.

For each 10ms frame, the "voiceless" category includes all frames for which no pitch can be detected. The "creaky" category includes all frames for which *H*1-*H*2<-15dB, or for which *H*1-*H*2<0 and r_X<0.7. All other frames are labeled with an objective category label called "modal." Fig. 5 illustrates an example of objectively labeled creaky voice on the sonorant [er]. The waveform in the top tier is divided into 10ms intervals in the bottom tier. The voiceless, non-creaky, or creaky label is assigned to each 10ms frame based on the above-mentioned criteria. Within each sonorant phone, whose boundaries we obtained through forced alignment, if more frames indicate creaky category than any other category, the phone itself is assigned creaky label ("_cr"). For our experiment, we do not consider the voice quality variation for obstruents such as stops and fricatives, and only sonorants are eligible to be assigned the creaky label.

[7] Because the input speech is inverse filtered so that the effect of resonant frequencies are minimized, if not completely eliminated, we didn't apply any correction regarding formants to H1-H2, as is suggested in Hanson (1997) (cf. H1*-H2*).

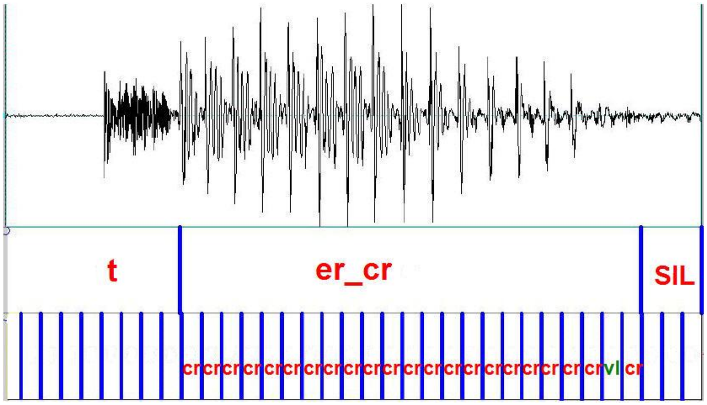


Figure 5: Example of a sonorant /er/ with objective creaky label

4. Voice quality distinction reflected in PLP coefficients

As discussed in §2, the acoustic measures we extracted (see §3) are correlated with the voice quality of creakiness. These features (i.e., H1-H2 and mean autocorrelation ratio) are not a standard input to speech recognition systems. Instead, PLP (Perceptual Linear Predictive) coefficients are usually used as standard input features. There are two ways of incorporating the features related to the voice quality into a speech recognition system: (1) appending the voice quality related features to the standard PLP coefficients, or (2) modeling phones of different voice qualities separately as allophonic variants, while not modifying standard feature vectors. In order to justify the latter approach, it is necessary first to determine whether the voice quality categories are predictive of the standard speech recognition feature vectors such as PLP. This section describes an experiment designed to determine whether or not PLP coefficients are sufficient to distinguish between creaky and non-creaky examples of any given sonorant phone.

The PLP (Perceptual Linear Predictive) cepstrum is an auditory-like cepstrum that combines the frequency-dependent smoothing of MFCC (mel-frequency cepstral coefficients) with the peak-focused smoothing of LPC (Hermansky 1990). In our work, thirty-nine PLP coefficients are extracted over a window size of 25ms with a timestep of 10ms. PLP coefficients, as shown in the second figure in Fig. 6, typically perform well for speech recognition purposes, even with noisy (low SNR) signals. In order to show that the voice quality distinction based on H1-H2 and the mean autocorrelation ratio is also reflected in the acoustic features used in speech recognition, such as PLP coefficients, this section reports the results of a validation test using SVM (Support Vector Machine) classification.

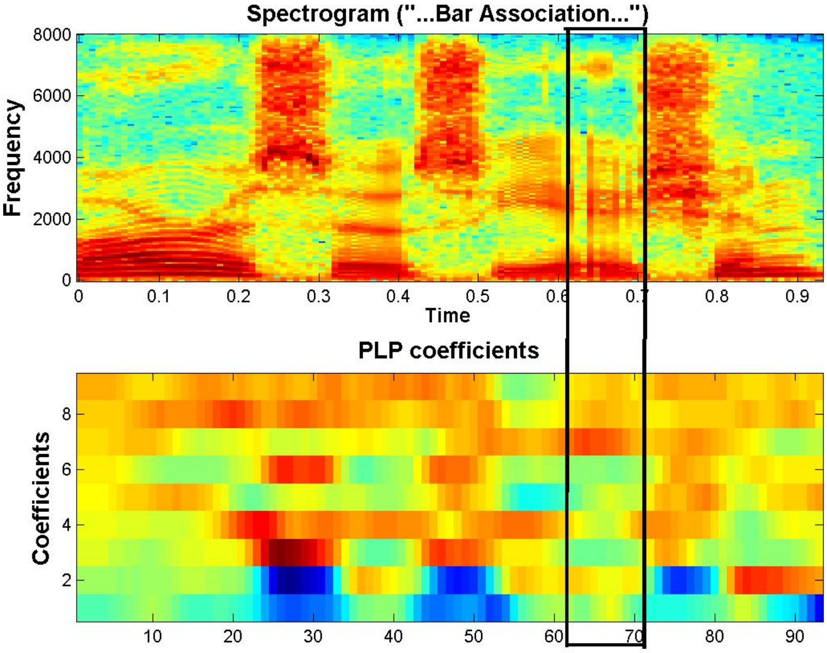


Figure 6: An example of spectrogram and graphical representation of the PLP coefficients. In the spectrogram shown in the first figure, the rectangular region between 0.6 and 0.7 in the x-axis of the upper figure indicates that the speech corresponding to [ei] in the word *Association* is produced with creakiness. This paper investigates whether the creakiness characteristic is reflected in the input feature vectors of PLP coefficients, which is graphically represented in the second figure.

SVM is a machine learning algorithm that seeks to find the optimal mapping function $y = f(x, \alpha)$, where y is an output category (e.g., either modal or creaky phones), x is an input feature vector (e.g., PLP coefficients), and α is a set of adjustable model parameters. The optimality is defined by minimizing the structural error of the classification. We use SVM with a non-linear kernel because we assume that the category boundary between model and creaky phones is nonlinear in the feature space of PLP coefficients.

We conduct an experiment to classify non-creaky phonation versus creaky phonation for each sonorant (i.e., vowel, semi-vowel, nasal or lateral). The phone-aligned transcription for each file is obtained using HTK (Young et al. 2005), and aligned against the voice quality label sequences given by the frame-level voice quality decisions described before. For each sonorant segment, if more frames indicate creakiness than the other voice qualities (i.e., modal or voiceless), the phone is labeled as creaky. We divide the 12 hour Switchboard subset into a training candidate pool (90%) and a testing candidate pool

(10%). Then for each sonorant phone from the training candidate pool, we extract a subset of the non-creaky tokens that is equal in size to the creaky tokens for the same phone, based on the creakiness label resulting from the decision scheme. These non-creaky and creaky tokens compose the training data for each sonorant. The testing data for each sonorant are similarly generated from the testing candidate pool, which also have equal numbers of creaky and non-creaky tokens and no overlap with the training data. We use the SVM toolkit LibSVM (Chang & Lin 2004) to train separate binary classifiers for each sonorant; each classifier distinguishes between creaky and non-creaky examples of the phone. Classifiers are tested using the testing data, for each sonorant separately. The classification accuracies obtained from the testing data for each sonorant are reported in Table 1.

Table 1: SVM classification of voice qualities for each phone. The first and third columns list the creaky (indicated by cr) versus non-creaky phone labels, in ARPABET notation. The second and fourth columns list the accuracy of a classifier trained to distinguish between creaky and non-creaky examples of the specified phone.

Phones		**Accuracy**	**Phones**		**Accuracy**
uh	uh_cr	74.47%	w	w cr	69.91%
dr	er_cr	73.26%	ih	ih cr	69.75%
aw	aw cr	73.26%	ow	ow cr	69.09%
eh	eh cr	71.93%	y	y cr	68.45%
ae	ae cr	71.52%	l	l_cr	68.23%
uw	uw_cr	71.42%	ao	ao_cr	68.04%
iy	iy_cr	70.51%	m	m_cr	67.79%
ey	ey_cr	70.50%	ax	ax_cr	67.24%
ay	ay_cr	70.37%	el	el_cr	66.85%
ah	ah_cr	70.14%	r	r_cr	66.36%
aa	aa_cr	70.13%	oy	oy_cr	63.24%
ng	ng_cr	70.05%	en	en_cr	58.19%
n	n_cr	70.03%			

As shown in Table 1, the PLP coefficients are correctly classified with an overall accuracy of 58% to 74% (with an average overall accuracy of 69.23%). Chance performance is 50%. An average of 19.23% improvement, relative to chance, suggests that the voice quality decision is reflected to some degree in the PLP coefficients. Based on this finding, we conclude that it should be possible to design a speech recognition system that distinguishes between creaky and non-creaky examples of each sonorant phone using only PLP coefficients as an acoustic observation.

5. Voice quality dependent speech recognition

The goal of a speech recognition system is to find the word sequence that maximizes the posterior probability of the word sequence $\mathbf{W} = (w_1, w_2, \cdots, w_M)$, given the observations $\mathbf{O} = (o_1, o_2, \cdots, o_T)$:

(5) $$\hat{W} = \arg\max_{\mathbf{W}} p(\mathbf{W} \mid \mathbf{O})$$

Using Bayes rule and the fact that $p(\mathbf{O})$ is not affected by $\mathbf{W}$,

(6) $$\begin{aligned}\hat{W} &= \arg\max_{\mathbf{W}} \frac{p(\mathbf{O} \mid \mathbf{W})p(\mathbf{W})}{p(\mathbf{O})} \\ &= \arg\max_{\mathbf{W}} p(\mathbf{O} \mid \mathbf{W})p(\mathbf{W})\end{aligned}$$

Sub-word units $\mathbf{Q} = (q_1, q_2, \cdots, q_L)$, such as phones, are usually essential to large vocabulary speech recognition, therefore we can rewrite formula (7) as:

(7) $$\begin{aligned}\hat{W} &= \arg\max_{\mathbf{W}} p(\mathbf{O} \mid \mathbf{W})p(\mathbf{W}) \\ &\approx \arg\max_{\mathbf{W}}[\max_{Q} p(\mathbf{O} \mid \mathbf{Q})p(\mathbf{Q} \mid \mathbf{W})p(\mathbf{W})]\end{aligned}$$

The general automatic speech recognition architecture is shown in Fig. 7. The post probability of each word sequence hypothesis $\mathbf{W}$ is calculated according to three components: the acoustic model $p(\mathbf{O} \mid \mathbf{Q})$, the pronunciation model $p(\mathbf{Q} \mid \mathbf{W})$ and the language model $p(\mathbf{W})$.

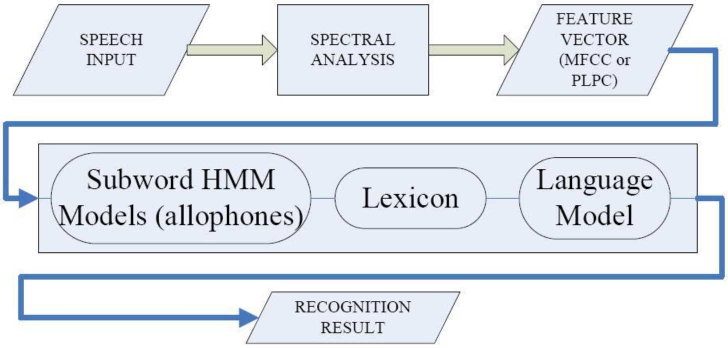


Figure 7: General automatic speech recognition architecture

In a typical speech recognition system, the observation vectors $\bar{O}$ are PLP (Perceptual Linear Predictive) coefficients or MFCC (Mel Frequency Cepstral Coefficients), plus their energy, all computed over a window size of 25ms at a time step of 10ms, and their first order and second order regression coefficients, referred to as delta and delta-delta (or acceleration) coefficients.

The acoustic model $P(\mathbf{O} \mid \mathbf{Q})$ is usually a set of left-to-right hidden Markov models (HMMs), each modeling the acoustics of a particular sub-word unit such as a phone, as in Fig. 8:

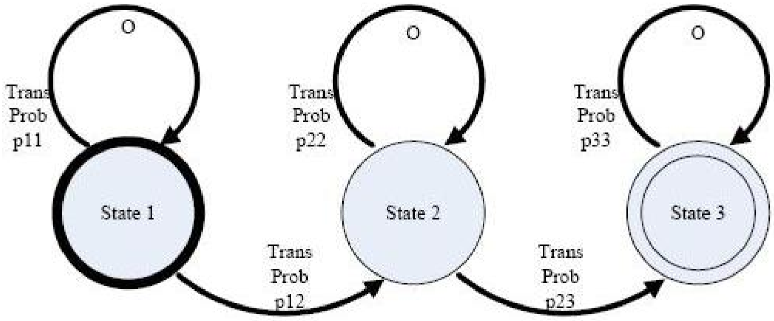


Figure 8: Left-to-right hidden Markov model

In a left-to-right HMM, state transitions occur from a state either to itself or to the following state. These state transition probabilities describe, from a probabilistic point of view, how long each part of the sub-word unit *q* should be. For each of the states, there is one Gaussian-mixture distribution describing the state-conditioned observation distributions.

The pronunciation model $p(\mathbf{Q} \mid \mathbf{W})$ typically maps a word to either phones or triphones (allophones in particular contexts). In this paper, we are using a deterministic pronunciation model, i.e. mapping each word to a fixed sequence of triphones.

The language model $p(\mathbf{W})$ is usually the *n*-gram model: the probability of a particular word in the word sequence is conditioned on the previous *n-1* words.

$$(8) \quad p(w_1 w_2 \cdots w_m) = p(w_1) \cdots p(w_{n-1}) \prod_{i=n}^{m} p(w_i \mid w_{i-n+1} \cdots w_{i-1})$$

For example, the simple bigram language model is as follows:

$$(9) \quad p(w_1 w_2 \cdots w_m) = p(w_1) \prod_{i=2}^{m} p(w_i \mid w_{i-1})$$

5.1 Baseline system

We build a triphone-clustered HMM-based speech recognition system as the baseline system using HTK (Young et al. 2005). This system uses a deterministic pronunciation model, also called dictionary, and a bigram language model, but a sophisticated acoustic model, which will be detailed in the following paragraphs.

Every phone is represented by a large number of partially independent triphone models. All triphones that represent the same base phoneme use the same transition probability matrix; we say that their transition probability matrices are "tied." The observation probability density functions associated with the first, second, or third state of any given pair of triphones may also be tied together, as shown in Fig. 9.

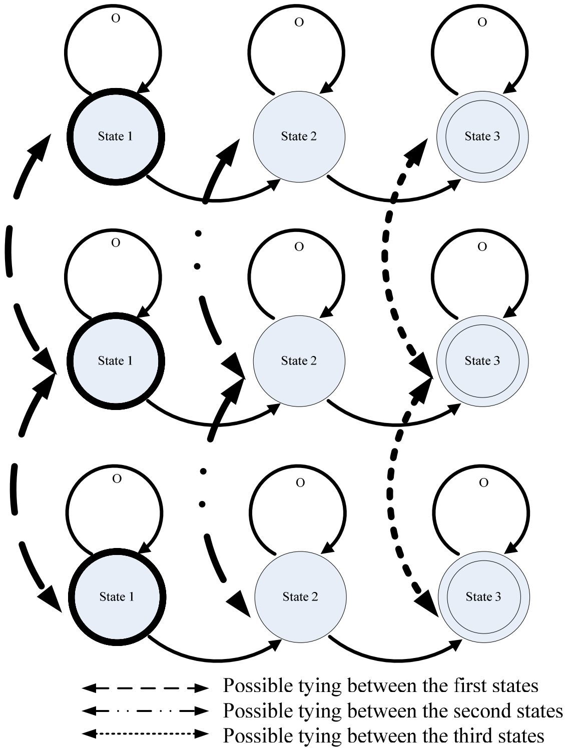

O
O
O
State 1
State 2
State 3
O
O
O
State 1
State 2
State 3
O
O
O
State 1
State 2
State 3
Possible tying between the first states
Possible tying between the second states
Possible tying between the third states

Figure 9: Tying options of counterpart states in HMMs representing allophones of the same base phoneme

Allophones of the same base phoneme are tied together in allophone sets. Each allophone set corresponds to one of the leaves in a binary tree. The phonetic binary clustering tree

(Fig. 10) begins with a root node comprising all allophones of a given base phoneme label. At each level of the tree, the allophones belonging to the next higher level are split into two categories based on a question about the phonological features of the left context phone or the right context phone. The tree is grown from root to leaf (or from top down in Fig. 10), with all corresponding states of allophones placed at the root node initially. At each non-leaf node, the splitting question is selected from a pool of binary questions to maximize the increase in the likelihood of training data given the model. In this way, phonetic contexts that induce the most allophonic variation are placed nearest to the tree root. Once the maximum likelihood increase at a particular node is smaller than a threshold, this node will not be further split and all states in that node will be tied together.

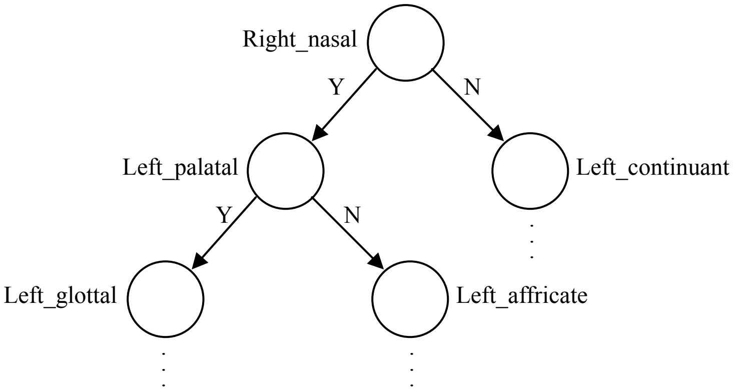


Figure 10: Binary clustering tree (an example of the near-root part of the binary clustering tree for the third emitting state of vowel /ae/)

It is necessary to deal with triphones unseen in the training data but maybe existing in testing data. These triphone models are synthesized, after the model is fully trained, by tying the states of the HMM to three particular states from seen allophones, chosen according to the unseen triphone's answer to binary questions in the clustering tree. In other words, a synthesized state is tied to all the states in a particular leaf node of the clustering tree.

After state tying is completed, the number of Gaussians in each mixture Gaussian observation distribution is repeatedly incremented, with further mean and variance estimation following each increment, thus achieving observation distributions that better reflect the characteristics of the allophones.

5.2 VQ-ASR system

The Voice Quality Automatic Speech Recognition (VQ-ASR) system incorporates into the baseline system binary voice quality information (creaky or non-creaky) for every sonorant phone.

Inclusion of Voice Quality Information: We use forced alignment to obtain phone boundaries for the phonemes specified in the canonical dictionary entry for each word listed in the Switchboard word transcription. This phone-aligned transcription is aligned against the voice quality label sequences given by the frame-level voice quality decisions described in §3.2. For a vowel, semi-vowel or nasal, if more frames indicate creakiness than the other voice qualities (i.e., modal or voiceless), a "creakiness label" is attached to this phonation (See Fig. 5).

Given these creakiness-labeled phone transcriptions and corresponding wave files, we use the Baum-Welch algorithm to do an embedded estimation of all the allophone HMMs involved in these transcriptions. For every training utterance, the HMMs corresponding to phones present in that utterance are concatenated according to the transcription, and estimated together instead of separately. Thus, we can get one HMM for each allophone, defined on its own phone identity and its context, both in terms of phonetics and voice quality. The creakiness of a phone is modeled as part of the phone's context, rather than being part of the base phoneme label, thus creaky and non-creaky versions of the same phoneme are eligible to be clustered together by the triphone clustering algorithm exemplified in Fig. 10. Fig. 11 illustrates how voice quality knowledge is incorporated in the training transcription.

Word transcription:	SAY	YOU	DID
Phone transcription: Creakiness labels:	s ey sp	y uw sp cr	d ih d sp cr
Creakiness-labeled phone transcription:	s ey sp	y uw_cr sp	d ih_cr d sp
Allophone transcription	s+ey s-ey sp y+uw_cr y-uw_cr sp d+ih_cr d-ih_cr+d ih_cr-d sp		
Allophone Transcription (phonetic / voice quality context)	s+ey s-ey sp y+uw_cr y_cr-uw sp d+ih_cr d_cr-ih+d_cr ih_cr-d sp		

Figure 11: Conversion from the word transcription to the transcription of allophones defined on phone identify and phonetic/voice quality context ("_cr" represents the "creakiness label".)

Recognition Dictionary with Voice Quality Information: To perform speech recognition using voice quality information, we need to map the voice quality dependent allophone sequences to word sequences. While we wish to take advantage of explicit acoustic modeling of voice quality variation, such variation does not impact word identity (in English). Therefore, we need a new dictionary containing all possible pronunciations of the same word, with all of the different possible voice quality settings. For example, for “bat b+ae b-ae+t ae-t” in the baseline system dictionary, as in Fig. (12a), the dictionary in a VQ-ASR system should have two entries “bat b+ae b-ae+t ae-t” and “bat b+ae_cr b_cr-ae+t_cr ae_cr-t”, as in Fig. (12c).

Word:	bat		
Phones:	b	ae	t
(a) Triphones:	b+ae	b-ae+t	ae-t
(b) Triphones:	b+ae	b-ae+t	ae-t
(with VQ Info)	b+ae_cr	b-ae_cr+t	ae_cr-t
(c) Triphones:	b+ae	b-ae+t	ae-t
(VQ context)	b+ae_cr	b_cr-ae+t_cr	ae_cr-t

Figure 12: Recognition dictionary with voice quality information (example: the word “bat”)

Reduction of the Number of Parameters: The number of triphones increases dramatically, as the creakiness label can be attached to one or both of the neighboring phones for each triphone. To reduce the number of parameters, we include allophones with different phonetic/voice quality context in the same binary decision tree in the triphone clustering process (Fig. 13). By tying transition matrices of all allophones, tying states of some allophones using a tree-based clustering technique, and synthesizing unseen triphones in the same way as the baseline system, we build the VQ-ASR system with an almost identical number of parameters to that in the baseline system, despite the increase in the number of triphones. This is necessary, because any increase in the number of model parameters will have a tendency to improve recognition performance, which would make the comparison between the VQ-ASR system and the baseline system less accurate.

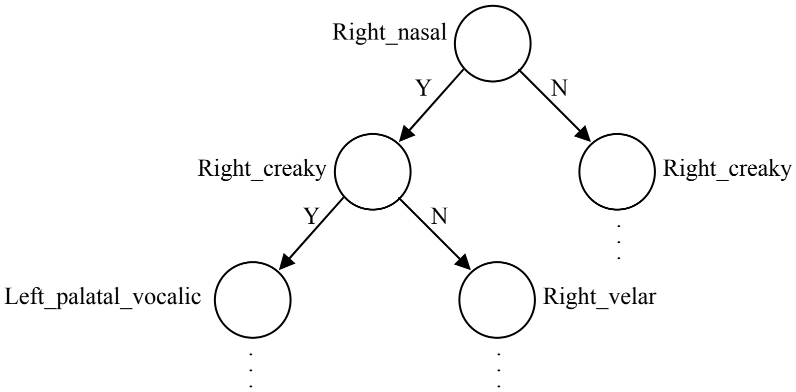


Figure 13: Binary clustering tree showing the effect of creakiness. (An example of the near root part of the binary clustering tree for the third emitting state of vowel /ae/, showing that creakiness context is more salient than most phonetic context.)

6. Experimental results

Word recognition accuracies of the voice quality dependent and voice quality independent speech recognition systems are shown in Table 2. In our experiment, both systems are prototype ASR systems, trained and tested on the 12 hour subset of Switchboard.[8] The comparison of the results in Table 2 is made under the condition of (i) tied transition probabilities for all allophones and (ii) an almost identical number of states for both systems. This allows for a stringent comparison between systems with a nearly equal number of parameters.

Table 2: Word recognition accuracy for the voice quality dependent and voice quality independent recognizers. The number of Gaussians in each Gaussian mixture is given in the first column. %Correctness is equal to the percentage of the reference labels that were correctly recognized. %Accuracy is a more comprehensive measure of recognizer quality that penalizes insertion errors.

Mixture	Baseline		VQ-ASR	
	%Correctness	%Accuracy	%Correctness	%Accuracy
3	45.81	39.28	46.42	39.35
9	52.77	45.31	52.77	46.01
19	52.88	46.82	55.41	48.63

[8] The two systems are designed to identify the impact of voice quality dependence, therefore not comparable to systems trained on much larger amounts of data (e.g., Luo & Jelinek 1999, Sundaram et al. 2000).

Two evaluation metrics are used: %Correctness and %Accuracy, defined as

$$\begin{aligned} &\%\text{Correctness}=\frac{N-D-S}{N}\times 100 \\ (10)\quad & \\ &\%\text{Accuracy}=\frac{N-D-S-I}{N}\times 100 \end{aligned}$$

where *N* is the number of tokens (i.e. words) in the reference transcriptions that have been reserved as a test dataset for the evaluation purpose, *D*, the number of deletion errors, *S*, the number of substitution errors, and *I*, the number of insertion errors. The %Correctness penalizes deletion errors and substitution errors deviating from the reference transcriptions; %Accuracy also penalizes insertion errors. Word error rate (WER), another widely used evaluation metric, is equal to 100-%Accuracy.

As seen in Table 2, when voice quality information is incorporated in the speech recognition system, the percentage of words correctly recognized by the system increases by approximately 0.86% on average and the word accuracy increases by approximately 1.05% on average. It is worth noting that as the number of Gaussians per mixture increases to 19, the improvement in the percentage of words correctly recognized increases to 2.53%, and the improvement in the word accuracy increases to 1.81%.

7. Discussion and conclusion

In this paper, we have shown that a voice quality decision based on H1-H2 as a measure of harmonic structure, and the mean autocorrelation ratio as a measure of temporal periodicity, provides useful allophonic information to an automatic speech recognizer. Such voice quality information can be effectively incorporated into an HMM-based automatic speech recognition system, resulting in improved word recognition accuracy.

As the number of Gaussian components per state of the HMM increases, the VQ-ASR system surpasses the baseline system by an increasingly greater extent. Given that the number of untied states and the number of transition probabilities in the HMMs in both systems are identical, it follows that the VQ-ASR system benefits more from an increasingly precise observation PDF (probability density function), compared to the baseline system. Although we don’t know why added mixtures might help the VQ-ASR more than the baseline, we speculate that there must be an interaction between the phonetic information provided by voice quality labels, and the phonetic information provided by triphone context. Perhaps the acoustic region represented by each VQ-ASR allophone is fully mapped out by a precise observation PDF to an extent not possible with standard triphones.

Similar word recognition accuracy improvements have been shown for allophone models dependent on prosodic context (Borys 2003). Glottalization has been shown to be correlated with prosodic context (e.g., Redi & Shattuck-Hufnagel 2001), thus there is reason to believe that an ASR trained to be sensitive to both glottalization and prosodic context may have super-additive word recognition accuracy improvements.

Tae-Jin Yoon
Department of Linguistics
University of Victoria
PO Box 3045
Victoria, BC V8W 3P4, Canada
tyoon@uvic.ca

Xiaodan Zhuang
xzhuang2@uiuc.edu

Jennifer Cole
jscole@uiuc.edu

Mark Hasegawa-Johnson
jhasegaw@uiuc.edu

Linguistic Patterns in Spontaneous Speech, 101-128
2009-1-025-005-000267-1

Prosodic Hierarchy as an Organizing Framework for the Sources of Context in Phone-Based and Articulatory-Feature-Based Speech Recognition*

Mark Hasegawa-Johnson[1], Jennifer Cole[1], Ken Chen[2], Partha Lal[3], Amit Juneja[4], Tae-Jin Yoon[5], Sarah Borys[1], and Xiaodan Zhuang[1]

University of Illinois at Urbana-Champaign[1]
Washington University[2]
University of Edinburgh[3]
Think-a-Move, Ltd.[4]
University of Victoria[5]

Large-vocabulary automatic speech recognition (ASR) models each word as either (1) a sequence of context-dependent allophones called "phones," or more rarely, (2) a matrix of articulatory units, called "features," in autosegmental tiers. Details of phone or feature inventory vary from system to system, but the requirements are easy to define: each phone (or each vector of articulatory features) must be both "acoustically compact" (the acoustic correlates of a phone or feature vector are predictable) and "phonologically compact" (the phone or feature correlates of a word, in context, are predictable). This paper proposes that the two goals of a phone inventory may be satisfied (1) by defining phones that are sensitive to prosodic context, or (2) by using prosodic context to constrain the temporal evolution of features. Five different example systems are described. Three phone-based systems include phone inventories that are sensitive to prosodic phrase context (phrase boundaries and phrasal prominence), foot-level context (lexical stress and vowel reduction), and disfluency context (filled pauses and interruption points). Two feature-based systems include feature synchrony constraints sensitive to word-level context (within-word vs. between-word featural asynchrony) and syllable-level context (consonant release vs. consonant closure acoustic observations). We report results showing that many of these systems have reduced word error rate (WER) of an ASR in at least one controlled experiment.[1] Computational complexity

* This material is based upon work supported by the National Science Foundation (NSF) under grants 04-14117 (Hasegawa-Johnson, Cole, and Shih) and 01-21285 (Jelinek and Khudanpur). Any opinions, findings, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

[1] All of the experimental results described in this article have been previously published in technical reports or conference papers, but only the results of §2 have been previously published in professional journals; a more extensive description of one of the results of §6 is also currently

limitations, and training data limitations, have thus far prevented the integration of all proposed context features into any single ASR application.

Key words: automatic speech recognition, prosody, disfluency, phones, articulatory features

1. Introduction

This paper proposes using the prosodic hierarchy as an organizing framework for the sources of phonetic context information in both phone-based and articulatory-feature-based ASR. The goal of this introductory section is to adequately define the terms in the preceding sentence, and to give some of the reasons why we believe it to be a promising paradigm for ASR research.

An automatic speech recognizer is a search algorithm governed by a probability mass function (PMF). The PMF is an estimate of the probability, $P(W \mid X)$, that a talker has produced the word sequence $W = [w_1, \ldots, w_L]$ given that the acoustic signal has short-time spectra $X = [\vec{x}_1, \ldots, \vec{x}_T]$. The goal of the search algorithm is to find the W that maximizes $P(W \mid X)$:

$$(1) \quad \hat{W} = \arg\max_{W} P(W \mid X)$$

Researchers studying the "search problem" try to find an algorithm that maximizes $P(W \mid X)$ as fast as possible; researchers studying the "training problem" try to find a function $P(W \mid X)$ that is as accurate as possible. Because the field is specialized in this way, the accuracy of a speech recognizer is determined by the accuracy of its PMF model. The goal of accurate speech recognition is therefore equivalent to the goal of finding a function $P(W \mid X)$ such that, in all cases, the correct words (the words the talker actually said) are also the ones that maximize $P(W \mid X)$.

For computational reasons, Eq. 1 is usually rewritten as

$$(2) \quad \hat{W} = \arg\max_{W} \left(\frac{P(W)p(X \mid W)}{p(X)} \right) = \arg\max_{W} P(W)p(X \mid W)$$

The *language model* PMF $P(W)$ and the *acoustic model* probability density function (PDF) $p(X \mid W)$ are complicated functions with millions of trainable parameters. The acoustic model, $p(X \mid W)$, is parameterized by two fundamentally different types of parameters:

under review. References to relevant technical reports and on-line documentation are provided in each section.

mode parameters and *mixture parameters*. Mode parameters represent the mean and variance of an acoustic mode (a set of similar acoustic spectra that occur in similar linguistic contexts; a mode is usually modeled using a Gaussian distribution, therefore the mean and variance of the mode are sufficient statistics). Mixture parameters represent the different ways in which acoustic modes can be combined to form any given word sequence. There are two different types of mixture parameters: "mixture weights" specify the probability of disjunctive mode selection at a specified time, while "transition probabilities" specify the probability of any given mixture sequence.

Most words are infrequent, therefore it is impractical to learn the mode parameters and mixture parameters of every word directly from training data. Instead, most large-vocabulary speech recognizers simplify the mixture problem by defining a finite countable set of context-dependent, segmental units, intermediate between the word and the acoustic signal, called "phones." A well-designed phone set has the following properties:

- ACOUSTICALLY COMPACT: The phone label predicts the acoustic spectrum. In other words, given a phone label q_m at time t, the distribution $p(\vec{x}_t \mid q_m)$ of acoustic spectra has low entropy.
- PHONOLOGICALLY COMPACT: The word sequence predicts the phone sequence. In other words, given a word sequence $W = [w_1, \ldots, w_L]$, the distribution $P(Q \mid W)$ of possible phone sequences $Q = [q_1, \ldots, q_M]$ has low entropy.

It is not easy to define a set of phone labels that is both acoustically and phonologically compact. Orthographically identical phones may be acoustically disparate, e.g., there are acoustically important differences between the ten different productions of /t/ typical of the words "top," "tree," "stop," "steep," "felt," "bat," "bats," "batman," "butter," and "button" (Zue & Laferriere 1979). Pronunciation depends on long-term context: an intonational-phrase-initial phone is different from an intonational-phrase-final phone, and a phone with phrasal prominence is different from a phone without phrasal prominence (Cole et al. 2007). The relevant acoustic context is the entire utterance: prosodic phrases at the end of a prosodic group are shorter than prosodic-group-medial phrases (Tseng et al. 2005).

In order to be acoustically compact, the phones used by an ASR must be context-dependent. The number of relevant contexts is quite large: it is not unusual for a typical ASR to use a phone inventory with tens of thousands of distinct phones. Each phone model represents a certain set of training examples: in order to specify the exact contexts in which those training examples occur, we need to define some notation. Standard notation makes a distinction between monophones, context-dependent phones (CD-phones), and states.

A “state” is an index into the table of parameterized acoustic probability distributions: given a unique state variable number or name, q, the recognizer is able to look up the parameters of the acoustic PDF $p(\vec{x} \mid q)$ in a parameter table. Most typically, the PDF $p(\vec{x} \mid q)$ is a diagonal covariance mixture Gaussian function (Juang et al. 1986). A mixture Gaussian PDF represents a D-dimensional acoustic feature vector, $\vec{x} = [x_1, \ldots, x_D]^T$, using a linear combination of K different Gaussian modes:

$$(3) \quad p(\vec{x} \mid q) = \sum_{k=1}^{K} c_{kq} \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi\sigma_{dkq}^2}} e^{-\frac{(x_d - \mu_{dkq})^2}{2\sigma_{dkq}^2}}$$

where the number of modes, K, and the dimension of the acoustic feature vector, D, are specified by the system designer, and all other parameters including the mixture weights c_{kq} and the mode parameters μ_{dkq} and σ_{dkq} are automatically learned from training data. Most ASR systems break each CD-phone into three temporally sequential states, so that the first state models the CD-phone onset and the third state models its offset (Jelinek 1976).

Each CD-phone is a context-dependent variant of exactly one monophone. There are typically 48 monophones in English (Lee & Hon 1989). The monophones correspond approximately, but not precisely, to phonemes. Non-phonemic monophones are created in order to represent unusually common and stable surface forms such as schwa (/AX/) and flap (/DX/). In this paper, monophones are expressed using two forms of notation: IPA notation (e.g., /noteʃən/) and two-letter ARPABET notation (e.g., /N OW T EY SH AX N/); to reduce confusion, the latter is written in capital Roman letters.[2] The contextual constraints acting on a CD-phone are specified using three delimiters: a preceding – denotes left context, a following + denotes right context, and a following – denotes long-term context. For example, if the code US means “unstressed syllable,” then the CD-phone /AY-F+OW_US/ is a statistical model that has been trained to represent examples of the monophone /f/ occuring immediately after /ɑʲ/, immediately before /o/, in an unstressed syllable. States are specified by augmenting the CD-phone label with a number, e.g., /AY-F+OW_US2/ is the second state of the CD-phone /AY-F+OW_US/, and $p(\vec{x} \mid \text{AY-F+OW\_US}\,2)$ is the corresponding parameterized distribution of acoustic feature vectors.

As suggested by the notation, left context and right context are special, because they are used more universally than other types of context. A CD-phone dependent only on local left and right context (no long-term context) is called an n-phone (e.g., triphone,

[2] For complete definitions of the ARPABET monophone inventory see, for instance, Parsons (1987), Lee & Hon (1989), Young et al. (2002), Hasegawa-Johnson (2005).

quinphone, or septphone; (Lee & Hon 1989)). For example, the triphone AY-F+OW represents an /f/ produced at the center of the 3-phone sequence /ɑʲfo/, as in the word "triphone." If there are N monophones, then the number of possible n-phones is N^n. No reasonably-sized training corpus contains enough data to robustly train N^3 triphone models, therefore left-context phones and right-context phones with similar effects on the center phone are typically clustered together using a binary classification tree learned from training data (Odell et al. 1994).

Standard phone notation, as introduced in the preceding paragraphs, suggests that the acoustic PDF $p(\vec{x} \mid q)$ is best indexed by some combination of the monophone label together with a series of context specifiers. Articulatory phonology (Browman & Goldstein 1992) provides a quite different way of thinking about the effects of context on the acoustic correlates of a word. In articulatory phonology, a word is not stored in memory as a sequence of phones; instead, a word is stored as a partially sequenced set of intended articulatory gestures. The partial sequencing of gestures has been modeled as a graph of violable and sometimes conflicting alignment targets, mediated by a control algorithm (Nam & Saltzman 2003); an alternative prior approach models phonology as a graph of pairwise temporal precedence relations between the onsets and/or offsets of articulatory states (Carson-Berndsen 1998). Temporal overlap between competing gestures may block the perfect implementation of either gesture, leading to phonological assimilation or reduction. There is evidence to suggest that assimilation and reduction are planned rather than passive processes (e.g., Gomi & Kawato (1996) demonstrate that locality assimilation in manual reaching movements is centrally planned), therefore articulatory phonology posits a continuous-valued mental representation called the "tract variable;" assimilation and reduction are planned during the mental transformation from gestures into tract variables (Saltzman & Munhall 1989).

Fig. 1 provides a schematic of the way in which overlap among conflicting gestures may cause phonological reduction and assimilation: in this case, reduction of the word "everybody" to the casual form "eruwai" (ɛrʊwɑʲ) attested in a conversational telephone speech database (Livescu 2005). The left half of Fig. 1 represents the phonological-to-articulatory planning process in the mind of a talker during production of the word "everybody" in citation form. Three intended lip gestures are shown for the word "everybody:" the CRitical gesture that defines the /v/, the CLosed gesture that defines the /b/, and the PRotruded gesture that defines the /ɔ/. Browman & Goldstein (1992) suggest that a gesture is specified in the lexical entry for any word if (and perhaps only if) it is necessary to distinguish that word from another word. Fig. 1 generalizes their claim slightly: we assume that the /ɔ/ inserts a LIP-PR gesture into this talker's lexical entry for the word "everybody," because /ɔ/ without the LIP-PR gesture would be /ʌ/, and despite the lack of any English word that would be confused with "everybody" if the

LIP-PR gesture were omitted. During all other phones, Byrd & Saltzman (2003) suggest that the phonological-to-articulatory transform is driven by the influence of a "default gesture;" Fig. 1 assumes that the default gesture for the lips, during speech, is rather more open than closed. The "lip opening" tract variable smoothly interpolates between the target positions of the specified gestures; the articulatory-to-muscle-control transformation then determines muscle commands, to the several muscles of the lips, necessary to generate the desired labial aperture. Similar processes generate motor commands to the tongue, jaw, soft palate, larynx, and lungs.

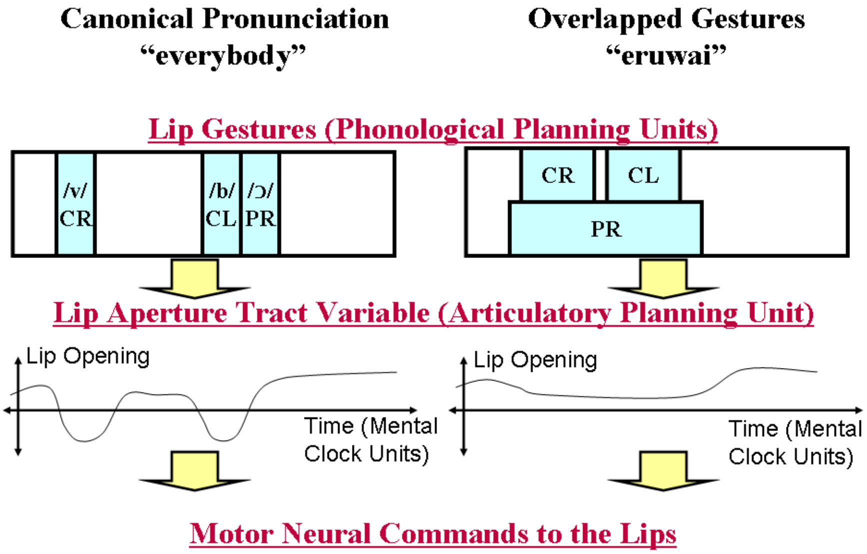

Canonical Pronunciation
"everybody"
Overlapped Gestures
"eruwai"
Lip Gestures (Phonological Planning Units)
/v/
CR
/b/
CL
/ɔ/
PR
CR
CL
PR
Lip Aperture Tract Variable (Articulatory Planning Unit)
Lip Opening
Time (Mental Clock Units)
Lip Opening
Time (Mental Clock Units)
Motor Neural Commands to the Lips

Figure 1: Overview of articulatory phonology. During rapid or casual speech, intended lip gestures are allowed to overlap in time (likewise tongue gestures, glottal gestures, et cetera). The phonological-to-articulatory transformation determines a target labial aperture ("tract variable") that serves as a compromise among the conflicting gestures. The articulatory-to-muscle-command transformation then generates motor unit commands that will achieve the target labial aperture. Lip gesture types shown in the figure are CLosed, CRitical (fricative), and PRotruded.

The right half of Fig. 1 shows the speech planning process during rapid or casual speech; the lip gestures for the phones of "everybody" have been allowed to overlap.[3] Conflicting gestures specifying that the lips should be simultaneously CLosed and

[3] For a comparable example of overlap among gestures within the same tract variable, see, e.g., Fig. 9 of Byrd & Saltzman (2003).

PRotruded cannot be simultaneously satisfied, therefore the phonological-to-articulatory transformation works out a compromise: the lips will be narrow but open.

In a standard ASR, a "phone" is defined to be a monophone, modified by context specifications. In an ASR based on articulatory phonology, on the other hand, a "phone" may be defined as a vector of simultaneously active gestures and tract variable settings. In articulatory phonology-based ASRs described in this paper, the vector of gestures and tract variables is never collapsed down to a single state variable. Instead, as proposed by Livescu & Glass (2004), the ASR state variable, q, is replaced by a vector, $\vec{q}=[q_L, q_T, q_G]^T$ of quasi-independent articulatory features (AFs) representing the lips, tongue, and glottis/velum. Each articulatory feature is intended to be a summary of all gestures currently acting upon a particular named articulator. The vector of all currently active AFs serves as an index into a table of parameterized acoustic PDFs, $p(\vec{x}\,|\,\vec{q})$.

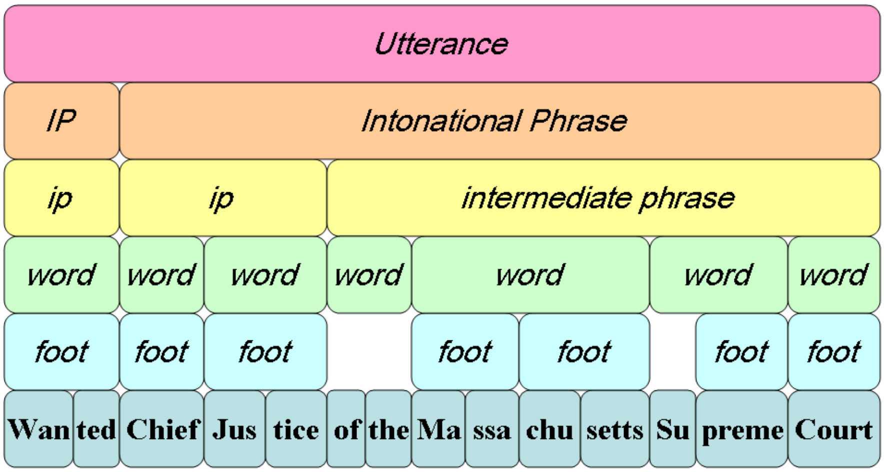


Figure 2: An example prosodic hierarchy with six levels, exemplified using a sentence from the Boston University Radio Speech Corpus (Ostendorf et al. 1995)

This paper demonstrates in §§3 and 5 that an AF vector is a useful replacement for the hidden state variable in ASR. An AF vector, however, has no explicit context specification: unlike the CD-phone label, the AF vector is not explicitly modified by triphone or prosodic context labels. It remains to be specified how we may represent prosodic and triphone context in an AF-based ASR.

Selkirk (1981) proposes that any given phonological or phonetic sound pattern (that establishes a dependency between sounds or restricts the occurrence of a sound) must be defined in terms of relationships among the units at a specified level of the *prosodic hierarchy* (Fig. 2). Each level of the hierarchy is the relevant context for a particular set of phonetic and phonological sound patterns (processes and constraints). Some of the processes and constraints that have been proposed to operate at each level of the hierarchy include:

- Sound patterns bounded within the **Utterance** include the generation of turn-taking cues. It has been hypothesized that end-of-turn is cued by word choice, and also by modulation of some of the same acoustic features that are used to signal other types of prosodic juncture, e.g., pause, duration, and pitch (Local et al. 1986, Ferrer et al. 2002, Gorman et al. 2007).
- Sound patterns bounded within the **Intonational Phrase** include boundary tones that mark the distinction between sentence types (e.g., question vs. statement), and possibly the specification of information structure distinctions such as theme vs. rheme (Steedman 2000).
- Sound patterns bounded within the **Intermediate Phrase** include the assignment of phrasal stress and pitch accent, the downstepping of pitch accents in "list intonation" (Beckman & Elam 1994, Yoon 2007), the re-setting of pitch register, and the temporal regularization that defines rhythm (Kim 2006).
- Sound patterns bounded within the **Prosodic Word** include the deletion or insertion of phones through processes related to syllabification, and many types of phonological feature assimilation.
- Sound patterns bounded within the **Foot** include the location of lexical stress, reduction and under-shoot of both vowels and consonants, flapping, and possibly rhythmic adjustments in the direction of isochrony (Kim 2006).
- Sound patterns bounded within the **Syllable** include the acoustic signaling of the phone itself. Stop consonants, for example, may be signaled by a consonant-vowel transition, a vowel-consonant transition, both, or neither.
- Disfluency may cause interruption and reset of any contiguous set of levels. For example, interruption of the word causes a word fragment; interruption of the intonational phrase causes pitch reset (Ostendorf et al. 2001, Cole et al. 2005).

This paper proposes using the prosodic hierarchy as an organizing framework for the sources of phonetic context information in both phone-based and articulatory-feature-based ASR. Specifically, this paper proposes two distinct methods for the explicit representation of prosodic context in ASR: one method that is appropriate for phone-based systems, and a quite different method that is appropriate for articulatory-feature-based systems.

Prosodic context may be incorporated into phone-based ASR by the use of long-term context specifications. Symbolically, the proposed scheme represents each phone as a vector of categorical features:

$$(4) \qquad \text{phone} = \begin{bmatrix} \text{monophone label} \\ \text{syllable context features} \\ \text{foot context features} \\ \text{word context features} \\ \text{prosodic phrase context features} \\ \text{utterance context features} \\ \text{disfluency context features} \end{bmatrix}$$

In implementation, the vector representation shown in Eq. 4 is collapsed into a single, rather long, CD-phone label. There are two ways to control the complexity of the resulting ASR. First, the differences among contextual variants of any given monophone may be constrained on the basis of phonetic knowledge, as described in §2. Second, the set of all CD variants of any given monophone may be clustered using the methods of (Odell et al. 1994). The methods for incorporating prosody into CD-phone-based ASR are reasonably well understood, and have been the subject of several published articles. Section 4 reviews the work of Bates & Ostendorf (2002) and Bates et al. (2007), who use the methods of Eq. 4 to incorporate word-level, foot-level, and syllable-level context into the phone definition. Sections 2 and 6 of this article describe our own previous work in this area, in which the methods of Eq. 4 are used to incorporate prosodic phrase context and disfluency context into the phone definition. In particular, §2 demonstrates that the use of a prosody-dependent acoustic model is not very effective unless combined with an explicit representation of prosody in the language model.

This paper proposes that prosodic context may be integrated into AF-based ASR by constraining the allowable sequences of articulatory features. The use of prosodic constraints in AF-based ASR is much less fully developed than the use of prosody-dependent phone models, but in general, it has the following properties. First, prosodic constituents (utterances, phrases, words, feet, or syllables) are specified symbolically in the language model and in the dictionary. Second, each prosodic constituent boundary imposes certain constraints on the articulatory feature trajectories as they cross the boundary. Section 3 describes, for example, a system in which all articulatory features must resynchronize at every word boundary, meaning that all articulators must simultaneously transition from one word to the next. We believe that this constraint is, in fact, too strict (examples of cross-word coarticulation are quite common in the literature (Beckman 1989) and in our ASR training data (Greenberg et al. 1996)), but results in the phonology literature suggest that some less strict type of synchronization constraint is active at the word boundary and/or at the boundaries of intermediate or intonational phrases. Similarly, §5 describes a system in which a syllable is implicitly modeled as a sequence of phonetic landmarks: an optional consonant release, a required syllable nucleus, and an optional consonant closure. A transition of the articulatory features from a closed vocal tract state to an open vocal tract state, and back again, necessarily generates a sequence of three landmarks; the likelihood of that particular articulatory feature trajectory is then evaluated using a set of classifiers (support vector machines) trained to detect release, nucleus, and closure landmarks in the acoustic signal. Articulatory feature systems have not yet been designed to incorporate prosodic phrase context, utterance context, or disfluency context; §7 very briefly sketches methods that may be effective, in the future, for the incorporation of phrase-level prosodic context into AF-based ASR.

The phone-based and AF-based approaches described in this paper are intended to be complementary rather than contradictory. Hickok & Poeppel (2000, 2007) have recently argued, based on an extensive review of the neurophysiological literature, that the robustness of human speech perception is supported by the existence of at least three parallel neural pathways, any one of which is capable of independently accessing the mental lexicon. They demonstrate that the dorsal pathway is responsible for the transformation of acoustic percepts into signals that touch upon the articulatory motor pathway; they argue that signals in this path may then access the lexicon by way of articulation. The right ventral pathway, they argue, is capable of accessing the lexicon using only prosodic cues, e.g., syllable count and stress pattern, though the right ventral pathway can also make use of phone-level cues if available. Finally, the left ventral pathway accesses the mental lexicon with few steps, if any, intervening between sound patterns and stored word forms; it is to this pathway that Hickok and Poeppel attribute most of the classical results concerning phonological neighborhood effects on lexical access. Parallel computation is effective in ASR, too, and has been proven to be useful in a large number of recent papers (e.g., Fiscus 1997, Schwenk & Gauvain 2000, Stolcke et al. 2000, Hain et al. 2000, Fiscus et al. 2000). Section 3 of this paper demonstrates a non-significant tendency for the lowest WER to be achieved by a system that combines the parallel outputs of a phone-based and an articulatory-feature based ASR.

2. Intonational and intermediate phrases

Intonational phrase (IP) boundaries are signalled by at least three types of cues: increased duration of phones in the rhyme of the phrase-final syllable (Wightman et al. 1992), a characteristic F0 movement called a boundary tone (Pierrehumbert 1980), and increased glottalization of phrase-initial phones (Dilley et al. 1996). Within an intermediate phrase (ip), typically at least one word receives phrasal prominence: the stressed syllable of that word will be produced with greater vocal effort than prosodically unmarked phones of the same type, resulting in greater intensity and with increased duration that extends throughout the stress foot (Turk & Sawusch 1997). There is often also a characteristic F0 movement associated with the accented syllable. This section considers, in particular, the use of increased phone duration as a cue for the detection of IP boundaries, and the use of F0 for the detection of phrasal prominence. Research in this area demonstrates that models of IP and ip context can reduce the word error rate (WER) of a speech recognizer.

By using speech data with manually transcribed intonational phrase boundaries and pitch accents, it is possible to train an automatic speech recognizer in which the prosodic context variable π_t for each phone takes one of four possible values: intonational phrase final vs. nonfinal, prominent vs. nonprominent (Chen et al. 2006). A phone in this system is defined to be phrase-final if it occurs in the rhyme of the syllable ending an intonational phrase, and nonfinal otherwise. A phone is defined to be prominent if it occurs in the lexically stressed syllable of a word marked as prominent in the prosodic phrase (i.e., marked with phrasal stress), and nonprominent otherwise. The prominent and nonprominent versions of each phone are allowed to differ only in the probability density function of an auxiliary normalized smoothed F0 observation, y_t; thus the joint probability density of the spectral envelope $\vec{x}_t$ and pitch y_t can be factored as $p(\vec{x}_t, y_t \mid c_t, \pi_t) = p(\vec{x}_t \mid c_t) p(y_t \mid c_t, \pi_t)$, where c_t is the triphone label. The spectral observation PDFs $p(\vec{x}_t, y_t \mid c_t, \pi_t)$ of the phrase-final and nonfinal versions of each triphone are not allowed to differ; only the model of phone duration is allowed to differ depending on intonational phrase position.

Table 1 shows WER of five different ASRs trained and tested using the Boston University Radio Speech Corpus (Ostendorf et al. 1995). The Radio Speech Corpus is a database of stories read, on the air and in the laboratory, by seven professional radio announcers. About 3.5 hours of speech have been prosodically transcribed using the ToBI (tones and break indices) prosodic transcription system (Silverman et al. 1992, Beckman & Hirschberg 1994). A baseline ASR trained using 90% of the prosodicaly transcribed portion of the Radio Speech Corpus, and tested using the other 10%, achieved WER of 24.8%, shown in the first row of Table 1. By incorporating prosody-dependent acoustic models, WER was reduced to 24.0%.

The relationship among syntax, prosody, and the word string is modeled in our system by a prosody-dependent bigram language model. A prosody-dependent bigram is an estimate of $p(w_m, p_m \mid w_{m-1}, p_{m-1})$. The prosodic label p_m carries two types of information: the phrasal prominence/nonprominence of word w_m, and the position of w_m within an intonational phrase. There are eight possible settings of p_m: a word may be prominent or nonprominent; the same word may be phrase-initial, phrase-final, phrase-medial, or it may be a one-word intonational phrase (both phrase-initial and phrase-final). The sequence $[p_{m-1}, p_m]$ takes on $|P|^2 = 64$ possible values, so in theory, a prosody-dependent bigram model learns 64 times as many parameters as a prosody-independent bigram model. In practice, most possible combinations of w_m and p_m never occur, so their probabilities are estimated by backing off to 1-gram and 0-gram (uniform) distributions; in our experiments, the actual parameter count of a prosody-dependent bigram model is slightly less than three times that of a prosody-independent bigram. A system using both prosody-dependent acoustic model and prosody-dependent language model, shown in the fourth row of Table 1, achieved WER of 23.4%—a significant reduction of word error rate in comparison to the baseline.

Table 1: Word error rate (WER), prominence error rate (PER), and intonational phrase boundary error rate (BER, in percent) with five different combinations of acoustic model and language model. Chance performance is 44.6% PER, 15.6% BER.

Acoustic Model	Language Model	WER	PER	BER
Prosody Independent	Prosody Independent	24.8	44.6	15.6
Prosody Dependent	Prosody Independent	24.0	45.9	15.0
Prosody Independent	PD Bigram	24.3	23.1	14.5
Prosody Dependent	PD Bigram	23.4	20.3	14.3
Prosody Dependent	PD Semi-factored	21.7	20.3	14.2

An empirically superior estimate of the prosody-dependent bigram probability may be trained by explicitly modeling the relationship between the prosodic tag, p_m, and the syntactic tag, s_m (Chen & Hasegawa-Johnson 2003). The syntactic tagset used in our first-pass ASR specifies the part of speech of word w_m. By explicitly modeling syntactic tags, the prosody-dependent bigram probability may be written as

$$(5)\ p(w_m, p_m \mid w_{m-1}, p_{m-1}) = \sum_{s_m, s_{m-1}} p(w_m, p_m, s_m, s_{m-1} \mid w_{m-1}, p_{m-1})$$

$$(6)\qquad \approx \sum_{s_m, s_{m-1}} p(p_m \mid s_m, s_{m-1}, p_{m-1}) p(s_m, s_{m-1} \mid w_m, w_{m-1}) p(w_m \mid w_{m-1}, p_{m-1})$$

The approximation in Eq. 6 is valid if we assume that, first, prosody is independent of the word string given knowledge of syntax, and second, that the syntactic tags are independent of prosody given knowledge of the word string. The first term on the right-hand side of Eq. 6, $p(p_m \mid s_m, s_{m-1}, p_{m-1})$, may be robustly estimated from a relatively small corpus, because the syntactic tagset and the prosodic tagset are both much smaller than the vocabulary. The second term, $p(s_m, s_{m-1} \mid w_m, w_{m-1})$, is the probability that a word sequence (w_{m-1}, w_m) implements syntactic tag sequence (s_{m-1}, s_m) ; in our experiments we assumed this mapping to be deterministic. The third term in Eq. 6, $p(w_m \mid w_{m-1}, p_{m-1})$, is a prosody-dependent semi-bigram probability, and is estimated directly from the Radio Speech Corpus, using backed-off maximum likelihood estimation. A system using Eq. 6 to represent the language model achieved our lowest WER to date on the Boston University Radio Speech Corpus—21.7%.

3. The word

Chomsky & Halle (1968) proposed that the domain of any given phonological process is bounded, with the domains of successive processes gradually expanding through a process of bracket erasure. In particular, they proposed that lexical stress assignment, phonotactics, and syllabification are determined within the boundaries of a lexical word. Selkirk (1981) noted, however, that resyllabification often occurs across word boundaries, and proposed the "phonological word" to be the domain of syllabification. A phonological word is most often coterminous with a lexical word in English, but is quite frequently longer than a lexical word in Japanese (e.g., Iwano & Hirose 1999) and Chinese (e.g., Huang & Lee 2006), and is occasionally shorter than a lexical word in Spanish (Peperkamp 1999). An example of a prosodic word composed of two lexical words is shown in Fig. 2, where the words "of the" have merged into a single prosodic word, allowing deletion of the final /v/ in "of," resulting in the open-syllable sequence /ə.ðə/. As in this example, resyllabification across a lexical word boundary may require phone deletion or substitution in order to avoid violating the phonotactic rules of the language. It is possible that, because of these resyllabification effects, phone deletion or substition effects across lexical word boundaries are more common within than between *prosodic* words, but we do not know of any published studies testing this hypothesis. Several published studies have proposed that cross-word coarticulatory effects are more common within than between prosodic phrases (Beckman 1989, Beckman & Elam 1994, Nakatani & Hirschberg 1994).

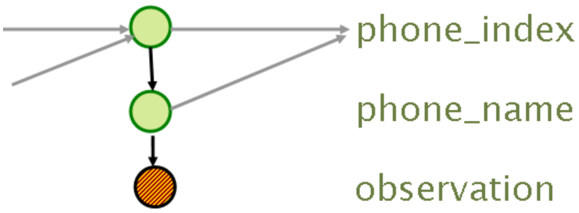


Figure 3: Dynamic Bayesian network (DBN) representation of a standard phone-based HMM speech recognizer.

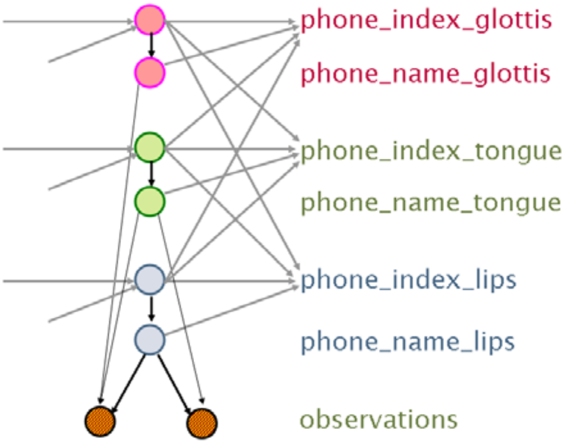


Figure 4: Dynamic Bayesian network (DBN) representation of a recognizer with three hidden state variables, separately representing the states of the lips, tongue, and glottis/ velum.

Coarticulation and assimilation are modeled, in most modern ASR systems, by means of the n-phone abstraction (e.g., triphone or quinphone). An n-phone is a phoneme-length segment (a consonant or vowel), but the n-phone label depends on the phonological features of n consecutive segments: for example, the triphone AY-F+OW represents an /f/ produced at the center of the 3-phone sequence /ɑ^{j}fo/, as in the word "triphone" (Lee & Hon 1989). In order to model the possibility that word boundaries may block coarticulation, many systems block the formation of triphones across a word boundary: for example, the /f/ in "my phone" may be represented by the biphone label F+OW instead of the triphone label AY-F+OW. Almost all modern systems use either cross-word triphones (in which triphone context extends across all word boundaries) or word-internal triphones (in which triphone context extends only within a lexical word), but Huang & Lee (2006) demonstrated that WER can be reduced by allowing cross-word triphones only when two lexical words are part of a single prosodic word.

Articulatory phonology has inspired a large number of recent ASR experiments (Richardson et al. 2000, Richmond et al. 2003, Livescu & Glass 2004). The model of Livescu & Glass (2004), for example, factors the "phone" into a set of three to eight parallel "articulatory features" (AF), modeled as the hidden variables in a dynamic Bayesian network (DBN). For computational reasons, all published articulatory-phonology based ASR systems (including the system of (Livescu & Glass 2004), and the system described in this section) prohibit cross-word coarticulation. Asynchrony among the different articulators is allowed during a word. At a word boundary, however, every articulator is required to simultaneously change state. For example, at a hypothesized boundary between the words "two" and "three" with no intervening silence, the tongue closure and the glottal devoicing movement would be required to occur simultaneously; for computational reasons, the ASR would not be allowed to consider the hypothesis that tongue closure and glottal devoicing occur asynchronously. The prohibition of cross-word coarticulation in these systems has been implemented as a way of controlling computational complexity, and it is certainly too strict to represent real speech phenomena (examples of cross-word coarticulation are quite common in the literature (Beckman 1989) and in our ASR training data (Greenberg et al. 1996)), but the work of Selkirk suggests that some kind of (looser) synchronization constraint may be appropriate at word boundaries, while the work of others (e.g., Beckman 1989) suggests that prohibition of coarticulation across prosodic phrase boundaries would be appropriate.

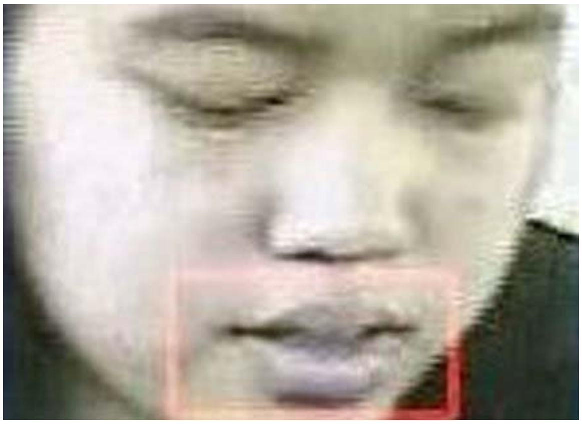


Figure 5: Asynchrony between audio and visual cues. The talker is preparing to begin saying the word "three;" there is not yet any audio signal. The tongue tip has been closed in preparation for the phoneme /θ/, and the lips have been rounded in preparation for the /r/.

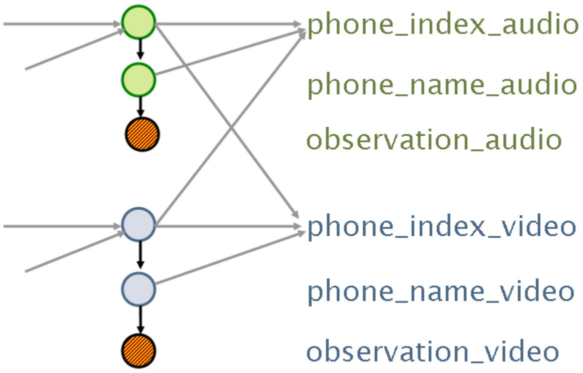


Figure 6: Coupled hidden Markov model (CHMM) designed to model asynchronies between the phone labels indicated by audio and visual speech observations. Each chain (audio and video) progresses through the same sequence of phones for any given word, but the two chains may progress at different rates.

The systems described in this section are based on the system of (Livescu & Glass 2004); all of these systems are implemented in GMTK (Zweig et al. 2002) using the notation of a dynamic Bayesian network or DBN. Fig. 3 shows a DBN representation of a standard hidden Markov model (HMM); Fig. 4 shows a DBN representation of a recognizer inspired by gestural phonology, with three different, conditionally independent articulators (the lips, tongue, and glottis/velum). The standard speech recognizer keeps track of two very different types of information about the phones at each time step: the **phone_index** specifies how far through the current word the talker has progressed, while the **phone_name** specifies which phone is actually being produced (which vowel or consonant it is). The **observation** (a perceptual LPC vector (Hermansky 1990)) is dependent on the value of the **phone_name**. In the models proposed by Livescu & Glass (2004), the **phone_name** is replaced by a set of three parallel labels: one label specifies the current state of the lips (wide, protruded, narrow, dental, closed, or silent), one label specifies the current state of the tongue (low back, high back, low front, high front, retroflex, palatal, palatal fricative, etc.), and the third label specifies the current state of the glottis and soft palate (unvoiced, voiced non-nasal, voiced nasal). The **observations** depend on the current settings of all three articulators.

Table 2: Word error rate, connected digit recognition, CUAVE development test data. Statistically significant differences are marked by a double line separating rows in the table.

System	WER
CHMM, up to 1 state of asynchrony allowed	22.8
Articulatory Feature system, 2 states asynchrony allowed	22.1
CHMM, up to 2 states of asynchrony allowed	21.8
ROVER system combination, three CHMM systems	20.1
ROVER system combination, two CHMM systems and one AF system	19.4

It has long been recognized that the visual signal may convey evidence of inter-articulator asynchrony that is not obvious in the acoustic signal. Fig. 5, for example, shows a sample frame from the silence preceding the word “three:” although the acoustic signal is still silent, two of the three phones in the upcoming word are already visible in the talker’s lips and tongue. It has been demonstrated many times that WER of an audiovisual speech recognizer may be reduced by explicitly modeling the asynchrony between audio and visual cues (e.g., Chu & Huang 2000, Neti et al. 2000, Zhang 2000). Asynchrony between audio and visual cues has most successfully been represented by the use of parallel HMMs: a “phoneme” model that generates audio feature observations, and a “viseme” model that generates video feature observations. One structure for managing the asynchrony between phoneme and viseme is the coupled HMM (CHMM) (Chu & Huang 2000). As shown in Fig. 6, a CHMM is a DBN with two parallel sets of phone labels: a **phone_name_audio** representing the phone that is audible in the acoustic signal, and a **phone_name_video** representing the phone that is visible in the image sequence. These two phone labels may fall out of synchrony if, for example, the video images clearly show the tongue producing a /θ/, but the audio signal clearly contains only silence, as shown in Fig. 5.

In July 2006, we developed (Livescu et al. 2007) an audiovisual speech recognition system based on the gestural phonology model of Livescu and Glass. The system that we developed is shown in Fig. 4. That system was compared to the performance of the CHMM shown in Fig. 6, on the task of connected digit recognition from audiovisual recordings.[4] Training and test data were drawn from the CUAVE corpus (Patterson et al. 2002). Audio features included PLP coefficients, energy, deltas, and delta-deltas. Video features included the 35 lowest-order coefficients from a discrete cosine transform of the

[4] In standard ASR technical descriptions, “connected digits” are digits spoken with a silent pause after each word. Digits spoken with no pause between words are called “continuous.” Connected speech recognition is generally considered to be easier than continuous speech recognition, but harder than isolated word recognition.

grayscale pixel values in a rectangle including the lips, and their deltas. Systems were trained using 60% of the available noise-free data. The number of Gaussians per mixture was increased until performance peaked on noise-free development test data (20% of the available data). Video and audio stream weights were then chosen in order to minimize WER on noisy development test data at six different SNRs (noise-free, 12dB, 10dB, 6dB, 4dB, and -4dB SNR), and the resulting WERs are reported in Table 2.

Results are shown in Table 2. The only statistically significant differences in this table are the difference between 20.1% WER and 21.8% WER, and the difference between 22.1% and 22.8% WER; all smaller differences are non-significant on this dataset. Trends shown in the table must be interpreted with caution, because they are not statistically significant, and because they have been obtained using development test data; confirmation of these results using independent evaluation test data was not completed. The trends shown in the table suggest, with the caveats already provided, that it would be worthwhile to pursue definitive support for the following conclusions. First, the CHMM seems to perform best when it is allowed to consider asynchrony between the states: as shown, allowing the audio and video phones to be asynchronous by two states (2/3 of a phone) is better than allowing only one state of asynchrony (1/3 of a phone). Similar results were achieved for the articulatory feature system. Second, it doesn’t seem to matter very much exactly how the asynchrony is represented: the CHMM and the Articulatory-Feature system have almost identical word error rate (21.8% vs. 22.1%; the difference is not statistically significant, and reverses polarity in one of the noise conditions). Third, however, the two systems make slightly different types of errors, and therefore it is possible for the two systems to correct one another. If all three of these speech recognizers are allowed to vote in order to determine the output word string (using the ROVER paradigm (Fiscus 1997)), word error rate is lower than the WER achieved by any one system alone. Furthermore, the ROVER combination of articulatory feature and CHMM systems has a tendency to be lower than the ROVER combination of three different CHMM systems (19.4% vs. 20.1% WER), suggesting that recognition accuracy may benefit from the use of two different methods to represent inter-articulator asynchrony.

All systems reported in Table 2 required the AF state variables to synchronize at every word boundary. It is common, in recent phone-based ASR systems, to allow two alternative pronunciations of each word: a version with cross-word triphones, and a version using only word-internal triphones (Hain et al. 2000, Young et al. 2002). Similar experiments were attempted using the AF-based ASR: models were developed that allowed the articulatory features to be asynchronous across the boundary between a word and its neighboring silence. The model that allowed asynchrony across word boundaries was considerably more computationally complex than the models reported in Table 2.

Because of the higher computational complexity, WER was only computed for the noise-free test condition; the resulting WER (7.5%) is significantly higher than the WER of any system in Table 2. Further research will seek to reduce the computational complexity and the WER of AF-based ASR with coarticulation across word boundaries.

4. The foot

The stress foot is the domain of lexical stress allocation, and of the strengthening or reduction of vowels and consonants (Turk & Sawusch 1997, Kim 2006). Lexical stress is deterministic, specified in the dictionary entry for all occurrences of a word, and therefore it is not difficult to use in ASR. In most standard English-language ASRs, for example, the dictionary entry for each word specifies whether any given vowel or alveolar stop should be implemented in reduced or unreduced form; reduced vowels are labeled as schwa (/AX/), and reduced intervocalic alveolar stops are labeled as flap (/DX/) (Lee & Hon 1989). Some systems also distinctly model fronted schwa (/IX/) and/or nasal flap (/NX/) (Zue et al. 1990). These forms of reduction are hard-coded in the dictionary, and may be present in the dictionary regardless of whether or not the dictionary explicitly labels the location of lexical stress.

In the absence of phrasal prominence, it is not clear whether stress-related differences other than vowel reduction are useful for speech recognition. Van Kuijk & Boves (1999) found that unreduced lexically stressed and unstressed vowels, without pitch accent, did not differ significantly in pitch, energy, or duration, and hence were indistinguishable in an automatic speech recognition system. Bates & Ostendorf (2002) and Bates et al. (2007), however, found that lexical stress can be used in automatic speech recognition as a form of optional context. In their study, triphone hidden Markov models were clustered into acoustically similar allophone clusters, as proposed by Odell, Woodland & Young (1994). In addition to the phoneme context questions proposed by Odell et al., however, Bates and Ostendorf also used questions about prosodic context (lexical stress, syllable position, and position in the word) to determine the clustering of allophones. The inclusion of prosodic context led to a statistically significant WER reduction. Hasegawa-Johnson (2006) has confirmed the results of Bates and Ostendorf using the training methods provided by a publicly available ASR toolkit.

5. The syllable

Syllable context impacts the acoustic implementation of a phone more than context at any other level. Indeed, any given articulatory gesture may lead to radically different spectrotemporal patterns, depending on its syllable context. Consider, for example, the word “backed” (Fig. 7). This word contains three stop consonants; because of their relative positions in the syllable, the places of articulation of these three stops are communicated by three very different types of acoustic information. The place of the final /d/ is communicated by an ejective burst spectrum. The place of the /k/ is communicated by formant transitions during the last 70ms of the vowel. The place of the initial /b/ is communicated by both a turbulent burst and by formant transitions during the first 70ms of the vowel, but experiments with synthetic speech (Delattre et al. 1955) and digitally modified natural speech (Nossair & Zahorian 1991) have shown that either of these cues may be excised without impairing listeners’ ability to understand the stop. The closure transition, burst spectrum, and release transition of a stop are thus redundant acoustic correlates; unambiguous presence of any one of these three acoustic patterns is enough to force listeners to hear the desired distinctive feature.

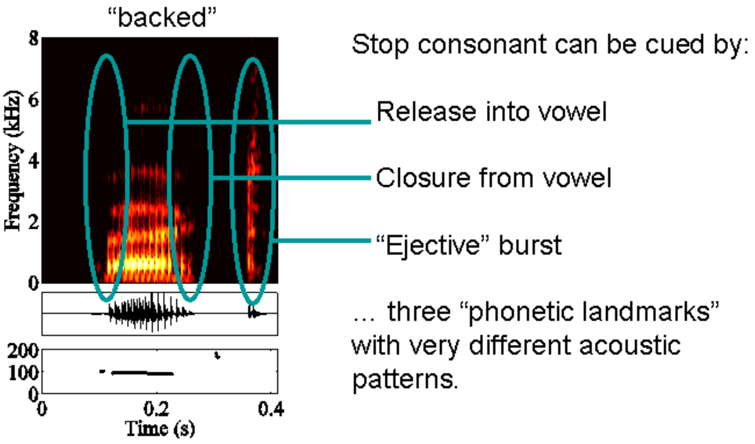


Figure 7: Redundancy of stop consonant landmarks: A stop consonant can be correctly recognized if a listener hears only the release (the /b/ in “backed”), only the closure (the /k/ in “backed”), or only an ejective release (the /d/ in “backed”).

Context at the level of the syllable is modeled, explicitly or implicitly, in every modern ASR. Triphones, for example (Lee & Hon 1989), implicitly distinguish between stop consonants that are signaled by the closure only (e.g., the /k/ in “backed,” whose triphone representation is /AE-K+D/), the release only (e.g., the /k/ in “miscast,” whose triphone representation is /S-K+AE/), or both (e.g., the /k/ in “backup,” whose triphone representation is /AE-K+AH/). Since most triphones do not explicitly represent syllable boundary, however, some acoustically important effects are not coded by triphones, therefore it has been proposed that acoustic models should be sensitive to the locations of syllable boundaries (Greenberg 1999). In the most extreme case, one may create an

ASR that uses syllables or demisyllables instead of phones as the fundamental building blocks of speech. The use of demisyllables as acoustic units is intuitively appealing, in part because it works so well in Chinese and Japanese. In English, however, the number of possible demisyllables is quite large, and the majority of possible demisyllables are rarely used, thus their acoustic correlates are not robustly represented in any reasonable-sized training corpus. Doddington et al. (1997) proposed solving the data sparsity problem by using syllabic acoustic units to augment a phone inventory rather than replacing it. Ganapathiraju et al. (2001) found that their best system included the following acoustic units: 200 monosyllable words, 632 common syllables, and triphones. In such a system, the "pronunciation" of any given word is given in terms of the largest available units: whole words if available, else syllables, else triphones.

Stevens et al. (1992) proposed a different method for representing syllable context. In the "landmark-based speech recognizer" they proposed, phones are replaced by four different types of acoustic speech recognition units: consonant closure landmarks, consonant release landmarks, syllabic peak landmarks, and intervocalic glide landmarks. The set of English landmarks is reasonably small: depending on the way in which they are enumerated, one typically finds that there are fewer than 1,000 acoustically distinct syllable-internal consonant-vowel and vowel-consonant biphones in English, and that all of them are well represented in a database the size of TIMIT (about 14 hours). To further simplify the task, Stevens et al. proposed detecting each landmark using a class-dependent modulation filtering algorithm, and labeling it using a series of binary distinctive feature classifiers. Landmark detection and distinctive feature classification algorithms have been developed using knowledge-based approaches (Espy-Wilson 1994, Liu 1995, Hasegawa-Johnson 1996, Bitar & Espy-Wilson 1996, Howitt 2000, Chen 2000, Pruthi & Espy-Wilson 2004), neural networks (Kirchhoff et al. 2000, King & Taylor 2000, Chang et al. 2001), and support vector machines (SVMs) (Niyogi & Ramesh 1998, Niyogi & Burges 2002, Juneja & Espy-Wilson 2003).

In July 2004, we trained and tested a number of different large-vocabulary continuous speech recognition (LVCSR) systems that fit the framework schematized in Fig. 8 (Hasegawa-Johnson et al. 2005). All LVCSR systems began with a high-dimensional multi-frame acoustic-to-distinctive feature transformation, implemented using SVMs trained to detect and classify landmarks. SVM inputs included MFCCs (computed using two different window lengths), formant frequencies and amplitudes (Zheng et al. 2004), knowledge-based acoustic parameters (Bitar & Espy-Wilson 1996), and multiscale spectrotemporal rate features (STRFs) (Mesgarani et al. 2004). Distinctive feature probabilities estimated by the support vector machines were then integrated using one of three different pronunciation models: a dynamic programming algorithm that assumes canonical pronunciation of each word, a DBN implementation of articulatory phonology, or a discriminative pronunciation model trained using the methods of maximum entropy classification. Log probability scores computed by these models were then combined, using log-linear combination, with the other word scores available in the lattice output of an HMM ASR, and the resulting combination scores were used to compute a second-pass speech recognition output.

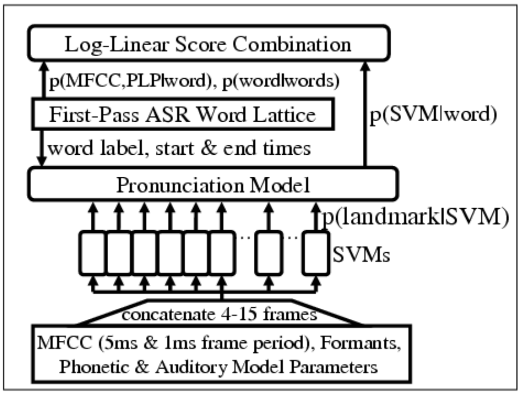

Log-Linear Score Combination
p(MFCC,PLP|word), p(word|words)
First-Pass ASR Word Lattice
p(SVM|word)
word label, start & end times
Pronunciation Model
p(landmark|SVM)
SVMs
concatenate 4-15 frames
MFCC (5ms & 1ms frame period), Formants,
Phonetic & Auditory Model Parameters

Figure 8: Schematic overview of landmark-based speech recognition systems implemented for large-vocabulary speech recognition by Hasegawa-Johnson et al. (2005)

A hybrid SVM-DBN landmark-based speech recognizer was created by combining the generative pronunciation model of (Livescu & Glass 2004) with the SVM acoustic observation probabilities described above. In the generative pronunciation model, hidden variables in a DBN represent features based on the tract variables of (Browman & Goldstein 1992), including the locations and/or degrees of opening of the lips, tongue, and glottis/velum. Each word's baseform pronunciations are mapped to tract variable trajectories. The DBN allows the tract variables to go through their trajectories asynchronously (while enforcing some soft synchrony constraints, encoded as distributions over degrees of asynchrony). The system developed in this way is similar to that shown in Fig. 4, with two key differences. First, the lips, tongue, and glottis/velum are allowed to take on "surface" values that differ from their canonical or "underlying" phone targets: for example, the variable **phone_name_lips** is divided into two hidden variables called, respectively, **phone_name_lips_underlying** and **phone_name_lips_surface**. Second, instead of **PLP observations**, the landmark-based speech recognizer observes the classification posterior probabilities computed by forty different SVMs trained to detect and classify landmarks.

Table 3 shows a sample of the word error rates obtained with this system on a three-speaker subset of the RT03 development test set. The baseline system in these experiments was the SRI EARS large vocabulary speech recognizer as of 2003 (Stolcke et al. 2003). It is worth noting that the WER of any speech recognizer is a moving target: the WER of the 2005 SRI system was approximately half that of the 2003 system. All rescoring experiments combined the log likelihoods of the SRI recognizer with log likelihoods of the DBN. Two rescoring experiments are reported in the table. In the first experiment, the DBN observes outputs of all SVM-based landmark detectors and classifiers. In the second experiment, the DBN observes the outputs of only the SVMs whose per-frame classification accuracy exceeds some reasonable threshold. The proposed methods show a trend (not statistically significant) toward reduction of WER on this development test dataset. Confirmation of these results using independent evaluation test data was not completed.

Table 3: Word error rates (%) in lattice rescoring experiments on a three-speaker subset of the RT03 development set. The last line of the table shows the WER achieved when the DBN observes only those SVMs whose per-frame binary classification accuracy exceeds a reasonable threshold.

System setup	WER
Baseline	27.7
SVM-DBN, all SVMs	27.3
SVM-DBN, high-accuracy SVMs only	27.2

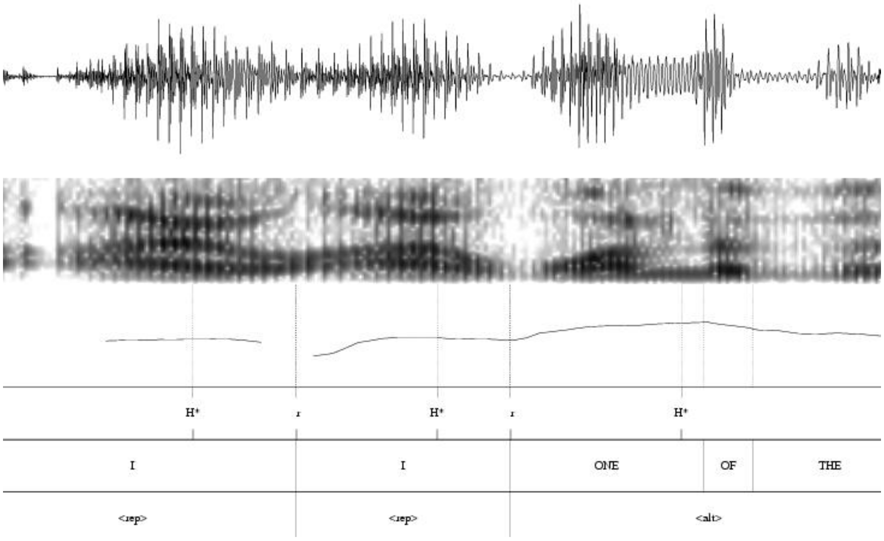


Figure 9: Transcription of prosody and disfluencies in the phrase “I, I, one of the...”

6. Disfluency

Disfluency can change the acoustic implementation of a phone, therefore the minimization of WER requires some representation of disfluency. Fortunately, disfluency is relatively easy to identify, in the following senses. First, linguistically naive transcribers are able to locate filled pauses and the interruption point of a repair or repetition disfluency with high levels of inter-transcriber agreement (Shriberg 2001, Meteer & Taylor 1995). Second, most disfluencies follow relatively stylized patterns of repair, repetition, and filled pause, and most disfluencies are therefore relatively easy to detect from an orthographic transcription of speech (Baron et al. 2002, Gupta et al. 2002, Kim et al. 2004, Lendvai et al. 2003). The key difficulties in the transcription of disfluency are: (1) if disfluency is not adequately modeled by the phone set of an ASR, disfluencies will be mis-transcribed as if they were fluent speech, causing a large number of speech recognition errors (Adda-Decker et al. 2003, Aylett 2003, Rose & Riccardi 1999), (2) all of the previous discussion refers to the most common patterns of disfluency, but some types of disfluency do not follow these patterns and are therefore difficult to transcribe (Shriberg 2001).

Fig. 9 shows a disfluency with a double reparandum: “I, I, one of the things I...” Fig. 9, like the remainder of this section, adopts the disfluency annotation system of Heeman and Allen (Heeman & Allen 1999). In their annotation system, the words being corrected are called the “reparandum” or REP, the correction is called the “alteration” (ALT), and filled pauses or meta-dialog are called the “edit” (EDT). In Fig. 9, the first reparandum is repeated, then finally repaired by the alteration. As shown, we find that most repair and repetition disfluencies in Switchboard contain no verbal EDT segment — many REP segments end in glottalization and/or elongation, but rarely in a verbal EDT segment. Conversely, most verbal EDT segments take the form of explicit filled pauses, most typically “uh” or “um” (Clark & Fox Tree 2002).

Disfluency is common in conversational speech. Of 1,100 words we have transcribed (Yoon et al. 2004, Cole et al. 2005), 40 are part of a reparandum, 37 are filled pauses, and 41 are part of an alteration, thus 10% of the words we have transcribed are part of a disfluency. This estimate is higher than most published estimates, perhaps because we include all words that are part of the reparandum or alteration, but most published studies estimate that at least 5% of the words in Switchboard are part of a disfluency (e.g., Shriberg 2001).

REP and ALT segments are not transcribed in most speech recognition training corpora, therefore it is difficult to train an ASR model of all aspects of disfluency. Two aspects of disfluency, however, are commonly transcribed in all speech recognition training corpora. First, filled pauses are usually labeled with unique lexical tokens: in the Switchboard corpus, for example (Godfrey et al. 1992), the words "UH" and "UM" are uniquely used to label filled pauses. Second, word fragments are often uniquely labeled. In Switchboard, for example, annotations specify the word that the talker was apparently trying to say (in the judgment of the transcriber); the unsaid portion is enclosed in brackets, e.g., the phone sequence /juniʔ/ might be transcribed using the word fragment "UNI[QUE]." Word fragments occur almost exclusively at the end of a disfluency reparandum, therefore word fragment labels specify the end (but not the beginning) of some (but not all) disfluency reparanda.

Filled pauses may be treated as regular lexical tokens in an ASR language model, forcing the language model to separately learn the lists of words which typically precede an "UH" or an "UM." Unlike the language model, the acoustic model of an ASR may benefit by giving "UH" and "UM" special treatment. The vowel in "UH" is acoustically similar to the vowel /ʌ/ in content words like "TUG," but the /ʌ/ of "UH" is usually weaker and longer. Similarly, the word "UM" is often produced with a drawn-out, low-intensity /m/. If the word "UM" is modeled by the phone sequence /ʌm/, then the aberrant statistics of the /m/ in "UM" will reduce the precision of the statistical model of /m/: because the phone model of /m/ is being used to represent both fluent and disfluent productions, it fails to compactly represent either. For these reasons, Greenberg, Hollenback & Ellis (1996) proposed representing the words "UH" and "UM" with the unique filled-pause phones /PV/ ("pause vowel") and /PN/ ("pause nasal").

The end of a REP segment—especially a REP that ends in a word fragment—is often glottalized. In Fig. 9, for example, glottalization is visible at both interruption points: the first REP segment ends in low-pitched creaky voicing, while the second REP segment ends in a glottal stop. Yoon et al. (this volume) have shown that WER of an ASR may be reduced by using an automatically labeled "creaky" vs. "modal" distinction as part of the definition of a phone.

7. Conclusions and future work

This paper has proposed using the prosodic hierarchy as an organizing framework for the sources of acoustically salient context information in ASR. Specifically, we have discussed five experimental systems, each of which divides the phone inventory into two or more subcategories as specified by the following prosodic and disfluency context features:

1. Position within intonational phrase (final vs. nonfinal)
2. Phrasal prominence (prominent vs. nonprominent)
3. Position within prosodic word (initial, medial, or final)

4. N-phone context (manner, place, and voicing of the preceding and following phones)
5. Lexical stress (primary stress, reduced, neither)
6. Syllable position (consonant release, consonant closure, syllabic nucleus, or intervocalic glide)
7. Fluency (filled pause vs. non-pause)
8. Voicing (creaky vs. modal)

Equation 4 suggests that all of the features above should be used to define a phone inventory. It is impractical, however, to divide a small speech training corpus into mutually exclusive subsets representing every possible combination of the features listed above. Instead, it is necessary to find some method of computing, and applying, an estimate of the specific acoustic transformations that relate one prosodic context to another.

Section 2 proposed using phonetic knowledge to define the most important acoustic differences among prosodic contexts. For example, in that section, the models of phrasally prominent and non-prominent examples of the same underlying phoneme are tied together in all acoustic dimensions but F0. Similarly, phrase final and nonfinal phones are tied together in all acoustic dimensions but duration.

Section 3 reviewed a common "tree-based splitting" approach to triphone context features, first proposed by Odell, Woodland & Young (1994). In that standard approach, the phone inventory of an ASR system is created through a tree-structured series of binary divisions of the training data. Each binary split is selected, from a list of candidate binary context features, in order to make the leaves of the new tree as acoustically compact as possible. The splitting process continues while each leaf of the tree contains a sufficient number of training examples. Bates & Ostendorf (2002) and Bates et al. (2007), proposed using a similar binary splitting method to model the acoustic salience of arbitrary prosodic context variables including syllable position, word position, and lexical stress. Borys (2003) proposed using the same method to model the acoustic salience of intonational phrase position and phrasal prominence. Yoon et al. (this volume) proposed using the same method to model the acoustic salience of voice quality labels.

Exhaustive splitting and tree-based splitting methods both work from the assumption that the "context-dependent phone" is an indivisible unit. Livescu & Glass (2004) have suggested, rather, that the scalar phone label should be split into a vector of AF labels, each representing the targets achieved by one articulator. Browman & Goldstein (1992) go one step farther, arguing that the phone should be replaced by three distinct set representations at each time *t*: a set of "gestures" that are intended or desirable at time *t*, a vector of "tract variables" that have been planned for production at time *t*, and a vector of articulator positions that are actually produced at time *t*. Most implemented computational models of articulatory phonology posit that the mapping from tract variables to articulator positions is usually trouble-free (in speech without pathology): most pronunciation variability comes from the mapping between gestures and tract variables.

All of the context variables discussed in this paper can be re-written in terms of articulatory phonology. For example, articulatory phonology greatly simplifies the representation of triphone context: All of the effects of triphone context are represented, in articulatory phonology, by the temporal overlap of competing gestures. The blocking of coarticulation across word or phrase boundaries may be represented, as suggested in §3, by forcing the gestures or tract variables to re-synchronize at the boundaries of prosodic words or phrases. The effects of syllable context may be represented, as suggested in §5, by developing distinct SVM or neural network classifiers designed to detect and classify the release and closure landmarks associated with any particular articulator.

Future work will try to develop comparable representations, in terms of articulatory phonology, for the effects of prosodic phrase context, prosodic group context, and disfluency. A promising method is suggested by the work of Byrd & Saltzman (2003). Byrd and Saltzman developed, based on the work of (Saltzman & Munhall, 1989), an algorithm for synthesizing articulator kinematics from hypothesized articulatory gestures. In their model, phrase boundaries are modeled by a π_T gesture (a “lengthening” gesture (Beckman & Edwards 1990)), whose function is to slow down the clock controlling the mapping between gestures and tract variables. Similarly, prominence is modeled by a π_S gesture (a “strengthening” gesture (Fougeron & Keating 1997)), whose function is to increase the magnitude of all tract variable excursions during its period of activity. There is a natural mapping between the context variables considered in this paper and the π_S and π_T gestures of Byrd and Saltzman: lexical stress and phrasal prominence are different types of π_S gesture, while utterance, intonational phrase, and intermediate phrase boundaries each generate a different type of π_T gesture. The Articulatory Feature (AF) models of §§3 and 5 provide a good starting point for the implementation of a prosody-dependent articulatory feature ASR, e.g., it may be possible to simply add two more hidden state variables representing π_S and π_T. In order for these ideas to become useful in automatic speech recognition, the biggest remaining unsolved problem seems to be the creation of a probabilistic representation of the “lengthening” and “strengthening” functions—that is to say, we need to somehow represent “lengthening” and “strengthening” as learnable context-dependent transformations of the mode parameters or mixture parameters of a statistical ASR.

Mark Hasegawa-Johnson
Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
405 N. Mathews Ave.
Urbana, IL 61801, USA
jhasegaw@uiuc.edu

Jennifer Cole
jscole@uiuc.edu

Ken Chen
kchen22@wustl.edu

Partha Lal
p.lal@sms.ed.ac.uk

Amit Juneja
amjuneja@gmail.com

Tae-Jin Yoon
tyoon@uvic.ca

Sarah Borys
sborys@uiuc.edu

Xiaodan Zhuang
xzhuang2@uiuc.edu

Linguistic Patterns in Spontaneous Speech, 129-146
2009-1-025-006-000268-1

Prosodic Features of Spontaneous Utterance-initial Phrases in Bernese and Valais Swiss German*

Adrian Leemann[1,2] and Beat Siebenhaar[3]
University of Tokyo[1]
University of Berne[2]
Universität Leipzig[3]

We analyze the prosody of utterance-initial phrases in spontaneous speech. The answers of 16 interviews were categorized into empty pauses, filled pauses, first phrases, and the rest of the utterance. Fundamental frequency (F0) measurements indicate two prevailing patterns: one pattern demonstrates a declination in F0 from the filled pause to the first phrase and to the rest of the utterance while the other features a low F0 in filled pauses followed by an F0-increase in the first phrase and again a declination for the rest. Therefore, modeling intonation must give special consideration to first phrases and preceding pauses.

Key words: prosody, conversation analysis, Swiss-German, turn-taking, intonation

1. Introduction

Spontaneous and prepared speech differ in several ways. Abercrombie (1965) characterizes the latter as having a standardized intonation pattern, little variation in tempo, pauses that are set according to grammatical structures and as possessing little or no disfluency. The former, in contrast, may feature the omission of syntactic elements as well as the overwhelming presence of fillers and hesitations. Abercrombie advocates the study of 'genuine spoken language' (ibid.: 9) in the same way Fox Tree believes that '[t]he phenomena that are the hallmark of spontaneous talk have often been thought of as unwanted elements of speech, unfortunate by-products of speaking on the fly. However, another way of viewing these phenomena is as an integral part of the communicative enterprise' (2000:376), a stance this study, too, affiliates itself with.

Intonation and speech rate seem to be the most researched aspects of prosody in contrastive analyses between spontaneous and prepared speech. In terms of pitch, Swerts

[*] This research in the framework of the project "Quantitative Approaches to Geolinguistics of Swiss German Prosody" (100011-116271/1) is founded by the Swiss National Science Foundation.

et al. (1996) found that the F0 tends to be higher in read-aloud tasks than in spontaneous speech; while in spontaneous speech, intonation is more varied (Syrdal 1996). Speech rate, too, is higher in conversational speech, which leads to reductions in vowels. Speech disfluencies, such as filled and empty pauses and hesitation phenomena, are, too, a typical characteristic of spontaneous speech—disfluencies per word in spontaneous English speech vary from 5-10% (Shriberg 1999). From a pragmatic and pycholinguistic perspective, fillers such as *uh* and *um* announce delays in speaking; they offer extra time for the speaker to search for the desired word or the adequate syntactic structure, which may not be accessible at that very moment (Clark et al. 2002). Shriberg (2001) further notes that disfluencies tend to occur predominantly in utterance-initial positions. The present paper explores these utterance-initial phrases on a prosodic level and within the framework of our current National Science Foundation (NSF) research project 'Quantitative Approaches to Geolinguistics of Swiss German Prosody'.

On most linguistic levels, Swiss-German dialects have been examined reasonably well; that is the case for many local dialects and for a geolinguistic comparison thereof. However, there is a lack of prosodic descriptions of the dialects. This is where the current project pitches in: in recording two Alpine and two Midland dialects, we try to work out a gross geolinguistic model that is geared at revealing the main prosodic features of these dialects. In four different places (Bern, Zürich, Brig, and Chur), 20-30 subjects are recorded. The data is collected via spontaneous interviews that include questions regarding the informants' goals after graduation, what they do in their next vacation etc. The prosodically most relevant parameters, time and fundamental frequency, are then extracted and modeled. The comparison of the subjects from each location allows for a distinction between region-specific and individual prosodic characteristics. The comparison between the different recording locations offers insight into the geolinguistic structure of prosody.

With respect to modeling the intonation and the timing of individual speakers and speaker groups, however, we encountered the following problem: it is a well-known fact that within a conversation setting prosody varies according to discourse structure. Brown et al. (1980) concluded that new topics, which are regularly introduced by question-answer pairs, are often presented in a comparatively higher pitch. Longer pauses, too, have been associated with shifts in topic (ibid.), and speech rate also varies according to discourse structure. The question thus arose of how such distinct discourse structure-related differences in prosody could be incorporated in the models resulting from our project.

From the above it follows that this study is not only phonetic in its nature but also conversation analytic. We want to scrutinize the acoustic correlates of utterance-initial phrases; phrases, which must be viewed in the larger context of the interview situation, because they are articulated by speakers who are randomly and spontaneously chosen by the researcher. This conversation analytic aspect will be attended to in the first section of this paper, followed by the discussion of the phonetic component. To establish the link between phonetic research, i.e. prosody research in our case, and conversation analysis (hereafter CA), it is then shown that, despite the vast amount of literature on prosodic features of phrase-final structures, it seems that little research has been conducted on the prosodic features of spontaneous utterance-initial phrases and phrase-initial features.

2. Conversation analysis and phonetics

2.1 Conversation analytic component

For Harvey Sacks conversation analysis is the study that 'seeks to describe methods persons use in doing social life ...' (1984:25). A core concept in CA is that of turn-taking, i.e. '[t]he talk of one party bounded by the talk of others ..., with turn-taking being the process through which the party doing the talk of the moment is changed' (Goodwin 1981:2). Sacks et al. (1974) designed a set of rules that account for places where a next turn can be anticipated. Among other things, these instances of possible turn-takings, which they refer to as transition-relevance places, are enacted with signals, such as discourse markers as well as syntactic and semantic features in the ongoing turn. More importantly, in the context of the present study, prosodic features such as pausing, duration of segments, and intonation constitute further turn-yielding signals (Taboada 2006:7). Couper-Kuhlen & Selting (1996:11) welcome the fact that CA, as a socially oriented approach towards the study of language, has acknowledged the importance of prosodic features in language-in-interaction, as opposed to studying prosody from a structuralist point of view.

This overlap between CA and phonetics is also appreciated by Local (2003:1), who adds that despite the large number of available corpora of spontaneous speech, surprisingly little has been used for further, talk-in-interaction, analyses. Local poses questions of the following nature '[h]ow do speakers/listeners manipulate fine phonetic detail in producing and interpreting the moment-to-moment flow of everyday conversation?' (ibid.). This is, in fact, one of the aspects conversation analysts have not addressed thoroughly enough, so Couper-Kuhlen & Selting (1996).

For the transcription process, conversation analysts more often than not apply Jefferson's transcription system (cf. Jefferson 2004—for German 'Gesprächsanalytisches Transkriptionssystem' see Selting et al. 1998), with the aim of capturing talk as it occurs in daily conversations, '... in all its apparent messiness ...' (Hutchby & Wooffitt 1998:75). Such an ambitious goal obviously meets with criticism. Kendon (in Hutchby

& Wooffitt 1998:76) points out that '[i]t is a mistake to think that there can be a truly neutral transcription system ... Transcriptions, thus, embody hypotheses.' In CA's defense, Hutchby & Wooffitt (ibid.) state two main goals of CA: first, to describe the dynamics of turn-taking and second to elucidate the characteristics of speech delivery, including prosodic features such as stress, pauses, enunciation, intonation, pitch etc. Such transcriptions include, for example, the measurement of pauses in tenths of seconds (ibid.: 81) or the most literal transcription of laughter as possible (ibid.: 83).

While Local & Kelly (1989:204) believe that pausal phenomena and audible respiratory activity are consistent in CA, they argue that tempo, pitch, loudness, vowel quality, voice quality etc. are often rendered inconsistently and arbitrarily in CA transcripts. Hutchby & Wooffitt (1998:77) take note of such a counter argument, yet state that if a CA analyst were to pay closer attention to phonetic phenomena, transcripts would go beyond the reader's understanding thereof. Moreover, conversation analysts believe that CA has a different aim, namely 'to get as much of the actual sound as possible into our transcripts, while still making them accessible to linguistically unsophisticated readers' (Sacks et al. 1974:734). Such a position is, however, problematic as fine phonetic observations can shed light on details in the analysis of a conversation that are not revealed if one adheres to such methodologies.

Smith & Clark (1993), for instance, discusses the nature of answer-questions adjacency pairs. In answering a question, the respondent often delays his answer. The timing of such delays is crucial, as it makes the delay open for interpretation on the part of the questioner. Did the respondent not understand the question, not retrieve the required information? Can she/he not formulate his response? In order to find answers to such questions, exact measurements of pausal duration, measurements beyond stopwatch-timing, need to be made. A further example is rising or falling intonation in an answer to a question. Rising intonation often denotes uncertainty on the part of the respondent, in contrast to falling intonation in an answer, which does not leave open such an ambiguous interpretation. The phonetic correlate can be minimal, its interpretation, on the other hand, may not be the one intended by the speaker.

This is, in part, where this study starts off. While CA transcripts may reveal that utterance-initial phrases are indeed prosodically different from the rest of the utterance, we want to illustrate what this looks like on a more detailed level. This is achieved by means of instrumental analyses (fundamental frequency measurements and pause duration measurements). In other words, we want to bring to light the prosodic features of utterance-initial phrases as provided by a speaker who, heteronomously, provides answers to our questions.

2.2 Phonetic component

First descriptions of German prosody date from the late 19th century with the emergence of phonetics. Even early monographs on Swiss-German dialects (Vetsch 1910, Wipf 1910 and others) hold sections on dialectal prosody. However, early impressionistic descriptions that included statements on the geolinguistic distribution of prosodic patterns (Bremer 1893, Sievers 1912) could not be verified until the present day (Gilles 2005). Based on perception models, Isačenko & Schädlich (1964) built models for sentence intonation which questioned the prevailing syntax based models (Bierwisch 1966). These were revised by communicative oriented descriptions (Féry 1993, Selting 1995). Actual research on linguistic-based German prosody mainly has three branches: prosody as part of pragmatics (e.g. Kohler 1991) and interaction (Selting 1995), phonological representation of intonation (e.g. Grice & Baumann 2002, Gibbon 1998), and variationist research on differences of regional prosody (e.g. Gilles 2005, Peters 2004, Siebenhaar 2004, Siebenhaar et al. 2004). In addition, speech technology views prosody from a technology-based perspective.

Results from studies that describe standard German — which in some regions is still more of a construct than a real norm, especially in prosody — can only be referred to as a background to our dialect data. Kohler's (1987, 1991) perception studies on the communicative function of pitch alignment as well as the descriptions and perceptive tests of the interplay between different phonetic aspects in determining phrase boundaries in spontaneous speech (B. Peters et al. 2005), are based on a German standard. However, it is not mentioned that this is the standard German spoken in Northern Germany. Correspondingly, Atterer & Ladd (2004) have shown that even in laboratory read speech there are prosodic differences in standard German that can be reduced to regional aspects. They have shown that speakers from southern Germany generally show later peaks than speakers from northern Germany do. Consequently, regional intonation has become a focus of research in the last decade as seen in publications by Gilles, Peters, Selting, and Auer. They describe intonation patterns that are specific to a region or that have a different communicative function, in one region as opposed to another. Their work is grounded in the description of the contours of final nucleus syllables in their functional distinction of termination and continuation. On the one hand, the comparison shows a geolinguistic difference between southern and northern regions with preferences for different patterns; on the other hand, regionally specific patterns become evident (Gilles 2005). However, the phonetic distinctions are not as apparent as they are on the segmental level.

Siebenhaar and co-workers (Siebenhaar 2004, Siebenhaar et al. 2004, Häsler et al. 2005) made first modern attempts at a description of Swiss-German dialectal prosody.

The prosody of interviews of three speakers from two different dialect regions is analyzed in such a way as to subsequently formulate models for a dialectal speech synthesis system. The aim was that the models should build a methodological basis with which to compare dialectal prosody. These analyses, indeed, indicate clear differences between the speakers. Unfortunately, the results could only partially be traced to dialectal reasons, as the sample was too small. However, it can be said that, as is the case in read speech (cf. Keller 1994:7), timing seems to be more stable than intonation. The timing models achieve a correlation with real spontaneous data that is nearly as high as that achieved with read speech. While in spontaneous speech timing is quite predictable and specific to every linguistic variety, intonation seems to be more variable and dependent on situational and/or functional factors.

The mentioned descriptions of German prosody focus on either timing, peak alignment of accents, or on the intonation contours of final nucleus syllables. Utterance-initial phrases, however, have not been researched extensively.

2.3 Prosodic features of utterance-initial phrases

With respect to turn-medial and turn-final phrases, Stephens & Beattie (1986) found that subjects who are presented with an audio recording were able to discriminate between turn-final and turn-medial utterances in the case of disagreements in conversation, thus highlighting the role of prosodic features in the regulation of turn-taking. In the analysis of an interview with Margaret Thatcher, Beattie et al. (1982) suggest that non-verbal turn-yielding signals include a low drop in pitch and loudness. Maclay & Osgood (1959:20) propose that pausal phenomena, too, serve to identify the end of phrases and sentences. In comparison to research on turn-medial/turn-final utterances, turn-initial phrases have not been investigated as thoroughly, possibly because their function is not considered relevant for turn-taking or for the constitution of a turn.

Much of the research in the context of prosodic features of utterance-initial phrases has evolved around analyses of fillers in terms of their intonation and duration (cf. Shriberg 1999, Swerts 1998), and not around the acoustic analysis of entire turn-initial phrases as such. Analogously, there is abundant literature on the pragmatic (cf. e.g. Smith & Clark 1993, Corley 2003, Clark & Fox Tree 2002) and social psychological (cf. e.g. Cook & Lalljee 1973, Siegman & Pope 1965) contextualization of such discourse markers. In describing the acoustic correlates of disfluent speech, Shriberg (1999) mentions that the lengthening of syllables before the actual point of interruption is a typical feature of everyday, disfluent speech. Despite the duration modification in lengthened syllables, however, the fundamental frequency remains largely unaffected. Further she shows how vowels of filled pauses, most often acoustically similar to schwa, are articulated significantly longer than where the same vowel occurs in fluent contexts. With respect to the intonation of filled pauses, Shriberg finds a low F0 and a linear or a slightly gradual fall in pitch. Swerts (1998), alternatively, examines a possible correlation between filled pauses and discourse structure. He shows that pause fillers are more likely to occur in initial-phrases, if preceded by major discourse boundaries. Additionally, he concludes that the fillers in initial-phrases are segmentally as well as suprasegmentally different from those in phrase-medial or phrase-final positions.

3. Methods

Given the assumption that utterance-initial phrases are prosodically different from utterance-medial or utterance-final phrases, it needs to be clarified how such a hypothesis can be tested. The data used in this study was retrieved from a corpus of spontaneous speech that was collected within the National Science Foundation (NSF) research project 'Quantitative Approaches to Geolinguistics of Swiss German Prosody' at the University of Berne. 25 subjects, all of whom attended grammar school and were aged between 18-22 at the point of the documentation, were recorded in Brig, which represents the Western alpine variety of Swiss German. 25 subjects were recorded in Berne, a city that stands for the Western midland dialects of the country. From this pool of recordings, 16 were used for analyses in this paper: eight Valais (four female, four male) and eight Bernese (four female, four male) recordings. Within the selection of the speakers' sex and dialect, the recordings were chosen by chance.

With the aim of extracting as much spontaneous language as possible, the subjects were asked to answer a number of questions as part of a spontaneous interview. The interview consisted of questions such as 'What do you think you will do once you have graduated?', 'What do you do in your spare time?', 'What does your next vacation look like?' etc. This form of interview is considered the most suitable method with which to collect naturally occurring language, since it sheds broader light on non-marked language use with a stranger. The interview constitutes roughly half of each of the 20-minute recording sessions. As said by Selting (1995:243ff), such questions present non-restrictive, open questions, along with a new or renewed focusing (WH-questions or verb-first questions). The subject's answer normally comprises the beginning of a narrative contribution to the conversation.

The conversation between the researcher and the informant was manually labeled with PRAAT (Boersma & Weenink 2006). If appropriate, the first label that pertained to the subject is a filled pause (FP). Secondly, the first phrase (PHRASE) was marked off from the rest of the utterance (REST). Possibly occurring non-filled pauses (#)

between the question and the PHRASE were labeled as well. Various measuring points were elicited. The most relevant ones are summarized below:

- A possible # preceding and/or following an FP.
- The type of FP: *aa*, *ää*, *mm*, *ja* = yes), *ja+* = 'yes' plus other particle), also 'well'), '*vowel*' = all other vowels), '*repetition*' = repetition of a word of the question), *nei* 'no'), 'conjunction', and *ts*. It has to be stressed that 'real' *ja*, *nei* 'yes/no') answers do not count as FP but as PHRASE.
- The sequence of #FP#, PHRASE, and REST i.e. which of these components was present in the answer to the interviewer's question).
- Type of literal question yes/no,[1] open).
- Fundamental frequency F0) mean for each FP, PHRASE, REST occurrence.
- differences in F0 in %) between FP/PHRASE and PHRASE/REST.
- FP duration in ms for each occurrence.
- Duration of # before FP in ms.

In the following sections, the findings of our study are described and discussed. Our concerns are fourfold: firstly, we explore whether or not utterance-initial phrases, as provided by the speakers in the sample at hand, are prosodically different from the rest of the utterance. Secondly, if there are differences, we question whether they can be attributed to the specific dialect regions. Thirdly, if they cannot be attributed to the speakers' dialectal backgrounds, we address what other factors they could be attributed to. Lastly, we want to attend to some of the repercussions of these findings for our current project and, from a broader perspective, on prosody research per se.

4. Data analyses

The following analyses focus on some of the relevant distinctions that could be extracted from our data. The first section deals with the use of pauses while the second addresses the fundamental frequency.

4.1 Use of pauses

First insights revealed a different use of filled and empty pauses between the two speaker groups.

[1] These yes/no questions were also intended to be open questions, yet it turned out that they were formulated as 'Can you tell me what you want to do after school?' which literally is a yes/no question.

4.1.1 Filled pauses and type of question

The first issue that needs to be addressed is whether or not the type of question — open question or literal yes/no question — has an impact on the type of FP (see above). In fact, all questions were intended to be open questions, yet some were formulated as yes/no questions, such as ‘Can you tell me what you want to do after school?’. These indirect questions were marked as literal yes/no-questions. The result of a contingency test is not significant. Thus, each type of FP is equally possible in the replies to both types of questions for speakers of both dialect groups.

4.1.2 Filled pauses and empty pauses

The total number of answers the students provided comprises 353, out of which the speakers delivered 251 turn-initiations with an FP, i.e. in 71.1% of the cases they provided some sort of hesitation marker in their answers. In 50% of these FPs there is a preceding #, in 27% the FP is preceded and followed by a #, in 17% the FP stood alone, and in 5% the FP follows the question directly without a #, but it is followed by a #. If we look at the type of FP that was used by the subjects, we get the following distribution, shown in Figure 1:

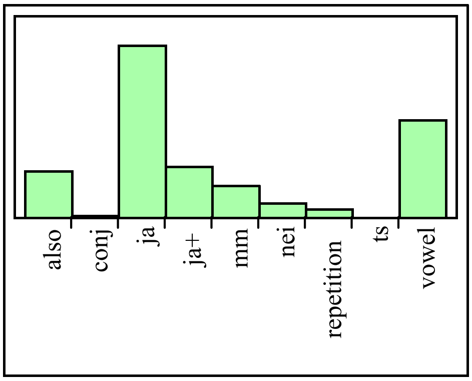


Figure 1: Frequency distribution of FP types.[2]

The distinction between FPs with no semantic meaning and FPs with minimal semantic meaning does not show any significant differences. With a frequency of nearly 40 %, *ja* is by far the most recurrent type of FP. Therefore, this positive feedback (which

[2] *ja+* stands for the filler *ja* and a subsequent sequence of sounds, such as /jaː əmː/ or /jaː dəs/ which corresponds to *yes*, *um* and *yes*, *that*.

is different from the *ja*, denoting an answer to a yes/no-question) indicates a generally positive attitude on the part of the interviewee towards the interviewer.

Further tests showed that there are differences between our Valais and Bernese speakers in terms of the types of FPs they use. Our Valais speakers use more *also*, and *ja+* than Bernese speakers do, while the Bernese fill their pauses with *ja* and *mm* more frequently. It also has to be considered whether there are differences in the use of FPs in their varying sequences, i.e. # following FPs, # preceding FPs, and the duration of # before FPs between Valais and Bernese speakers. A contingency analysis of the sequences of FPs (FP#, FP, #FP#, and #FP) between Bernese and Valais speakers suggests that they make use of or omit # differently, cf. Table 1.

Table 1: Contingency table of FPs by Bernese and Valais speakers

	#FP	#FP#	FP	FP#
BE	82 52.9%	30 19.4%	36 23.2%	7 4.5%
WS	53 46.1%	43 37.4%	12 10.4%	7 6.1%

The contingency table indicates that #FP and FP alone are used more often by the Bernese speakers with nearly 53% and 23% respectively — while the Valais speakers show a large number of #FP#, namely 38%. Overall, the Valais speakers use significantly more #s to frame the FP than Bernese speakers do. In fact, these differences are significant for # before FP and after FP.

A further test is used to address another issue, namely whether the durations of #s before FPs provided by Valais speakers are quantitatively different from those of the Bernese speakers. An ANOVA indicates that this is indeed the case; it turns out that the #s before the FPs of the Valais speakers (777 ms) are significantly longer than those of the Bernese speakers (601 ms). When looking at the duration of the FPs, we find a mean of 480 ms. The Valais speakers' FPs tend to be longer than those of the Bernese speakers, but the difference is just under significant. This result is interesting because it accentuates the utterance-initial position. Preliminary explorations of pause duration in the whole interview of a reduced data set have shown that Bernese speakers make more and longer inter-utterance pauses than Valais speakers.

4.1.3 Summary and discussion

If we recapitulate these findings, we get the following differences between Bernese and Valais speakers: the type of FP used and the duration thereof for the two dialect groups is different. Both groups use or omit # differently. The Valais speakers have more # after and before FPs, and the Valais speakers' #s before the FPs tend to be longer. The reasons for these differences remain unclear. Valais speakers may have had more production difficulty than Bernese speakers, possibly they were less comfortable with the discussed topics, or they may have been less honest in answering the questions as opposed to the Bernese speakers. These are some of the aspects that need to be taken into consideration for the interpretation of filled and empty pauses, so Fox Tree (2002). Siegman & Pope (1965) assume that interview questions that are low in specificity may correlate with caution and hesitation markers in the interviewee's speech. However, frequency counts show that both groups are asked a nearly identical number of open or yes/no questions. Therefore, it may be that the interviewers' questions in the Valais recordings were perceived as having a low degree in specificity. While such explanations need to be viewed as possible causes, it seems more likely, however, that the results can in fact be ascribed to the different dialect regions. This different use of pauses can be viewed as evidence for regionally different communicative behavior, which should be taken into account by conversation analysts.

4.2 Fundamental frequencies

At this stage, we need to contemplate the fundamental frequency of FPs in the larger context of the FP-PHRASE-REST sequence. The following calculations were made by using the F0% values, where the basis for every subject is its mean F0 of the REST. This is beneficial in that subjects with different fundamental frequencies (especially male and female subjects) can directly be compared. The results indicate that there are two distinct patterns of F0 distribution from FP to PHRASE to REST. On the one hand, there is a gradual declination in F0 as the speaker progresses from the FP to the PHRASE to the REST. An example of this pattern is the F0% distribution of subject BE01m given in Figure 2:[3]

[3] BE01m stands for Bern, subject Nr. 1, male; WS stands for Wallis, i.e. Valais.

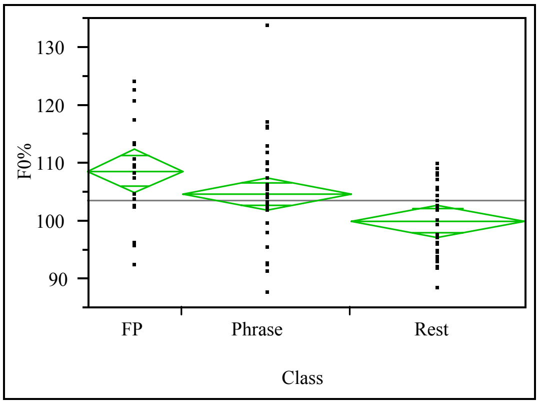


Figure 2: Gradual declination of F0% from FP to PHRASE to REST for BE01m[4]

Out of 16 speakers, five seem to adhere to this pattern of F0 declination. In stark contrast, however, all the other speakers pursue a different pattern of F0 modification in the course of the FP to the PHRASE to the REST: low F0 in FP to high F0 in PHRASE to low F0 in REST, which may take a shape as depicted in Figure 3 (subject WS25f):

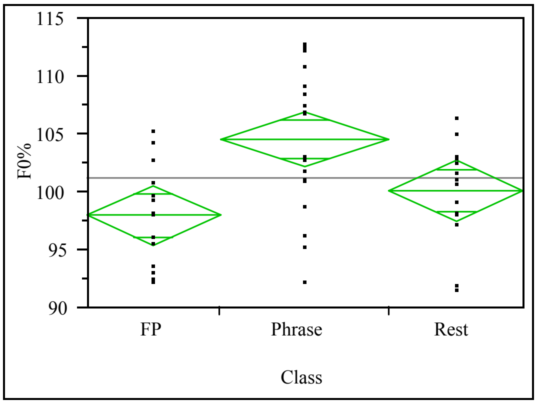


Figure 3: Low – high – low F0%-pattern from FP to PHRASE to REST for WS25f

The fact that all observed speakers can be placed in either of these groups is quite extraordinary. Both patterns show a declination in F0 from PHRASE to REST. This declination is not dependent on the fact that the first phrase of an utterance is more likely to be a continuing phrase with a high boundary tone, which raises the mean F0 of

[4] What looks like a diamond in these figures indicates the confidence interval (95%), while the dots show the dispersion, and the diamond-central line the median.

that first phrase compared to the remaining phrases of an utterance. High and low boundary tones are not clearly differentiated as a function of continuing or terminal phrase such as this is the case in standard German and many other languages (cf. Wipf 1910:23); yet, what causes an FP to be higher or lower in F0 from the following first PHRASE? The following section addresses factors that may affect these F0 patterns, including regional variation, type of FP, sex, and the type of question.

4.2.1 Regional variation

In order to find out whether regional variation may be the cause for these patterns, the relative overall means (%) of FP-PHRASE-REST of the two groups were compared. Whether there are significant differences in F0 between FP-PHRASE-REST in the Bernese group and in the Valais group was tested with a oneway ANOVA.

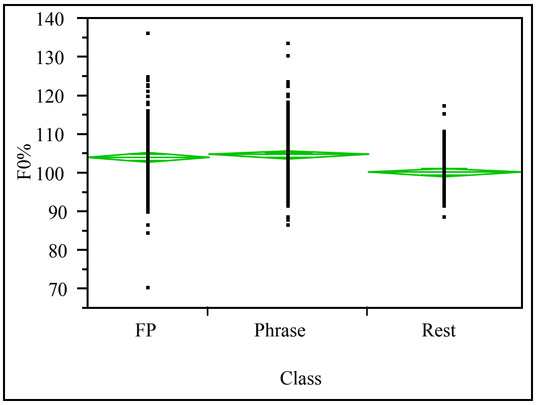


Figure 4: Oneway Analysis of F0% by FP, PHRASE, REST, Bernese speakers

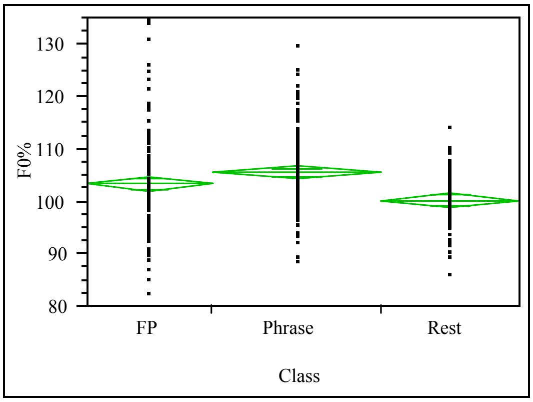


Figure 5: Oneway Analysis of F0% by FP, PHRASE, REST, Valais speakers

While for the Bernese speakers (Figure 4) the PHRASE and the FP demonstrate significant differences from the REST, the FP is not significantly different from the PHRASE. The t test confirms that the differences between the PHRASE and the REST, the FP and the REST, as well as between the PHRASE and the FP are significant within the Valais group of speakers (Figure 5). These results allow for the conclusion that in the sample at hand, utterance-initial phrases are indeed different in terms of their F0 as opposed to the following phrases. More importantly, however, the results show that whether or not a speaker follows the low-high-low or the declination pattern is not contingent upon the speaker variety.

4.2.2 Type of filled pause

A contingency table (Table 2) should reveal a possible connection between the type of FP and the rise or fall of F0 from FP to PHRASE. As Bernese and Valais subjects show the same tendencies of F0 patterning, both groups were examined together, which allows the running of a Chi-Square-Test with all cells having expected counts above 5. This Chi-Square test results in a significant difference between the cells, cf. Table 2.

Table 2: Contingency table of rise or fall from FP to PHRASE by FP type
FP types with a frequency lower than 10 were excluded.

Count Col %	*also*	*ja*	*ja+*	*mm*	'*vowel*'	
fall	3 12.0%	47 51.7%	11 37.9%	11 61.1%	13 25.0%	85
rise	22 88.0%	44 48.4%	18 62.1%	7 38.9%	39 75.0%	130
	25	91	29	18	52	215

Results indicate that for every type of FP, both F0 movements from FP to the following PHRASE (rise and fall) are possible. For *mm*, however, a falling movement is more likely, while for a '*vowel*', *ja+*, and *also*, a rising movement is more probable. An assumed interrelation between fall and rise on the one hand and open and literal yes/no questions (cf. §3.2.4) on the other does not show significant differences, except for *ja+*. A literal yes/no question that is answered with a *ja+* can equally be realized with a higher or lower fundamental frequency than the following PHRASE, while the answer to an open question is normally (10/11) realized with an FP lower in F0 than the following PHRASE. At this point, a finer distinction of the FPs' communicative functions could provide a more precise picture.

4.2.3 Sex

To find out whether the speakers' sex has an impact on the F0 of the FP, an ANOVA is run of the F0% factored by the speakers' sex. The results show that the Bernese speakers do not show significant differences between male and female speakers. The Valais speakers, conversely, demonstrate differences in the F0 of FP between the two sexes that are below a threshold level of .05, see Figure 6:

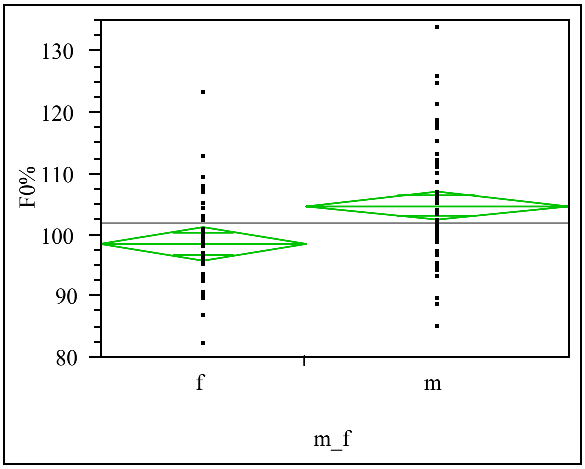


Figure 6: Oneway Analysis of F0% of FP by female/male Valais speakers

The diagram suggests that the Valais women's fundamental frequency of the FPs is below the fundamental frequency of the REST, while that of the men is somewhat higher. Yet, this difference between Valais men and women does not depend on the fact that only the men keep to both patterns while only the women draw on the low-high-low pattern. The men do not show any F0% differences that are contingent upon the patterns they adhere to. Thus, in the case of these eight subjects, it is the difference between the men and the women that is significant.

4.2.4 Type of question

It was of further interest to explore whether the type of question has an influence on the F0 of the FP and on the first PHRASE. First, tests showed that no correlation exists between the type of question (literal yes/no, open) and the F0 of the FP. Second, there is a clear correspondence between the type of question and the F0 of the first PHRASE. For both varieties, the mean F0 value for answers to open questions was significantly (3 %) higher than in answers to yes/no questions. Furthermore, it turned out that Valais speakers showed a distinct pattern of behavior in relation to the succession of FP and first PHRASE in answers to open questions as opposed to yes/no questions (cf. Table 3).

Table 3: Contingency table of rise or fall from FP to PHRASE by question type for Valais speaker

Count Col %	Literal yes/no	open	
fall	26 44.8%	13 25.0%	39
rise	32 55.2%	39 75.0%	71
	58	52	110

While the F0 could equally rise or fall from the FP to the first PHRASE for answers to yes/no questions, the answers to open questions showed a significantly higher share of rising movements. The Bernese subjects demonstrate the same tendency, yet without significant differences.

Additionally, a possible connection between the question types and the F0 pattern (low-high-low or declination pattern) was tested. It turned out that answers to yes/no questions have a similar rise or fall from FP to PHRASE for both groups. Differences can be detected in answers to open questions (cf. Table 4), however, where speakers with a declination pattern do not show a preference for rise or fall, while speakers with a low-high-low pattern prefer a rise. This means that the speakers who prefer a declination pattern answer yes/no and open questions similarly, while the subjects that adhere to the low-high-low pattern prefer a fall in F0 from FP to PHRASE when answering yes/no questions.

Table 4: Contingency table of rise or fall from FP to PHRASE by F0 Pattern for open questions

Count Col %	Speakers with a declination pattern	Speakers with a low-high-low pattern	
fall	25 53.2%	21 24.4%	46
rise	22 46.8%	65 75.6%	87
	47	86	133

4.2.5 Summary and discussion

The above results can be summarized as follows: we encounter two patterns of F0 from FP-PHRASE-REST. On the one hand, there is a gradual declination pattern in F0, while on the other hand there is a low-high-low pattern in F0. This pattern could not be attributed to either Valais dialect or Bernese dialect. Yet it can be said that, in any case, utterance-initial phrases are significantly higher than the remaining phrases. Concerning the type of FP, results show that every type of FP could either be higher or lower than the first PHRASE but that a simple '*vowel*', *ja+* or *also*, were predominantly followed by a rise to the PHRASE, while *mm* was followed by a downstep to the PHRASE. Speakers of both dialect groups show the same behavior. Answers to open questions are normally realized with a rise from FP to the first PHRASE, while answers to yes/no questions are equally realized with a rise or a fall. This distinction is significant for Valais speakers. While Bernese speaker show the same conduct, it is not statistically significant.

It seems that the rise or fall from FP to PHRASE depends on the literally-uttered type of question, yet only in correlation with the speakers' dialect background. However, the type of F0 pattern he or she adheres to generally seems to play an equally important role. With respect to the first discovery, one can, of course, only conjecture. It could be that the Valais speakers' pitch range, which is wider than other Swiss-German (including the Bernese) dialects, is one of the causes for the significant differences of the Valais speakers vis-à-vis the Bernese. The second finding may be an issue for conversation analysts in that the drawing of conclusions based on the observation of fundamental frequency changes in utterance-initial positions is exacerbated. The F0 pattern and the interpretation thereof may be due to individual differences or differences related to the speaker's specific dialect background.

5. Conclusion

The present study on prosodic aspects of utterance-initial phrases and utterance-initial filled pauses has touched upon an area of research that, in the context of Swiss German and German in general, has largely been unprecedented. First, it was shown how the speakers use filled and empty pauses differently. The key hypothesis was whether utterance-initial phrases show differences in prosody from the rest of the utterance. This hypothesis can be verified for the present sample. The two speaker groups show significant differences in F0 between the FP, the PHRASE, and the REST. A low-high-low pattern as well as gradual declination pattern in F0 was detected. These patterns cannot be ascribed to regional variation nor to the speakers' sex; yet the number of pauses and their duration do indicate regional differences.

The repercussions of these findings for our current NSF project are considerable. If an intonation and/or timing model of each of the dialects is established, these utterance-initial phrases must either be excluded from the model, so as not to skew its explanatory power, or they must be accounted for in detail and included within the models in a sensible way. From a general prosodic perspective, these findings underline the fact that models of prosody should take into consideration the context of an utterance within the conversation. The models ought to be adapted accordingly. From a CA perspective, these results show that it is sensible to have precise prosodic information, as minor differences on the suprasegmental level may alter the meaning of an utterance. Further, the interpretation of phonetic information must be viewed with respect to regional differences, as speakers embedded in one dialectal context can make use of phonetic distinctions differently than speakers of another dialect area, with a different communicative behavior, do. In addition, it would be interesting to examine individual choices of phonetic patterns.

It is evident that this piece of research only scratches the surface. A thorough analysis of our data from a conversation analytic point of view would allow for the assignment of different communicative functions to various filled pauses, which would in turn provide further explanatory power to different phonetic patterns. Also, the different functions of the questions and answers within the communicative context could be considered. Despite this need for further research within our project, the paper at hand shows a possible scenario for further work on the interface between phonetics and conversation analysis.

Adrian Leemann
University of Tokyo
Hirose & Minematsu Labs
School of Engineering
New Bldg. No.2 103C1
7-3-1 Hongo, Bunkyo-ku
Tokyo 113-8656, Japan
leemann@gavo.t.u-tokyo.ac.jp
adrian.leemann@isw.unibe.ch

Beat Siebenhaar
Universität Leipzig
Institut für Germanistik
Beethovenstraße 15
04107 Leipzig
Germany
siebenhaar@uni-leipzig.de

Linguistic Patterns in Spontaneous Speech, 147-166
2009-1-025-007-000269-1

Linguistic Patterns Detected Through a Prosodic Segmentation in Spontaneous Taiwan Mandarin Speech

Yi-Fen Liu
National Tsing Hua University

Shu-Chuan Tseng
Academia Sinica

This paper proposes that spontaneous speech, segmented into perceptually coherent prosodic constituents, is able to provide plentiful linguistic information in which clear patterns can be observed. We present pioneering studies with empirical and quantitative evidence, supporting the notion that prosodic units can be useful for the automatic processing of spontaneous speech. High inter-labelers' consistency proves the applicability of human prosodic segmentation. A series of results on spontaneous Taiwan Mandarin speech suggest that linguistic patterns found in different linguistic aspects can in theory be used for processing and understanding spontaneous speech. In an automatic POS tagging experiment, it is demonstrated that transcripts with annotations of prosodic boundaries achieved a slightly better performance than the original transcripts with only the speaker turn annotation. Making use of prosodic boundaries, we can deal with the problem of disfluency more directly. With regard to lexical and discourse cue phrases, we also found them produced frequently and regularly at the prosodic boundaries.

Key words: prosodic segmentation, labeling, POS tagging, cue phrases

1. Introduction

The content of spontaneous speech is composed of words just like written texts. But well-formed phrases, clauses, and sentences are not always used in spontaneous speech. Especially in conversation, a number of means other than the "words" may help achieve an active communication such as mimic, gesture, and prosody. Spontaneous interaction between conversation partners determines how, when, and if sentences are to be completed. Therefore, how to reprocess spontaneous speech content into well-formed sentences is an essential issue and task in developing algorithms for processing spontaneous speech. This is also because spontaneous speech contains disfluency, repetition, and abridged sentences (Shriberg 1999, Tseng 2006a). This paper introduces the notion of prosodic units as an intermediate unit to investigate spontaneous speech. Prosodic units were perceptually identified and a high inter-labelers' agreement was achieved. Importantly, these perceptually identified prosodic units are useful for automatic POS tagging, producing better results than the un-segmented turn transcription texts. Furthermore, prosodic units are marked at the boundary by particular words. These can be discourse items specifically found in spontaneous speech, or items which reflect important syntactic positions such as sentence- or clausal-initial and clause-final ones. This pioneering and experimental work uses prosodic units for segmenting spontaneous speech and takes into account syntactic and lexical notions for language modeling on discourse structure. Results introduced in this paper clearly support the notion that prosodic units are significant in many aspects of language processing and useful for spontaneous speech processing.

1.1 Intonation units

The idea that spoken utterances can be phrased or grouped into smaller units has been proposed in various studies. Among them, intonation unit (IU), primarily defined as a unit presenting a piece of meaning/concept, has been widely used in the studies of conversation analysis. IU, viewed as a special case of phonological phrases (Selkirk 1984), is a sequence of words combined under a single, coherent intonation contour, often separated by a pause or marked by a lengthening of the final syllable, a shift upward in overall pitch level at every IU beginning, or a perceived loudness (Chafe 1994, Du Bois et al. 1993). The main criterion for identifying an IU is that it should be perceptually judged as an intonationally coherent unit, often regarded as a reflection of concept. Intonation units are based on prosodic characteristics of spoken utterances, mainly intonation, also suggesting a possible relationship between prosodic units to other aspects of language, such as syntax and semantics. So, the prosody-syntax interface may be observed through studies on the correspondence between prosodic units (PU) and grammatical units (GU), e.g. phrases or clauses (Croft 1995, Tao 1996, Park 2002).

Given that intonation units are practically utilized as intermediate units for conversation analysis, an interesting question arises. Can intonation units also contribute to spontaneous speech processing and improve the results of automatic speech recognition of spontaneous speech? If yes, how can it be incorporated into a model of speech processing? Would a prosodic model be able to segment spontaneous speech into pieces of concept? The work done by Shriberg et al. (2000) and Hirschberg et al. (2004) suggest that cues on prosody are highly informative for speech segmentation and recognition. The models performed even more efficiently while combined with lexical information. On the basis of these previous research results, we suggest using prosodic units, defined as a perceptually coherent prosodic constituent, to segment our spontaneous Mandarin data.

1.2 Punctuation marks and cue phrases

Fine-grained and advanced works on processing and understanding written texts have been done in the field of natural language processing during the past two decades. Syntactic processing such as word segmentation and Part-of-Speech (POS) tagging for Mandarin Chinese have been developed and achieved fast processing time and high performance (Tsai & Chen 2003, Xia & Cheung 2006). For discourse structure processing, punctuation marks (comma, period…etc.) are often useful features for targeted category association. Two rhetorical parsers utilizing punctuation marks and lexical cue phrases (‘because’, ‘for example’ … etc.) at sentence boundary in texts obtain a high precision rate on rhetorical relation assignment; one for English (Marcu 2000); one for Chinese (Cheng et al. 2006). For spoken language, it is a challenging task to decode the phonetic information and then process the syntactic and discourse contents of spontaneous speech eventually. The function of punctuation marks in written texts are to a high degree similar to that of prosodic units in spontaneous speech, as they all provide structural information for segmenting the content of the texts or the speech.

2. Labeling prosodic units in spontaneous Mandarin speech

This section describes the details of the data, followed by a brief introduction to the operational principles for labeling prosodic units. Results of an inter-labelers’ consistency experiment will be presented subsequently.

2.1 Data

For the labeling consistency experiment, the data produced by one female speaker in a one-hour long conversation are used. In total, 583 speaker turns are processed in the first processing stage, equivalent to 4,101 words and 5,917 syllables. The data are extracted from the Mandarin Conversational Dialogue Corpus (MCDC), collected in a Chinese conversational corpus project (Tseng 2004). Detailed information about the corpus and transcription convention can be found at the website http://mmc.sinica.edu.tw. Because the data are long, free conversations, a wide variety of spontaneous speech phenomena are marked. In particular, prosodic variations are rich. Some of the prosodic features can be captured in the issue of intonation units (Chafe 1994, Tao 1996), such as pitch reset, prolongation and pauses. Although these types of cues have previously been applied to multiple speakers’ data, these prosodic cues also work well as in the case of single speaker’s data.

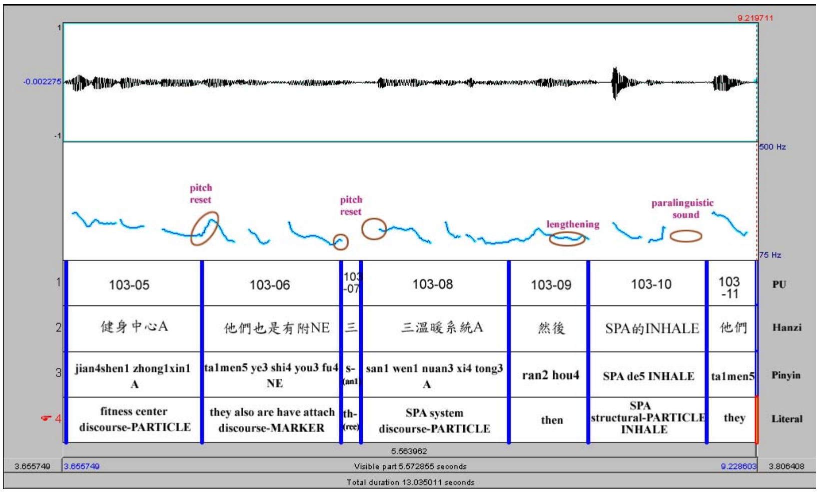
2.2 Principles for labeling prosodic units

A prosodic unit is defined as a perceptually coherent prosodic constituent. A number of the prosodic cues characterized in ToBI and IU are also adopted in our work, coherent contour type, pitch reset, final syllable lengthening, and disjunction of adjacent words (like break, pause, and laughter). In our labeling guidelines, we add one more feature to identify prosodic units (PU): tempo alternation. Supported from the studies on spoken Czech and Mandarin in which speech tends to start fast at the beginning of IU and to end slowly (Dankovičová 1997, Tseng 2006b), temporal variability proves to be a useful cue for unit boundary identification.

If the labelers perceive a coherent prosodic constituent with the help of the following features, they add a PU boundary in the PU tier in addition to the content tier, as illustrated in Fig. 1. Please note that tonal contrast is not a valid principle for identifying PU (however, sometimes it is difficult to distinguish tone from intonation in spoken Mandarin).

a. **Pitch reset**: a shift upward in overall pitch level. In other words, a new prosodic unit may begin with a pitch value higher than the ending pitch value of the previous prosodic unit. While applying this definition, we sometimes encountered with difficulties resulting from lexical tones. When a new prosodic unit begins with a syllable associated with a high-beginning tone, it is hard to distinguish the PU effect from the tone effect.
b. **Lengthening**: lengthening of syllables, changes in duration. Usually, when a syllable is lengthened, a kind of prosodic ending effect will be perceived. But when the lengthening cue is not clear enough, and the whole stretch of speech is more likely to be one single coherent intonation contour, the entire speech stretch will be annotated as one single prosodic unit.
c. **Alternation of speech rate**: changes in rhythm within the same speaker turn. In Taiwan Mandarin, it is often observed that speakers begin their prosodic units with a faster tempo. Especially when the initial words are highly frequent function words or connectives such as “then” and “so”. In this case, a syllable merger often occurs.
d. **Occurrences of paralinguistic sounds**: disjunction or disruption of utterances such as pauses, inhalation, and laughter. Pauses are the most salient cue for identifying prosodic units. Approximately 40 percent of the prosodic units are marked by final pauses. Pauses are further classified into three types: short break (labeled as BREAK), longer pause (labeled as PAUSE) and silence (labeled as SILENCE).[1]

[1] Details please refer to Tseng (2004).



1	103-05	103-06	103-07	103-08	103-09	103-10	103-11	PU
2	健身中心A	他們也是有附NE	三	三溫暖系統A	然後	SPA的INHALE	他們	Hanzi
3	jian4shen1 zhong1xin1 A	ta1men5 ye3 shi4 you3 fu4 NE	s-(an1)	san1 wen1 nuan3 xi4 tong3 A	ran2 hou4	SPA de5 INHALE	ta1men5	Pinyin
4	fitness center discourse-PARTICLE	they also are have attach discourse-MARKER	th-(ree)	SPA system discourse-PARTICLE	then	SPA structural-PARTICLE INHALE	they	Literal

Figure 1: Prosodic units in spoken Mandarin

The prosodic units illustrated in Fig. 1, processed by PRAAT (software for analyzing phonetics, very often used by phoneticians), are part of a female speaker turn. The content is illustrated below with the respective principles (in parentheses) used for the identification.

PU: 103-06	他們	也	是	有	附	NE[2]	**(Pitch reset)**
	ta1men5	ye3	shi4	you3	fu4	NE	
	they	also	are	have	attach	discourse-MARKER	

PU: 103-07	三	**(Pitch reset)**
	s-(an1)	
	th-(ree)	

PU: 103-08	三溫暖	系統	A	**(Pitch reset)**
	san1wen1nuan3	xi4tong3	A	
	SPA	system	discourse-PARTICLE	

PU: 103-09	然後	**(Lengthening)**
	ran2hou4	
	then	

PU: 103-10	spa	的	INHALE	**(Paralinguistic sounds)**
	steam bath	de5	INHALE	
	steam bath	structural-PARTICLE	INHALE	

[2] All discourse particles and discourse marker are transcribed with capital letters (Tseng 2004, 2006b).

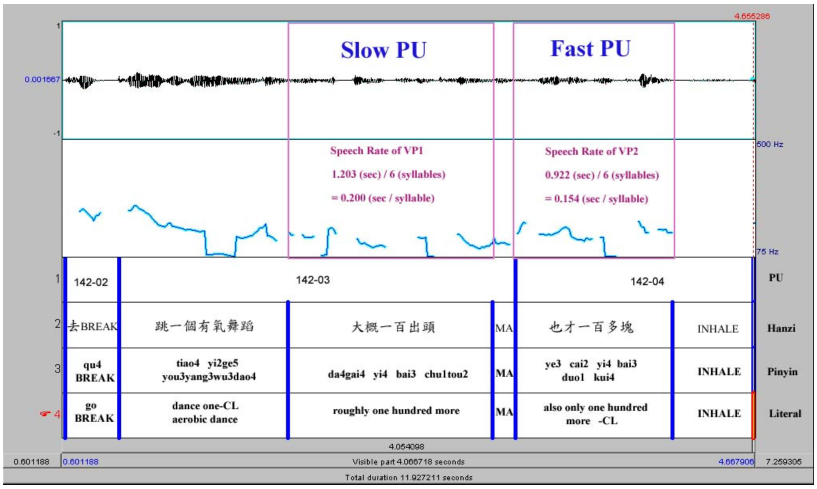


Figure 2: Prosodic units in spoken Mandarin

In Fig. 2, the ending boundary of PU 142-03 is perceived with the help of an upward shift in pitch and an abrupt change in speech rate. The word chunks in PU 142-04 are spoken much faster than similar word chunks in the previous PU.

PU: 142-02	去	BREAK	**(Paralinguistic sounds)**
	qu4	BREAK	
	go	BREAK	

PU: 142-03	跳	一個	有氧舞蹈	[$_{\text{VP1}}$	大概	一百出頭]	
	tiao4	yi2 ge5	you3 yang3 wu3 dao3		da4 gai4	yi4 bai3 chu1 tou2	
	dance	a classifier	aerobic dance		roughly	a bit more than one hundred	
	MA						**(Pitch reset, speech rate)**
	MA						
	discourse-PARTICLE						

PU: 142-04	[$_{\text{VP2}}$也	才	一百多塊]	INHALE	**(Paralinguistic sounds)**
	ye3	cai2	yi4 bai3 duo1 kui4	INHALE	
	also	only	more than one hundred dollars	INHALE	

2.3 Labeling consistency

For the inter-labelers' consistency experiment, we used the speech produced by a female speaker as the training material. In total, the speech of 150 entire speaker turns was labeled in terms of the principles defined for prosodic units by three professional labelers simultaneously. After three stages of annotation and discussion, a version of the prosodic segmentation of the 150 speaker turns was finalized. The precision rate was over 90% for all three labelers (Table 1). Table 2 shows that over 80% of labeled PU-final boundaries are consistently recognized by all three labelers. As we considered this to be acceptable, additional data were labeled independently.

Table 1: Precision rate of prosodic segmentation

Turn101-150	Labeler-01	Labeler-02	Labeler-03
# of PUs labeled	210	217	213
# of finalized PUs	218	218	218
# of correctly labeled PU-final boundary compared with finalized PUs	196	207	195
Precision rate (%)	**93%**	**95%**	**92%**

Table 2: Inter-labelers' consistency

Turn101-150	Labeler-01	Labeler-02	Labeler-03
# of PUs labeled	210	217	213
# of consistent PU-final boundary	178	178	178
Consistent Rate (%)	**85%**	**82%**	**84%**

2.4 Single speaker's dataset

The first dataset consists of one speaker's speech produced in four different scenarios. In Academia Sinica, four spoken corpora of Taiwan Mandarin were collected. In the MCDC, a speaker talked with a stranger in a free conversation, whereas in the MTCC and MMTC the speaker talked with a familiar person in a topic-oriented and task-oriented corpus setting, respectively. The speaker was subsequently asked to read news items. The labeling results are summarized in Table 3.

Table 3: Single speaker's dataset

Speaking situation	Corpus	# of turns	# of PUs	# of words	# of syllables
Free conversation with stranger	MCDC	583	**1,506**	4,104	5,917
Topic-oriented conversation with a familiar person	MTCC	64	**412**	1,582	2,271
Task-oriented conversation with a familiar person	MMTC	47	**114**	306	467
Read speech	READ	1	**71**	326	565
Total		695	**2,103**	6,318	9,220

2.5 Multiple speakers' dataset

The second dataset consists of 16 speakers' speech, extracted from the MCDC. This dataset was generated by another project on directional complements in spoken Taiwan

Mandarin, which consist of all complete speaker turns containing directional complements. In Table 4, the annotation result of prosodic units in this dataset are summarized.

Table 4: Multiple speakers' dataset

Speaker	Gender	Age	# of PUs	# of words	# of syllables
MISC-07	Female	29	333	1,156	1,717
MISC-08	Male	25	1,037	4,163	6,105
MISC-09	Female	37	269	1,213	1,800
MISC-10	Male	35	301	1,126	1,594
MISC-11	Female	16	377	1,630	2,414
MISC-12	Female	17	155	633	949
MISC-15	Male	40	760	3,227	4,622
MISC-16	Female	46	511	1,888	2,720
MISC-23	Female	30	144	645	932
MISC-24	Female	35	1,461	8,084	11,762
MISC-25	Male	35	686	2,871	4,343
MISC-26	Male	23	702	2,727	4,091
MISC-57	Male	43	554	2,769	4,078
MISC-58	Female	45	676	2,844	4,181
MISC-59	Female	37	227	975	1,458
MISC-60	Male	24	368	1,574	2,348
Total			8,561	37,525	55,114

3. Sociolinguistic features on prosodic segmentation

3.1 Multiple speakers' dataset: PU size

The size of prosodic units in words and the general statistics are shown in Table 5. Some prosodic units like one-word response fillers, the filler MHM, are excluded to obtain a better measurement of unit size in the mean length. As a result, all the B columns show that speakers share the same preference in terms of the length of words in prosodic units, as the means center in a narrow range. On average, the number of words per PU is from 3.5 (words/per PU) to 5.6 (words/per PU). The reports (Chafe 1994, Tao 1996) on the mean length of an EIU (English intonation unit) and a MIU (Mandarin intonation unit) are 4.8 and 3.5 words/per IU respectively, which fall into a similar range. It also shows that perceptually judged prosodic phrasing may not only simplify automatic speech recognition but may also capture the types and the length of language processing units across different languages.

Table 5: Unit size in the multiple speakers' dataset
(A = overall data, B = data excluding one-word response fillers)

Speaker	**# of PUs**		**# of words**		**Mean (words/PU)**		**Median(words/PU)**	
	A	B	A	B	A	B	A	B
MISC-07	333	323	1,156	1,146	3.5	**3.5**	3	**3**
MISC-08	1,037	1,030	4,163	4,156	4.0	4.0	3	3
MISC-09	269	269	1,213	1,213	4.5	4.5	4	4
MISC-10	301	301	1,126	1,126	3.7	3.7	3	3
MISC-11	377	364	1,630	1,617	4.3	4.4	4	4
MISC-12	155	148	633	626	4.1	4.2	3	3
MISC-15	760	755	3,227	3,222	4.2	4.3	4	4
MISC-16	511	507	1,888	1,884	3.7	3.7	3	3
MISC-23	144	144	645	645	4.5	4.5	3.5	3.5
MISC-24	1,461	1,452	8,084	8,075	5.5	**5.6**	5	**5**
MISC-25	686	684	2,871	2,869	4.2	4.2	4	4
MISC-26	702	688	2,727	2,713	3.9	3.9	3	3
MISC-57	554	550	2,769	2,765	5.0	5.0	4	4
MISC-58	676	669	2,844	2,837	4.2	4.2	4	4
MISC-59	227	226	975	974	4.3	4.3	3	3
MISC-60	368	367	1,574	1,573	4.3	4.3	4	4
Total	8,561	8,477	37,525	37,441				

Furthermore, we categorize all the identified prosodic units from multiple speakers in terms of gender and age. The distribution of prosodic unit size in words is shown in Fig. 3. Because the data are free conversations, the size of a PU varies to a great extent. A general declination in unit size from one word to 24 words is observed. The declination is obvious in both gender and age. But no clear distinction in terms of the PU size is found between male and female speakers. In addition, no differences were found across generations. Interestingly, we found that the majority of the PUs contains no more than five words. PUs longer than five words make up only 30% in total.

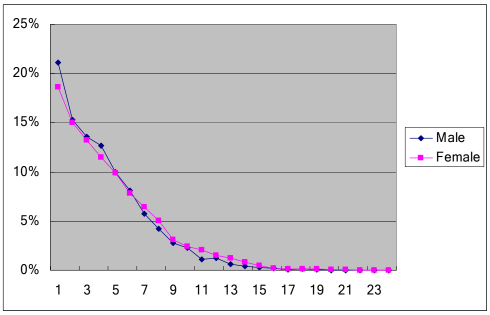


Figure 3a: Unit size in gender

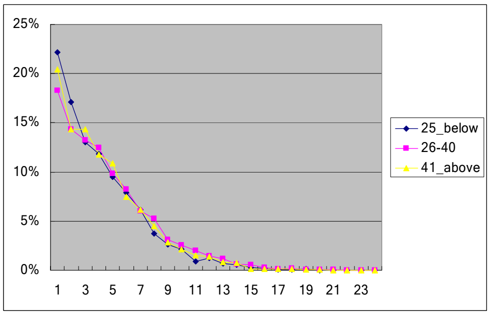


Figure 3b: Unit size in age

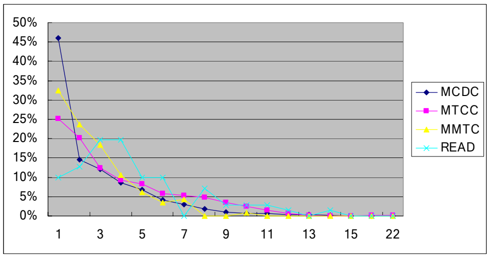
3.2 Single speaker's dataset: PU size

The same analysis on unit size is undertaken for the data produced by one female speaker in different speaking situations. As shown in Table 6, the speaker's behavior of speech production is quite different. While speaking to a stranger, the prosodic unit size averages at 3.2 words; however, while speaking to a friend, the mean length of prosodic units is much longer, 4.2 words. Moreover, when she gives direction instructions in a task-oriented conversation to her friend, the mean length of prosodic units is reduced to 2.8 words; but raised to 4.6 words when she reads the allotted paragraphs with clear-cut punctuation marks. Thus, the result suggests that the unit size distribution may provide clues to who the speaker is talking with. Interestingly, the correlation between unit size and speaking situations provides empirical evidence that sociolinguistic behavior is also reflected through the length of prosodic units. The size of prosodic units in some sense reflects to what extent the speakers strengthen themselves to plan and organize their speech to make it coherent and complex, both semantically and prosodically.

Table 6: Size of prosodic units of a single speaker in different speaking situations (A = Overall data, B = Data excluding one-word response fillers)

Corpora	**# of PUs**		**# of words**		**Mean (words/PU)**		**Median (words/PU)**	
	A	**B**	**A**	**B**	**A**	**B**	**A**	**B**
MCDC	1,506	1,154	4,104	3,752	2.7	**3.2**	2	**3**
MTCC	412	365	1,582	1,535	3.8	**4.2**	3	**3**
MMTC	114	105	306	297	2.7	**2.8**	2	**2**
READ	71	71	326	326	4.6	**4.6**	4	**4**
Total	2,103	1,695	6,318	5,910				

Both graphics in Fig. 4 are similar and drawn from the same datasets produced by the female speaker. They both depict the proportion between the unit size in words and different speaking situations, although we exclude prosodic units consisting of one-word response fillers such as MHM in Fig. 4b, because the one-word response fillers are produced very often which may affect the unit size distribution greatly. The data in Fig. 4 shows that the speaker produces less fragmentary prosodic units but longer, complete prosodic units in read speech. In the other three corpora, most prosodic units are one-word or two-word units. When talking with a stranger, frequent responses may indicate politeness. Therefore, one-word PUs in the MCDC data are found more frequently than in the MTCC. However, with the exception of the one-word PUs, the MCDC and MTCC share a similar pattern of distribution. The Map Task data have more short PUs than the others, as the speaker gives instructions, rather than statements in the MMTC scenario.



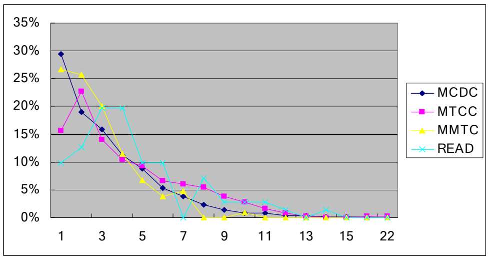


Figure 4a: Unit size, overall data

Figure 4b: Unit size, without one-word fillers

4. Syntactic processing: automatic word segmentation and POS tagging

4.1 Can prosodic segmentation help in automatic syntactic processing?

Prosodic units can be consistently recognized by professional labelers. But can they contribute to automatic syntactic processing to help decode the content? This issue can be investigated by applying automatic word segmentation and POS tagging system to the PU character sequences and the speaker TURN character sequences separately.

The original transcripts contain the orthographic transcription without word boundaries and any punctuation marks. They are only sequences of characters, separated by speaker turns. The PU annotated transcripts add PU boundaries to the original TURN transcripts. A comparative study on the results of these two data types can serve to illustrate the role prosodic segmentation plays in NLP research. For NLP, transcription of the whole speaker's turn is the first available text obtained from the audio data.

It is important to note that Chinese does not use blanks to separate words, nor are there clear morphological criteria to define the main verbs of sentences. In addition, sentences or utterances are not practical for the purpose of NLP work, because spontaneous speech often contains interruptions and disfluencies. Therefore, in order to test the validity of prosodic segmentation in syntactic processing, we applied the automatic word segmentation and POS tagging system[3] developed for the Academia Sinica Balanced Corpus (CKIP 1995) to the original, unprocessed TURN transcriptions and to the PU annotated transcriptions. Our purpose was to evaluate which advantages and disadvantages the prosodic segmentation will cause with regard to the automatic syntactic processing.

[3] The online CKIP tagging system can be found at http://ckipsvr.iis.sinica.edu.tw.

Table 7: Results of syntactic processing on TURN and PU

Consistenct	**Inconsistenct**	
	Word segmentation	POS tagging
36,140 (96.31%)	419 (1.12%)	966 (2.57%)
96.31%	**3.69%**	

Table 7 shows that out of the 37,525 words in the PU annotated transcription only 1,385 (3.69%) words are processed differently, compared with the result of the original TURN transcription. Four hundred and nineteen of them result from word segmentation difference; 966 occurrences are tagged with different POS. The POS system of SIMPOS_19[4] is adopted for running the POS tagging experiment, which is a revised version of the CKIP simplified POS_13. The remaining 96% of the words are consistently tagged by the CKIP system. This result supports the notion that the prosodically identified PU can serve as an intermediate unit between words and speaker turns, since the task of the automatic POS tagging of the PU annotated transcription is as good as the original TURN transcription. Furthermore, we want to analyze the result of inconsistently tagged words to see which version of the transcriptions works better.

4.2 Preference for prosodic segmentation

Table 8: Inconsistency on different units

Types of PU	**Inconsistency on word segmentation**				**Inconsistency on SIMPOS tagging**			
	both incorrect	**TURN correct**	**PU correct**	**Total**	**both incorrect**	**TURN correct**	**PU correct**	**Total**
Disfluency	16 (7.62%)	8 (4.76%)	193 (**87.62%**)	217 (100%)	7 (7.45%)	33 (35.11%)	54 (57.45%)	94 (100%)
Boundary effect	12 (6.42%)	36 (19.27%)	154 (74.31%)	202 (100%)	38 (4.36%)	187 (21.44%)	647 (**74.20%**)	872 (100%)
Disf + BE	28 (7.01%)	44 (12.15%)	**347** **(80.84%)**	419 (100%)	45 (4.66%)	220 (22.77%)	**701** **(72.57%)**	966 (100%)

In Table 8, 1,385 words are tagged differently. 311 among them occur in prosodic units which can be further classified as a result of disfluency. As indicated in the statistics, 80.84% of the differently word-segmented results and 72.57% of the differently

[4] For the mapping table between our revised SIMPOS_19 and CKIP SIMPOS_13, please see the Appendix. There are six more SIMPOSs in our revision than in CKIP SIMPO_13. In our revision, for specific POS, we prefer keeping the original POS unchanged. Four of them (bold-faced in Appendix) are changed. Two of them (underlined in Appendix) are added to obtain the speech style of spontaneous speech. For detailed description POS classification and annotation criteria, please refer to http://ckipsvr.iis.sinica.edu.tw.

POS-tagged results are correct in the PU annotated transcription. In general, this result shows that segmenting turns into smaller prosodic units works better than the original TURN transcription in the task of syntactic tagging. With regard to the occurrences distribution of those two types of inconsistencies, the statistics shows that 647 (74.20%) out of 966 POS tagging inconsistencies are correctly tagged in the annotated PU transcription. A clearer distributional preference is shown in Fig. 5a. The word segmentation inconsistencies show that smaller units have clearly more advantages, especially in the case of disfluency, since 193 (87.62%) are correctly segmented in the PU annotated transcription (also shown in Fig. 5b).

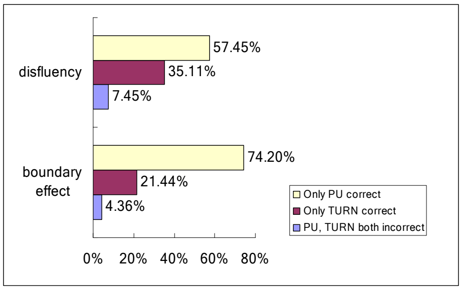


Figure 5a: POS inconsistency

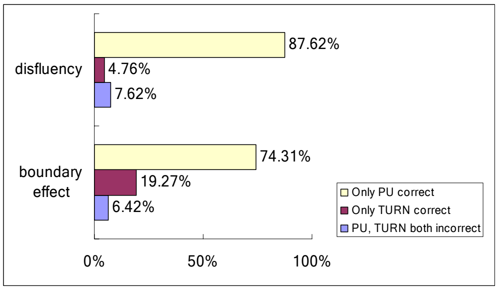


Figure 5b: Word segmentation inconsistency

In addition to disfluency, we are also able to observe how words are used differently in spoken discourse and how they change due to the variations, through a prosodic segmentation. To take the word *jie2guo3* (result) as an example, the same occurrences of *jie2guo3* are tagged differently in the PU and TURN transcripts. While in the TURN transcription, *jie2guo3* is segmented and tagged as a noun (result), it is tagged as an adverb (as a result) in the PU transcript. This means that the function as a discourse marker of certain lexical words in a spoken discourse will become clearer if we study them in the framework of prosodic units.

4.3 Lexical preference at prosodic boundaries

We have shown that prosodic segmentation works better than the un-annotated TURN transcriptions for spontaneous speech. We now look at whether regular lexical patterns exist at prosodic boundaries. If such patterns are found, then this will aid in detecting prosodic boundaries by means of lexical items. As the proportion in Table 9 indicates, most of the discourse items (including fillers, markers, interjections and particles) tend to be the boundary items, either in the initial or final position. Some lexical items, such as *le5*, *ma1*, *er2yi3* and *de5* occurring in the unit-final position, are often identified as sentence-final particles. In the CKIP tagging system, the tagger treats the prosodic unit boundary as a sentential boundary. Out of the 37,525 words produced by all 16 speakers, the most frequently used conjunctions make up only 51 types, compared with 288 different types of adverbials. Calculating the occurrences, 65.22% of conjunctions prefer unit-initial positions; most of them are listed in Table 9. Moreover, some adverbials, as the six listed in Table 9, also prefer the initial position in prosodic units. As their tokens are too few, adverbials are not as good a cue for boundary identification as conjunctions. The proportion shows that the rest of the syntactic categories seldom occur at a boundary, but often occur in a medial position. It also suggests that the prosodic marking at PU boundary functions in a similar way to the punctuation marks (comma, period and so on) in written texts.

Table 9: Unit position preference on fillers, discourse markers and 19 SIMPOS

Categories	Representative discourse / lexical items	PU-initial	PU-medial	PU-final
Fillers	MHM(HMHM), NHN(HNHN), UHN(HN)	**79.07%**	－	－
Markers	NA NE NAGE NEGE NEIGE ZHEGE SHEME SHENME	**68.87%**	－	－
Discourse interjections (I)	啊 (A), 喔 (O), 嗯 (EN)	**71.48%**	－	－
Discourse particles (T)	啊 (A), 啦 (LA), 喔 (O), 嘛 (MA), 哪 (NA), 吧 (BA), 呀 (YA)	－	－	**75.11 %**
Sentence final particles (T)	了 (le5), 嗎 (ma1), 而已 (er2 yi3), 的 (de5)			
Conjunctions (C)	要不然 (otherwise), 不然 (otherwise), 或者 (or), 不管 (no matter), 因爲 (because), 所以 (so), 可是 (but), 但是 (but), 不過 (however), 如果說 (if), 連 (and even)	**65.22%**	－	－
Adverbials (ADV)	然後 (then), 其實 (actually), 結果 (as a result), 也許 (maybe), 甚至 (even), 尤其 (especially)	－	**65.60%**	－
SHI, V_2, Vt, Vi, P, ASP, DE,CL, DET, N, POST, A, FW, b			**64.66%**	

5. Cue phrases in prosodic segmentation

5.1 Discourse and lexical cue phrases

In the studies of discourse structure for written texts, lexical cue phrases such as conjunctions and adverbs are often mentioned with regard to their function of marking specific locations relevant to the discourse structure. In spoken language, especially Mandarin Chinese, discourse markers and particles are highly essential as far as the discourse segmentation is concerned. Therefore, we analyze two different kinds of cue phrases: discourse and lexical cue phrases. Table 10 lists the most frequent discourse items identified at prosodic boundaries in our own data. These discourse items are associated with specific discourse functions such as hesitation, doubt expression, uncertainty etc. They are associated with the interaction between the conversation participants. Because they often occur at prosodic boundaries, they are not only related to the discourse structure, but also to the prosodic structure.

Table 10: List of discourse cue phrases

Types of DCP	Discourse items
Fillers	MHM MHMM MHMHM MHMHMHM NHN NHNN NHNHN NHNHNHN UHN UHNN UHNHN UHNHNHN
Markers	NA NE NAGE NEGE NEIGE ZHEGE SHEME SHENME
Particles	A AI AN BA E EI EN EP EIN HAI HAN HE HEI HEINHEN HO HON HWA O ON OU LA LIE LEI LO MA NOU NO WA SAI YA YE YEI YI YOU

In addition to the discourse cue phrases mentioned above, we also analyze the often used lexical cue phrases which are specifically related to the rhetorical functions. Rhetorical structure has been studied in the framework of the dominance tree structure (Marcu 2000) and cross-dependent graphics (Wolf & Gibson 2005) for written texts. In this study, we want to examine whether they are also relevant in the context of conversation. Cheng et al. (2006) identified a set of lexical items which served as crucial cues for the coherence of conversations in their rhetorical parser for Mandarin texts, mainly adopting the lexical cue phrases used in Cheng & Tian (1992), which are associated with eight frequently identified rhetorical relations, as shown in Table 11. The underlined pairs often appear in sequence in both texts and speech, e.g. on the one hand—on the other hand. In the following analysis, we will study whether there is any relationship between the rhetorical functions and the prosodic structure of these lexical cue phrases.

Table 11: List of lexical cue phrases

Types of RR	Lexical items
Joint	同時 (meanwhile), 同樣 (in the same way), 另外 (besides), 此外 (in addition), 也 (also), 一方面 (on the one hand)_另一方面 (on the other hand), 第一 (first)_第二 (second), 首先 (first of all)_其次 (secondly), 不在於 (is not in)_是在於 (but in)
Contrast	但是 (but), 但 (but), 可是 (but), 可 (but), 相反 (on the contrary), 然而 (however), 幸而 (fortunately), 不過 (however), 其實 (actually), 儘管 (even though), 儘管如此 (even though), 儘管這樣 (even though), 雖然 (although)

Sequence	後來 (then)
Alternative	還是 (or)，或者 (or)，要麼 (or)
Elaboration	而且 (moreover)，並且 (moreover)，並 (moreover)，還 (moreover)，更 (moreover)，甚至 (even)，何況 (furthermore)，況且 (moreover)，這就是說 (this means)，也就是說 (in other words)，所謂 (this is what we called)，意思是 (this means)
Cause-effect	因此 (therefore)，所以 (thus)，結果 (as a result)，由此看來 (concluded from this)，因爲 (because)，原因是 (the reason is)，由於 (due to)
Condition	那樣 (in that way)，否則 (otherwise)，不然 (otherwise)，那麼 (in that way)，要不 (if not so)，如果 (if)，如果這樣 (if so)，如果不這樣 (if not so)，如果不那樣 (if not)，假使 (if)_才 (only)，要是 (if)_就 (then)，不管這樣 (regardless)，這樣 (if so)，只要這樣 (only if)，只有這樣 (only if)，除非這樣 (only if)
Example	如 (like)，像 (like)，例如 (such as)，比如 (such as)，譬如 (for example)，舉例來說 (for instance)

5.2 Cue phrases in prosodic segmentation: multiple speakers' dataset

Fig. 6 shows the proportion distribution of the discourse/lexical items in the multiple speakers' dataset in terms of their position within prosodic units. Fillers are often associated with understanding or backchannel functions between the conversation partners and they often occur in a single prosodic unit. Discourse markers are often used while speakers hesitate for what message they are going to deliver next and they are often located in the unit-initial position. Also in the case of the single speaker's dataset, as shown in Fig. 7, discourse markers tend to be located at boundaries (initial, or final), rather in the middle of prosodic units. Generally, the particles used for indicating a speakers' attitude prefer the unit-final position. The preference for all three groups of discourse cue phrases to occur at prosodic boundaries is observed in both datasets. As this result indicates, discourse items are highly correlated with prosodic boundaries. Lexical cue phrases associated with rhetorical functions such as cause-effect, contrast and sequence relations frequently appear in the unit-initial position, as shown in Fig. 6. These lexical items mark both the rhetorical functions and the prosodic structure. This strengthens their function as cue phrases for discourse structure in conversation. However, for the other types of lexical cue phrases, no clear relationship between the rhetorical function and the prosodic marking can be found.

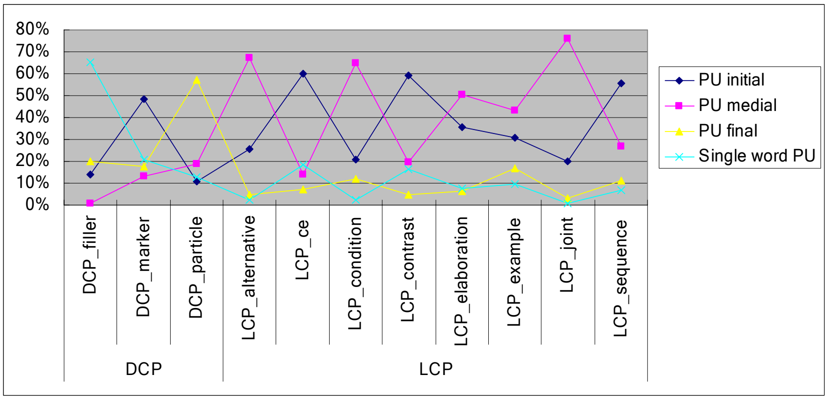


Figure 6: Distribution on cue phrases—multiple speakers

5.3 Cue phrases in prosodic segmentation: single speaker's dataset

As in the multiple speakers' data, discourse cue phrases are highly prosodically marked. In Fig. 7, lexical cue phrases associated with the cause-effect and contrast functions are still clearly prosodically marked, whereas the function 'sequence' is not prosodically marked as we observed in the multiple speakers' data. However, the items associated with the example relation are frequently used at the beginning of a prosodic unit. This may suggest that the use of prosody and rhetorical functions are sometimes speaker-specific. We need further investigation to study the lexical items which do not occur at unit boundaries to better understand their use in conversation.

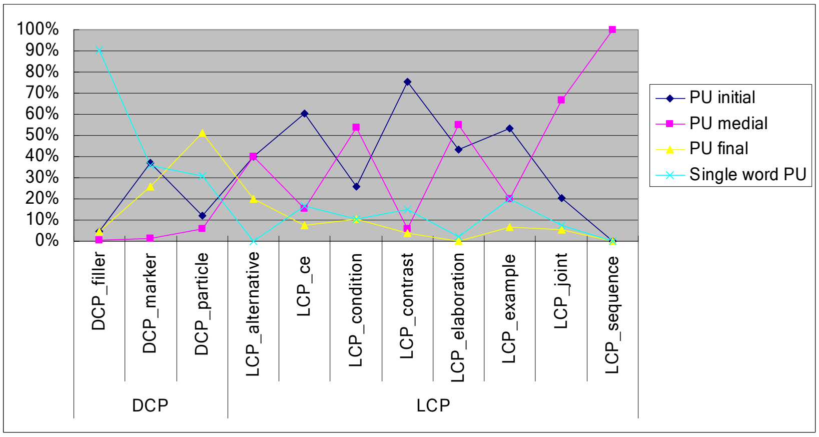


Figure 7: Single speaker's dataset

Prosodic units represent a kind of prosodic segmentation in spoken language. Rhetorical functions are associated with concepts that should have a certain interactive effect on the conversation partners. There should be a kind of interrelationship between prosodic phrasing and rhetorical functions. The surface-form-based approach for a rhetorical parser has been proposed and proved useful by Marcu (2000) for written texts. In our study, we have shown that a number of cue phrases indicating rhetorical functions are also prosodically indicated.

6. Conclusion

We tried to find an intermediate unit between words and sentences which can be operationally defined and practically applied to understand and process spontaneous speech. This paper proposes the notion of prosodic units for this purpose, for which a high inter-labelers' agreement was achieved. It has also been shown that this prosodic unit works well in an automatic POS tagging experiment. PU boundaries are often marked by specific syntactic categories and lexical items. The result also demonstrates that PU is directly associated with important discourse phenomena in spontaneous speech such as disfluency, discourse particles and markers, and fillers. PU is a unit that can be identified by applying relevant acoustic-prosodic features to improve automatic speech recognition algorithms. An efficient parser can be developed by adopting the PU-related syntactic categories and cue phrases to deal with spontaneous speech. Currently, the data is being labeled in terms of the syllabic boundaries to obtain more acoustic-prosodic cues which should be relevant to PU boundaries.

Appendix

CKIP POS_48	**CKIP SIMPOS_13**	**Revised SIMPOS_19**	**CKIP POS_48**	**CKIP SIMPOS_13**	**Revised SIMPOS_19**
A	A	A	VA	Vi	Vi
b	-	b	VAC	Vt	Vt
Caa	C	C	VB	Vi	Vi
Cab	POST	POST	VC	Vt	Vt
Cba	POST	POST	VCL	Vt	Vt
Cbb	C	C	VD	Vt	Vt
D	ADV	ADV	VE	Vt	Vt
DE	**T**	**DE**	VF	Vt	Vt
Da	ADV	ADV	VG	Vt	Vt
Dfa	ADV	ADV	VH	Vi	Vi
Dfb	ADV	ADV	VHC	Vt	Vt
Di	ASP	ASP	VI	Vi	Vi
Dk	ADV	ADV	VJ	Vt	Vt
FW	FW	FW	VK	Vt	Vt
I	**T**	**I**	VL	Vt	Vt
NAV	NAV	NAV	**V_2**	**Vt**	**V_2**
Na	N	N			
Nb	N	N			
Nc	N	N			
Ncd	N	N			
Nd	N	N			
Nep	DET	DET			
Neqa	DET	DET			
Neqb	POST	POST			
Nes	DET	DET			
Neu	DET	DET			
Nf	M	CL			
Ng	POST	POST			
Nh	N	N			
SHI	**Vt**	**SHI**			
T	T	T			
P	-	P			

Yi-Fen Liu
Institute of Information Systems and Applications
National Tsing Hua University
101, Sec. 2, Kuang-fu Road
Hsinchu 300, Taiwan
yifenliu@gmail.com

Shu-Chuan Tseng
Institute of Linguistics
Academia Sinica
130, Sec. 2, Academia Road
Nankang, Taipei 115, Taiwan
tsengsc@gate.sinica.edu.tw

Linguistic Patterns in Spontaneous Speech, 167-192
2009-1-025-008-000270-1

Prolongation of Clause-initial Mono-word Phrases in Japanese[*]

Yasuharu Den
Chiba University

In this paper, we focus on a particular type of prolongation in Japanese, prolongation at the end of clause-initial mono-word phrases (CIMWPs). We first investigate various syntactic and acoustic features in relation to the prolongation (PR) rate of CIMWPs. We show that the PR rate of CIMWPs is affected by word class, duration of the preceding pause, and presence of a succeeding pause, but that there is no reliable effect of clause complexity. We then take up a subclass of CIMWPs, the clause-initial conjunction *de*. In modeling the duration of *de*, we show that succeeding pauses and fillers, as well as preceding pauses, are important factors.

Key words: prolongation, clause-initial mono-word phrases, clause-initial conjunction, mixed-effects model

1. Introduction

In spontaneous speech, speakers may prolong their speech segments anywhere in an utterance. This phenomenon has recently attracted considerable attention in the study of spontaneous speech (Eklund 2001, Den 2003, Lee et al. 2004). Eklund (2001), studying prolongation in Swedish, showed that the occurrence of prolongations varied as a function of phonological type, position in the word, lexical factors, and word class. He also examined cross-linguistic factors of prolonged speech, comparing Swedish data with Tok Pisin data. He showed that while there were similarities between the languages, there were also significant differences at the segmental and distributional levels. Lee et al. (2004) studied prolongation in spontaneous Mandarin and showed that prolongations were often found in word-final, phrase-final, and utterance-medial positions. They also

[*] This paper is a revised version of a paper presented at the International Symposium on Linguistic Patterns in Spontaneous Speech 2006, and includes a supplementary §4, which was originally written for a project supported by a Grant-in-Aid for Scientific Research (B) (No.16320060) supported by the Japan Society for the Promotion of Science, led by Kikuo Maekawa during the period of 2004-2006. I thank the anonymous reviewers and the attendants of the symposium for their useful comments and discussions.

found that prolongations were particularly frequent in transitive verbs, adverbs, nouns, and particles, in contrast to intransitive verbs, aspectual adverbs, and adjectives, which were rarely, or never, prolonged.

Based on a corpus analysis of spontaneous monologues in Japanese, Den (2003) reported that Japanese speakers use several strategies in prolonging speech segments. These strategies are summarized as follows:

- Japanese speakers frequently prolong utterance-initial, mono-moraic words, which are typically discourse markers and which are distributed complementarily to fillers.
- They sometimes prolong the final vowels of utterance-initial content words, including nouns and demonstrative nouns serving as the topic of an utterance.
- They sometimes prolong the final vowels of phrase-final content words, and often prolong the final vowels of phrase-final function words, the latter being typically followed immediately by a silent pause.
- They sometimes prolong the vowels at the disruption point of word fragments, which are often restarted from the beginning immediately after the disruption, resulting in a word repetition.

Though exhaustive, these results do not make clear what factors are involved in the production of prolonged speech segments.

In this paper, we focus on a particular type of prolongation in Japanese, and investigate syntactic and acoustic factors that may affect the production of such prolonged speech. The type of prolongation we take up here is prolongation at the end of a clause-initial mono-word phrase (CIMWP). A CIMWP is a word that appears at clause-initial position and that constitutes a phrase by itself. An example is *de* in the following utterance.

De<H>	minna-	ga	iku-	no-	nara,	zya	watasi-	mo-	to-	yuu	koto-	de, ...
and	everyone	NOM	go	NM	if	then	I	too	QT	say	thing	be

And, if everyone goes, then I do too, . . .

<H> at the end of the first word *de* indicates a non-lexical prolongation. Because the conjunction *de* by itself constitutes a ‘bunsetsu’ phrase in Japanese, this example forms a typical instance of prolongation of CIMWPs, which we particularly focus on in this study. According to Den (2003)’s findings, prolongations of CIMWPs are rather frequent; the first and the second of the aforementioned strategies include prolonged CIMWPs, and the last one is also concerned with them.

In elucidating the factors affecting the production of prolonged CIMWPs, it is worth discussing the significance, in communication, of the clause-initial position. In communication with others, due to the pressure of temporal imperative, speakers may sometimes start a constituent that has not been fully construed (Clark & Wasow 1998). This would be more likely to happen at clause-initial position where the cognitive load of planning the content, and formulating the structure, of an utterance is severe. On these occasions, speakers need to deal with an upcoming possible delay in speech production. They may use a filler or preliminary commitment to the next constituent, which is involved in a repeated word, in order to inform listeners of an anticipated long pause (Clark & Wasow 1998). Den (Den & Clark 2000, Den 2001, 2007) showed that prolongation at the interruption point of a repeated word in Japanese, particularly one appearing at the beginning of an utterance, may serve as such a trouble-announcing function. It may be natural that we assume prolonged speech in general, whether it involves a repeated word or not, also has a similar function.

Several predictions follow from this assumption. First, if prolonged CIMWPs have a trouble-announcing function, they would be more frequent when the utterance, or clause, to be construed is more complex. This is true of fillers at clause-initial position; Watanabe et al. (2006) showed that the rate of fillers at clause-initial position tends to be higher when the clause gets longer. Second, as Watanabe et al. (2006) also reported, the presence of a deep syntactic and discourse boundary enhances the rate of clause-initial fillers appearing at such boundaries. This tendency may also be replicated in clause-initial prolongations, since the presence of a syntactic and discourse boundary may be associated with the cognitive load of speech production. Third, if prolongation has something in common with a filler concerning a trouble-announcing function, the two forms of disfluency would be complementarily distributed, a tendency that has already been suggested by Den (2003). That is, when a CIMWP is prolonged at the end, it would be less likely to be immediately preceded or followed by a filler. Finally, if prolonged CIMWPs have a trouble-announcing function, frequent pauses would be observed after them.

The aim of the current study is to verify these predictions by testing the effects of various syntactic and acoustic factors, and to examine relative contributions of these factors by means of a multivariate statistical model. Our focus on prolongation of CIMWPs, rather than on prolongation of clause-initial words or phrases in general, is for the purpose of equating the structural configuration in which it occurs. Prolonged CIMWPs are structurally uniform in the sense that they all appear at clause-initial as well as phrase-final positions, both being significant conditions for prolongation to occur. Furthermore, prolonged CIMWPs occupy about 60% of prolonged clause-initial words in our data. The dominance of prolonged CIMWPs among the entire prolongations at clause-initial position suggests a good starting point for a deep and precise understanding of the phenomena.

The rest of the paper is organized as follows. In §2, we describe the spoken language corpus used in the current study as well as the annotations made on it. In §3, we examine several syntactic and acoustic features in relation to the prolongation rate of CIMWPs, based upon the subjective judgment of the transcribers, which was already supplied in the corpus. In §4, we further our analysis by making use of duration data, focusing on a subclass of CIMWPs, the clause-initial conjunction *de*. We construct a statistical model to predict the duration of *de* at clause-initial position using several syntactic, acoustic, and discourse features. In §5, we discuss the implications of our findings.

2. Data and annotation

2.1 Data

We analyzed a part of the *Corpus of Spontaneous Japanese* (Maekawa 2003). From among the entire data, we selected 177 monologues in the Core data, which come with hand-corrected annotation of clause units, 'bunsetsu' phrases, long and short unit words, and phonetic segments. The data were classified into two groups according to recording source: academic presentation speech (APS) and simulated public speech (SPS). APS is the live recording of academic presentations for several academic societies covering the fields of engineering, social science, and humanities. SPS, on the other hand, is studio recorded speeches of paid layman speakers, of about 10-12 minutes, on everyday topics presented in front of a small audience and in a relatively relaxed atmosphere. The speakers were 24 females and 46 males in APS and 54 females and 53 males in SPS, ranging in age from their early twenties to late sixties, with a median at the mid thirties. The speech data amounted to ca 40 hours, and the morphological data to ca 410,000 short unit words excluding fillers. Table 1 shows the summary statistics of the data relative to speech type and gender of the speaker.

Table 1: Summary statistics of the data

	APS		SPS		
	Female	Male	Female	Male	Total
No. of sessions	24	46	54	53	177
Duration (hrs)	7.0	11.8	9.8	10.1	38.6
No. of clauses	3141	5375	4645	5030	18191
No. of phrases	29736	49596	42761	43118	165211
No. of words	75564	127030	103445	107509	413548

2.2 Annotation

For the selected 177 monologues, the boundaries of clauses, phrases, and words were provided in the corpus. Words were segmented in two different ways: *short unit words* (SUWs) and *long unit words* (LUWs) (Ogura et al. 2004), where SUWs correspond roughly to entry words in a Japanese dictionary and LUWs cover compound words. For these words, parts of speech, based on a traditional Japanese grammar, were given. The starting and ending times of the words in the recorded discourse were also precisely indicated by phonetic segmentation labels, except for those words concealed for privacy. Fillers, such as *eeto* and *ano*(*o*), occupied the position of genuine words in the original data, but we treated them as part of the pause preceding a word.[1]

Phrases were segmented in terms of 'bunsetsu,' which is a widely used notion in Japanese linguistics. A 'bunsetsu' phrase consists of a content word (LUW) possibly followed by one or more function words including those that have been grammaticalized (Ogura et al. 2004). We used these phrases only for the purpose of identifying clause-initial mono-word phrases.

Clauses were identified as *clause units*, which were segmented based mainly on syntactic criteria (Takanashi et al. 2004). For clause units, boundary types, according to the form of the final word in a clause, were given. Three boundary types were distinguished: (i) absolute boundary, at which the clause unit ends with a verb, adjective, or auxiliary verb in a conclusive or imperative form, a final particle, or a quotative particle without being followed by a main verb, (ii) strong boundary, at which the clause unit ends with a coordinate conjunctive particle, and (iii) weak boundary, at which the clause unit ends with a subordinate conjunctive particle.

For a part of the data (15 speakers in APS and 25 speakers in SPS), discourse boundaries were also provided in the corpus. The segmentation of discourse was performed by trained annotators based on Grosz & Sidner (1986)'s theory of discourse structure, and the purpose of each discourse segment, or sub-segment, was supplied (Takeuchi et al. 2004). These annotations will be used in the analysis presented in §4.

Non-lexical prolongations of vowels, at any position in a word, were indicated by an <H> tag in transcriptions. The assignment of these tags was based on the transcribers' intuition, and we did not revise them. In this paper, we focus on prolongations occurring at the final mora of a clause-initial word that constitutes a phrase by itself, and do not look at those occurring at word-internal position or in a non-clause-initial word.

[1] All the fillers, marked by an F tag in transcriptions, including those in a phrasal form, e.g, *eeto-desu-ne*, were treated in the same way. Only when they appeared at the end of a clause were they regarded as genuine words.

3. Analysis 1: Prolongation rate and syntactic and acoustic features

3.1 Outline of the analysis

In this section, we examine several syntactic and acoustic features in relation to the prolongation (PR) rate of CIMWPs. First, we compare the PR rate across the word classes of the CIMWPs and across the boundary types of the preceding clause units. Word classes, such as content and function words, have been found to influence the PR rate (Eklund 2001, Den 2003, Lee et al. 2004). Since we are concerned here only with clause-initial position, at which a content word usually appears in Japanese, we consider a rather limited range of word classes. The boundary type of the preceding clause unit may also influence the rate of disfluency. Watanabe et al. (2006) found such an influence for the rate of fillers at clause-initial position.

Second, we investigate the effect of the complexity of clause units. Clark & Wasow (1998) claimed that the complexity of the following constituents does affect speakers' planning load and consequently the ratio of disfluencies. In this line, Watanabe et al. (2006) showed that the rate of fillers at clause-initial position tends to be higher when the clause gets longer. A similar effect may be observed for prolongation. We obtain the complexity of clause units initiated by CIMWPs, and examine whether it correlates with the PR rate of CIMWPs.

Third, we turn our attention to the acoustic environment where prolonged CIMWPs occur. We focus on pauses and fillers that immediately precede or succeed a CIMWP, investigating their correlations with the PR rate of CIMWPs. Den (2003) reported that the prolongation of utterance initial mono-moraic words is distributed complementarily to fillers. If the tendency is replicated in the current study, it would be expected that the rate of prolonged CIMWPs is likely to be negatively correlated with the rate of preceding fillers. Den (2007) showed that pauses at the interruption point of repeated words sometimes become long, insisting that the first token of a repeated word, which is often prolonged at the end, is used by the speaker to inform listeners of upcoming delay in speech production (cf. Clark & Wasow 1998). If prolongation in general has such a trouble-announcing function, it would be expected that prolonged CIMWPs tend to be followed by a pause.

Finally, to determine the relative contributions of these factors, we apply a multivariate statistical model to our data. The dependent variable in our analysis is binary: prolonged or non-prolonged. Thus, an adequate method may be a logistic regression model, which is a variant of the multiple regression model, extended so as to allow dependent variables that obey a binomial distribution. This method, however, is not adequate, because in our data multiple cases were sampled from a single speaker, which violates the assumption of the independence of cases required by the method. For such data, generalized linear mixed-effects models (GLMMs) are applicable. Particularly, we apply a GLMM with binomial distribution and logit link that contains only one random effect representing deviations, due to individual differences, from the overall intercept, also known as a logistic regression model with a random intercept.[2] We generate a regression model that optimally fits the observations in the data, and compare the contributions of various variables by means of a model selection technique.

3.2 Method

3.2.1 Data selection

For the 177 monologues, we extracted instances of CIMWPs in the following way. For each clause unit, the initial word (SUW) was first extracted, and it was then retained as a CIMWP when it constituted a ‘bunsetsu’ phrase by itself. When the preceding or succeeding pause could not be measured, due to the presence of a concealed word for privacy, the instance was excluded from the analysis. This procedure left us 8029 instances of CIMWPs; 3716 instances (53.1 instances per speaker) in APS and 4313 instances (40.3 instances per speaker) in SPS.

3.2.2 Measurements

In this analysis, we used seven independent variables: (i) the word class of the CIMWP, (ii) the boundary type of the preceding clause unit, (iii) the complexity of the current clause unit, (iv) the duration of the preceding pause, (v) the presence of a preceding filler, (vi) the presence of a succeeding pause, and (vii) the presence of a succeeding filler.

In order to decide how many word classes to distinguish, we first obtained the distribution of CIMWPs with respect to their parts of speech. Fig. 1 shows the mean frequency of CIMWPs per speaker (bars) and the number of speakers who produced more than 4 instances of CIMWPs (dots) for each part of speech in APS and SPS. For reliability, we required that the data used for the analysis contained at least 5 instances for each speaker. Based on this result, we decided to distinguish only two word classes: conjunction and others. After removing the data for speakers who had less than 5 instances for either word class, 169 monologues (69 speakers in APS and 100 speakers in SPS) were retained for the analysis.

[2] We used the lme4 package for the statistics language R.

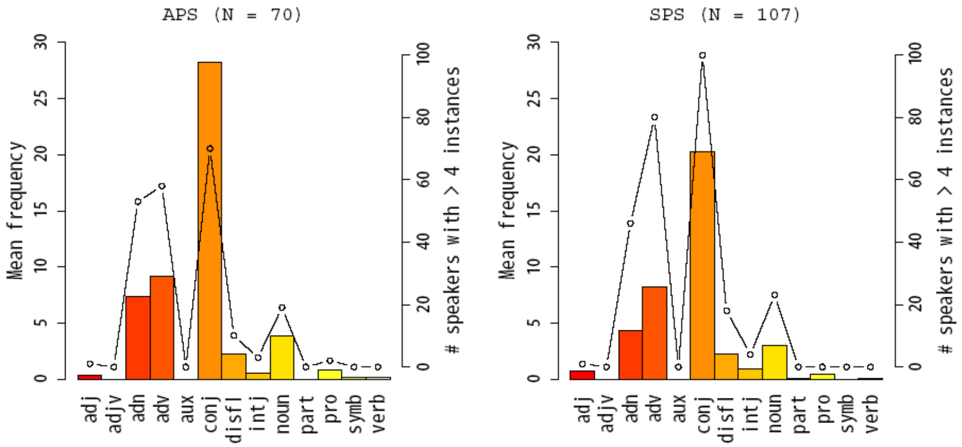


Figure 1: Distribution of CIMWPs with respect to their parts of speech. Bars represent the mean frequency of CIMWPs per speaker and dots represent the number of speakers who produced more than 4 instances.

Next, in order to decide how many boundary types to distinguish, we obtained the distribution of CIMWPs with respect to the boundary types of the preceding clause unit. Fig. 2 shows the mean frequency of CIMWPs per speaker (bars) and the number of speakers who produced more than 4 instances of CIMWPs (dots) for each boundary type in APS and SPS.[3] Based on this result, we decided to distinguish only two boundary types: absolute boundary and others, hereafter referred to as 'sentential' and 'clausal' boundaries, respectively. The criterion of 'more than 4 instances for both boundary types' removed 23 monologues, leaving us 146 monologues (53 speakers in APS and 93 speakers in SPS) for further analysis.

Complexity of the current clause unit was measured by the number of short unit words (SUWs) in the clause unit. There are many other ways to measure clause complexity. For instance, the number of syntactic nodes in a parse tree can be a good measurement of clause complexity. Here we used the number of words as the measure of clause complexity because it is the simplest method and is known to correlate highly with other measurements such as the number of syntactic nodes (Wasow 1997). We also used SUWs, instead of LUWs, in counting the number of words because LUWs are sometimes too long to be

[3] The distributions for APS and SPS differed considerably. The frequency of absolute boundaries was far higher in APS than in SPS. This may be attributed to the difference of the overall distributions of the clause boundary types between the two speech types; due to a formal speaking style, the speakers of APS more often used auxiliary verbs in a conclusive form, e.g., *masu*, *ta*, and *desu*, at the end of a clause unit, whereas the speakers of SPS also used a conjunctive particle *te* very frequently.

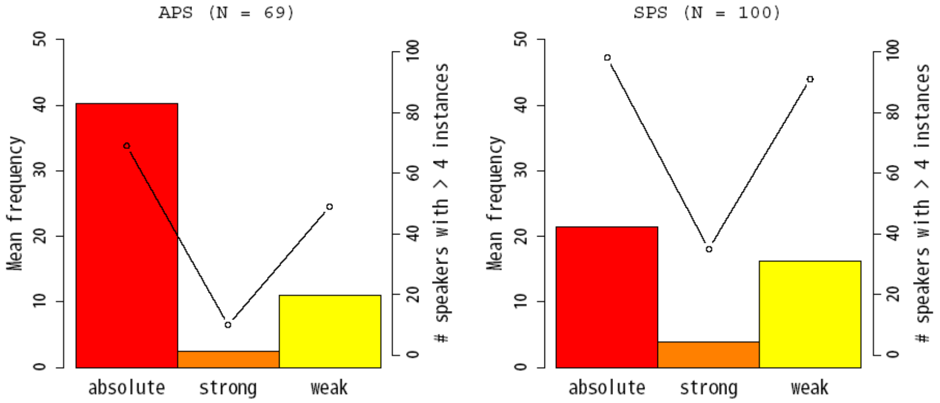


Figure 2: Distribution of CIMWPs with respect to boundary types. Bars represent the mean frequency of CIMWPs per speaker and dots represent the number of speakers who produced more than 4 instances.

viewed as a unit of speech production.[4] In the calculation of clause complexity, we always ignored fillers.

Duration of the preceding pause was measured as the duration of the region between the start of the CIMWP and the end of the preceding word. The values were log-transformed by the formula $\log(x+1)$ when used in statistical analyses. When a filler immediately preceded the CIMWP, its end point was used as the boundary of the pause region. In this case, a binary variable representing the presence of a preceding filler was also set to 1; otherwise, this variable was set to 0.

For the succeeding pause, a binary variable representing the presence of a pause was used, instead of representing the duration by a continuous value. This is because no pauses immediately follow the CIMWP in the majority of the cases (80.0% in APS and 75.1% in SPS), which yielded a very skewed distribution of the durations. Another binary variable was used to represent the presence of an immediately succeeding filler.

The dependent variable was occurrence or non-occurrence of prolongation at the end of the CIMWP, or the rate of prolongation, i.e., the PR rate. When the PR rate was the target of statistical analysis, the values were angular-transformed by the formula $\sin^{-1}\sqrt{x}$.

[4] For instance, *kokuritu-kokugo-kenkyuu-syo* (National Institute for Japanese Language) is a single long unit word, whereas it consists of four short unit words.

3.3 Results

3.3.1 Overall PR rate

For each speaker, we calculated the PR rate as the number of prolonged CIMWPs divided by the total number of CIMWPs. Fig. 3 shows the mean (angular-transformed) PR rate relative to speech type and gender of the speaker. A two-way ANOVA revealed a significant main effect of speech type (*F*(1,142)=46.4, *p*<.001), but no significant main effect of speaker's gender or interaction between the two factors (*Fs* <1). The speakers of SPS produced far more prolonged CIMWPs than the speakers of APS (14.8% vs. 2.5% on average).[5] Taking this into account, we will present the results of the subsequent analyses separately according to the two speech types.

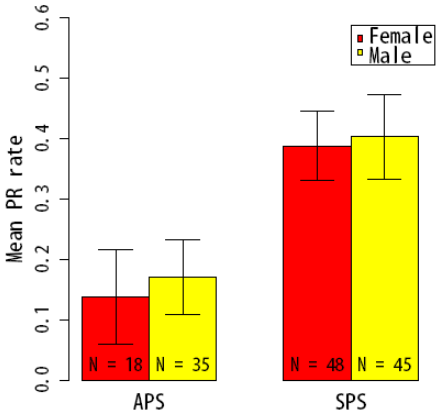


Figure 3: Mean (angular-transformed) prolongation rate of CIMWPs relative to speech type and gender of the speaker. Error bars indicate 95% confidence intervals.

3.3.2 Word class

We calculated the PR rate for each word class, i.e., conjunction and others. Fig. 4 shows the mean (angular-transformed) PR rates relative to the word class of the CIMWP in APS and SPS. Paired t-tests revealed significant differences in the PR rate between the two word classes in both APS (*t*(52)=2.49, *p*<.05) and SPS (*t*(92)=10.98, *p*<.001). The PR rate for conjunction was far higher (2.5% in APS and 23.9% in SPS on average) than that for the other word class (less than 0.7% in APS and 2% in SPS on average).

[5] Throughout the paper, means are first calculated on the transformed scale and then inversely transformed back onto the original scale.

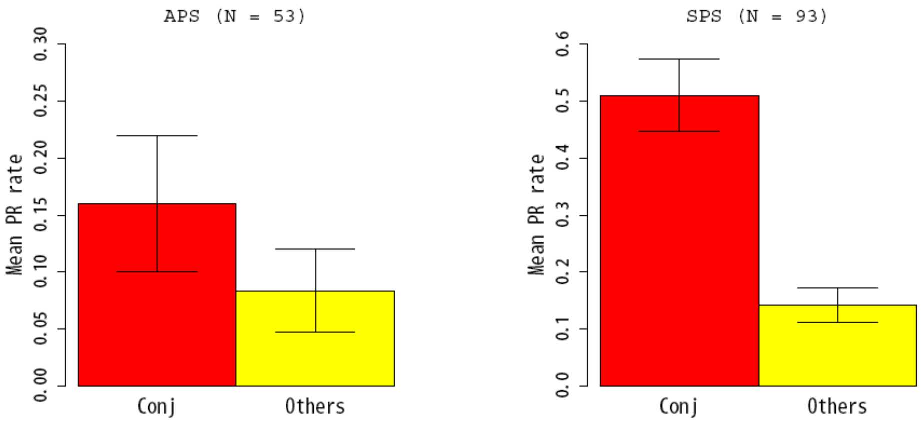


Figure 4: Mean (angular-transformed) prolongation rate relative to the word class of the CIMWP.

3.3.3 Boundary type

We calculated the PR rate for each clause-boundary type, i.e., sentential and clausal. Fig. 5 shows the mean (angular-transformed) PR rates relative to the boundary type of the preceding clause unit in APS and SPS. Paired t-tests revealed a marginal difference between the two boundary types in APS (*t*(52)=−1.87, *p*<.07), and a significant difference in SPS (*t*(92)=−4.23, *p*<.001). The PR rate for sentential boundaries was higher than that for clausal boundaries (2.5% vs. 1.3% in APS and 16.3% vs. 8.5% in SPS on average), although this tendency was less reliable in APS.

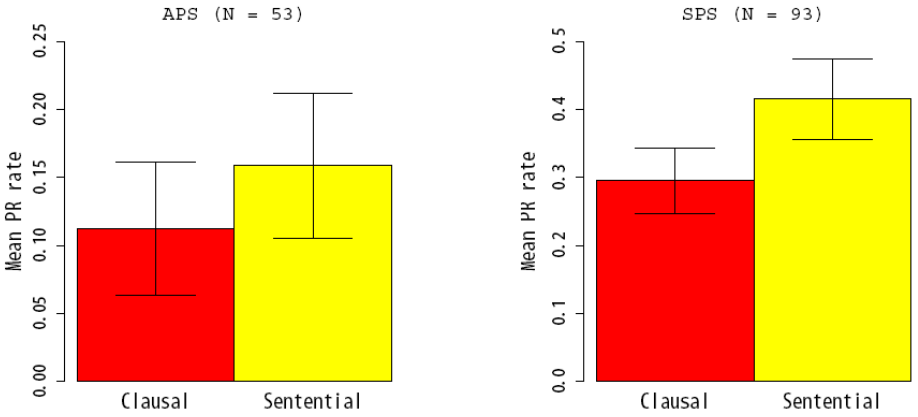


Figure 5: Mean (angular-transformed) prolongation rate relative to the boundary type of the preceding clause unit.

3.3.4 Clause complexity

For each speaker, we obtained the median of clause complexities, measured by the number of SUWs in the clause unit, and correlated it with the PR rate of that speaker. Fig. 6 shows the scatter plots between the median clause complexity and the (angular-transformed) PR rate in APS and SPS. Tests for Pearson's correlation revealed no significant correlation between the median clause complexity and the PR rate in either APS (*r*=.03, *t*(51)=.23, *p*=.82) or SPS (*r*=.10, *t*(91)=.93, *p*=.36).

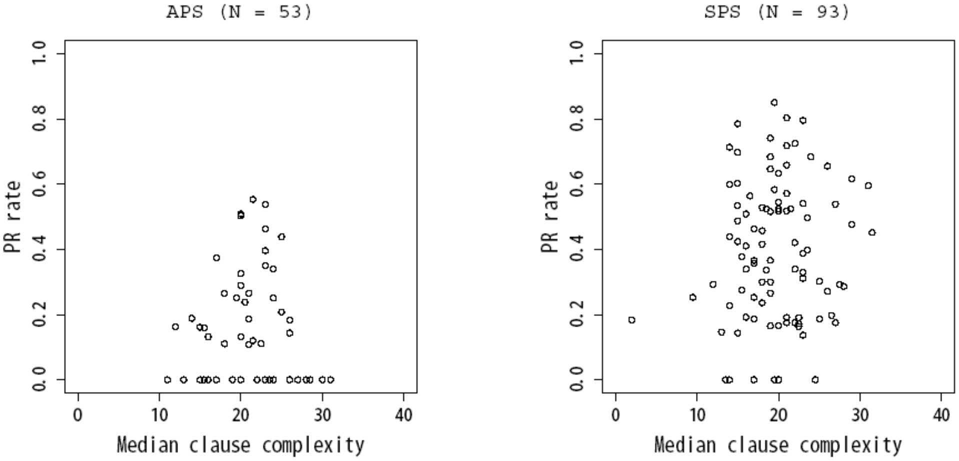


Figure 6: Scatter plot between the median clause complexity and the (angular-transformed) prolongation rate.

3.3.5 Duration of a preceding pause

For each speaker, we obtained the median of the durations of preceding pauses, and correlated it with the PR rate of that speaker. Fig. 7 shows the scatter plots between the median (log-transformed) duration of the preceding pause and the (angular-transformed) PR rate in APS and SPS. Tests for Pearson's correlation revealed weak but reliable correlations between the median duration of the preceding pause and the PR rate in both APS (*r*=.33, *t*(51)=2.50, *p*<.05) and SPS (*r*=.38, *t*(91)=3.93, *p*<.001). Speakers who produced a longer pause, on average, immediately before the CIMWP showed a higher rate of prolonged CIMWPs.

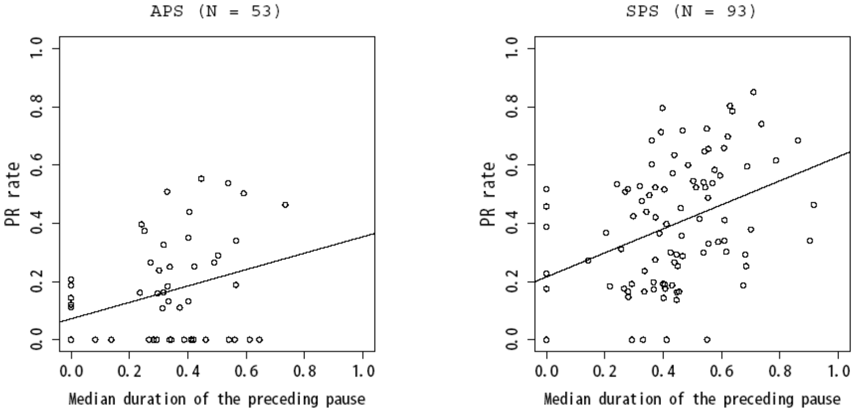


Figure 7: Scatter plot between the median (log-transformed) duration of the preceding pause and the (angular-transformed) prolongation rate. Lines represent the regression lines.

3.3.6 Presence of a preceding filler

For each speaker, we obtained the rate of fillers immediately preceding the CIMWP, and correlated it with the PR rate of that speaker. Fig. 8 shows the scatter plots between the (angular-transformed) rate of preceding fillers and the (angular-transformed) PR rate in APS and SPS. Tests for Pearson's correlation revealed weak and marginal correlation between the rate of preceding fillers and the PR rate in APS (*r*=−.24, *t*(51)=−1.7, *p*<.09), and weak but reliable correlation in SPS (*r*=−.27, *t*(91)=−2.70, *p*<.01). The correlations were negative, meaning that the speakers who produced a filler immediately preceding the CIMWP at a higher rate showed a lower rate of prolonged CIMWPs.

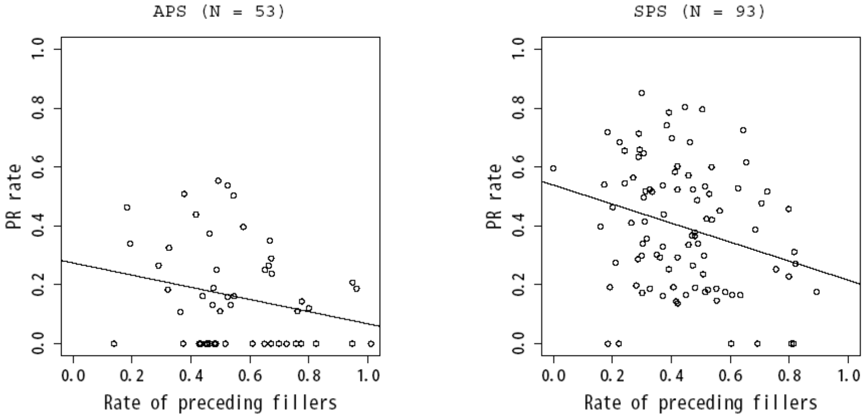


Figure 8: Scatter plot between the (angular-transformed) rate of preceding fillers and the (angular-transformed) prolongation rate.

3.3.7 Presence of a succeeding pause

For each speaker, we obtained the rate of pauses immediately succeeding the CIMWP, and correlated it with the PR rate of that speaker. Fig. 9 shows the scatter plots between the (angular-transformed) rate of succeeding pauses and the (angular-transformed) PR rate in APS and SPS. Tests for Pearson's correlation revealed no correlation between the rate of succeeding pauses and the PR rate in APS (*r*=.08, *t*(51)= −.61, *p*=.55), but a reliable correlation was found in SPS (*r*=.26, *t*(91)=2.58, *p*<.05). The PR rate was higher in SPS when the speaker produced a pause immediately succeeding the CIMWP at a higher rate.

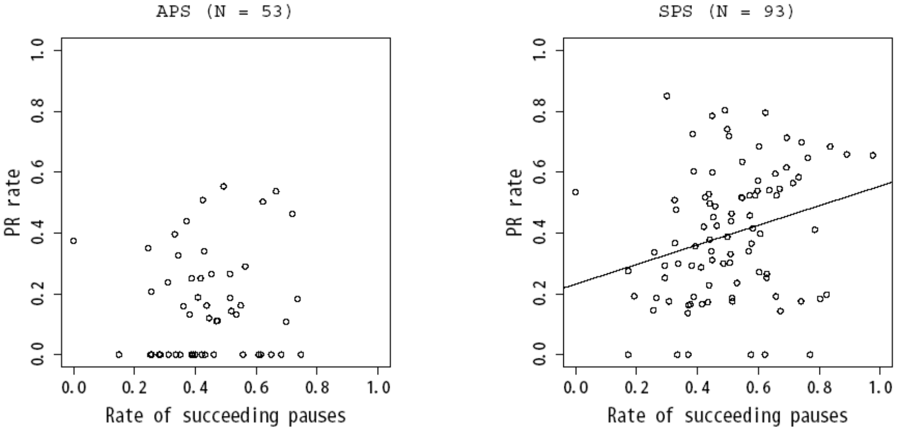


Figure 9: Scatter plot between the (angular-transformed) rate of succeeding pauses and the (angular-transformed) prolongation rate.

3.3.8 Presence of a succeeding filler

For each speaker, we obtained the rate of fillers immediately succeeding the CIMWP, and correlated it with the PR rate of that speaker. Fig. 10 shows the scatter plots between the (angular-transformed) rate of succeeding fillers and the (angular-transformed) PR rate in APS and SPS. Tests for Pearson's correlation revealed no correlation between the rate of succeeding fillers and the PR rate in either APS (*r*=−.03, *t*(51)=−.20, *p*=.84) or SPS (*r*=.09, *t*(91)=.86, *p*=.39).

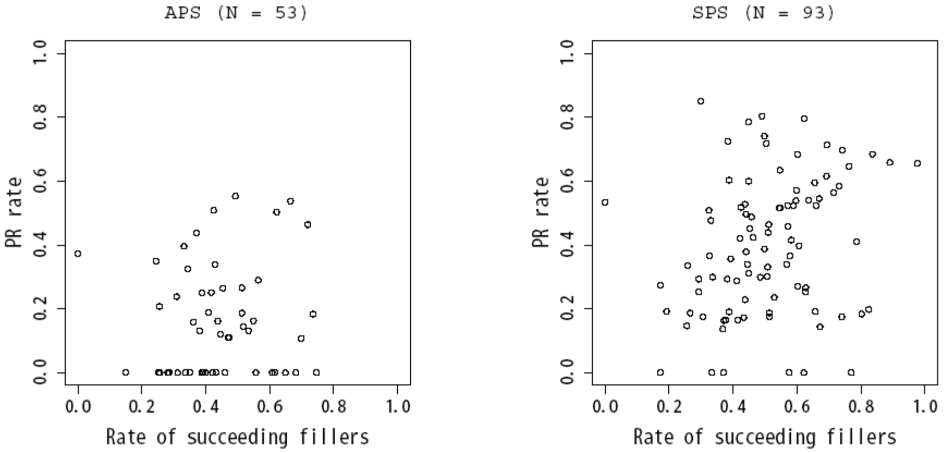


Figure 10: Scatter plot between the (angular-transformed) rate of succeeding fillers and the (angular-transformed) prolongation rate.

3.3.9 Multivariate model

To determine the relative contributions of various variables, we applied a logistic regression model with a random intercept, and compared the contributions of the variables by means of a model selection technique. We used, as explanatory variables, the seven variables that we have thus far analyzed separately. No interaction terms were used, only main effects were considered. We applied two methods of model selections: forward selection and backward selection. In forward selection, starting from the minimal (null) model that contained only an intercept as well as a random intercept, we repeatedly fit a model to the data, adding a term to the model one by one. When no significant improvement in log-likelihood was achieved by the addition of another term, we reached an optimal model. Conversely, in backward selection, starting from the maximal model that contained all main effects and an intercept as well as a random intercept, we repeatedly fit a model to the data, deleting a term from the model one by one. When significant declination in log-likelihood was yielded by the deletion of another term, we reached an optimal model.

Table 2 shows the model selection steps in APS and SPS. Each row provides information about the degree of freedom (df) and the log-likelihood of the fitted model, as well as the χ^2 statistics and the *p* value obtained in the likelihood ratio test between the models in the current and the previous rows. In both speech types, the forward and the backward selections converged on the same optimal model. The optimal models for APS and SPS, however, were different from each other. Both models included the variables concerning word class (`class`), the (log-transformed) duration of the preceding pause (`pre.pause`), and the presence of a succeeding pause (`suc.pause`), but the model for APS, in addition, contained the variable concerning boundary type (`boundary`), while the model for SPS, instead, contained the variable concerning the preceding filler (`pre.filler`). The variables concerning clause complexity (`complex`) and the succeeding fillers (`suc.filler`) did not survive in either model.

Table 2: Model selection steps. The upper half shows the steps for forward selection and the bottom half shows the steps for backward selection.

APS					**SPS**				
Model	df	−2logLik	χ^2	*p*	Model	df	−2logLik	χ^2	*p*
Minimal	2	1131.6			Minimal	2	3468.4		
`+class`	3	1092.4	39.2	< .001	`+class`	3	2977.5	490.9	< .001
`+suc.pause`	4	1063.9	28.5	< .001	`+suc.pause`	4	2908.0	69.5	< .001
`+pre.pause`	5	1057.0	6.9	< .01	`+pre.pause`	5	2866.9	41.1	< .001
`+boundary`	6	1053.2	3.8	< .05	`+pre.filler`	6	2863.0	3.9	< .05
Maximal	9	1046.8			Maximal	9	2856.9		
`-suc.filler`	8	1047.4	0.6	= .43	`-suc.filler`	8	2858.2	1.3	= .26
`-complex`	7	1050.1	2.7	= .10	`-boundary`	7	2860.2	2.0	= .16
`-pre.filler`	6	1053.2	3.1	= .08	`-complex`	6	2863.0	2.7	= .10

Table 3: Estimated coefficients for the optimal models. SE denotes standard error.

APS					**SPS**				
	Coef.	SE	*z*	*p*		Coef.	SE	*z*	*p*
(Intercept)	−4.97	.34	−14.82	< .001	(Intercept)	−4.35	.21	−20.74	< .001
`class=C`	1.35	.26	5.16	< .001	`class=C`	2.32	.15	15.28	< .001
`boundary=S`	−.51	.26	−1.96	< .05	`pre.filler`	.40	.20	1.99	< .05
`suc.pause`	1.04	.19	5.34	< .001	`suc.pause`	.90	.11	8.18	< .001
`pre.pause`	.56	.18	3.06	< .005	`pre.pause`	1.26	.19	6.69	< .001
Random intercept: $\sigma_1^2 = 2.54$					Random intercept: $\sigma_1^2 = 1.04$				

Table 3 shows the estimated coefficients for the optimal models. The rows indicated by `class=C` and `boundary=S` give the coefficients of the dummy variables that represent the value of the word class being 'conjunction' and the value of the boundary type being 'sentential,' respectively. In both speech types, word class had the greatest contribution; the word class being 'conjunction' largely increases the probability of a prolonged CIMWP (odds ratio: 3.9 in APS and 10.2 in SPS). The presence of a succeeding pause also had a relatively great contribution in both cases; when followed by a pause, CIMWPs have more probability of being prolonged (odds ratio: 2.8 in APS and 2.5 in SPS). The duration of the preceding pause contributed to the model considerably in SPS, but the contribution was smaller in APS; the longer the preceding pause, the higher the probability of prolongation. Note, however, that only this variable is continuous, not binary, and, thus, it is difficult to compare the contribution of this variable with those of other variables directly from Table 3. The contributions of the remaining variables, boundary type in APS and presence of a preceding filler in SPS, were not so large, compared to the contributions of word class and the succeeding pause.

3.4 Discussion

We have identified several factors that affect prolongation of CIMWPs in Japanese. Univariate analysis revealed the following:

1. CIMWPs were far more often prolonged when they were conjunctions than when they were of the other word class.
2. CIMWPs were more often prolonged when they occurred at a sentential boundary than when they occurred at a clausal boundary, although the tendency was less reliable in APS.
3. There was no reliable evidence that the clause complexity, measured by the number of words, affects the PR rate of CIMWPs.
4. A longer preceding pause facilitated the prolongation of CIMWPs, whereas a high rate of preceding fillers decreased the PR rate of CIMWPs.
5. There was a reliable tendency in SPS for the PR rate to become higher in proportion to the rate of succeeding pauses, but no such tendency was observed in APS; the rate of succeeding fillers had no effect on the PR rate of CIMWPs for either speech type.

Some of these results coincide with the predictions we formulated in §1, while others do not. The boundary effect, i.e., the hypothesis that the presence of a deep syntactic and discourse boundary enhances the rate of disfluency at such a boundary, was supported, as in the second observation above. In contrast, the complexity effect, i.e., the hypothesis that the more complex an utterance is, the more disfluency is involved in the beginning of the utterance, was not supported, as in the third observation. As for the trouble-announcing function of disfluency—some types of disfluencies can serve as a signal for the speaker to inform listeners of upcoming delay in speech production—, we have predicted that prolonged CIMWPs would be distributed complementarily to fillers, and that they would be often followed by a pause. The former prediction was supported for preceding fillers, as in the fourth observation, but it was not supported for succeeding fillers, as in the fifth observation. The latter prediction was only partly supported for the half of the data, i.e., SPS, as in the fifth observation.

The multivariate analysis also indicated the insignificance of clause complexity and the succeeding filler. These variables did not improve the fitness of the optimal models. Word class and the presence of a succeeding pause were the two most influential variables in our multivariate models. Particularly, the presence of a succeeding pause in APS was shown to have a relatively great contribution by multivariate analysis, although its effect had not been significant in univariate analysis. The result obtained in the univariate analysis concerning the preceding pause was reinforced by the multivariate analysis; the duration of the preceding pause and the PR rate of CIMWPs had a positive correlation. The results for boundary type and the preceding filler seemed contradictory between the univariate and multivariate analyses. In the univariate analysis, sentential boundaries were shown to be more associated with prolonged CIMWPs, whereas the multivariate model for APS estimated a negative coefficient for the variable concerning the sentential boundary, meaning that sentential boundaries are less associated with prolonged CIMWPs. Likewise, the presence of a preceding filler was shown to decrease the PR rate of CIMWPs in the univariate analysis, but the multivariate model for SPS revealed a converse correlation. These discrepancies between the results in the univariate and multivariate analyses may be understood on the basis of the fact that in the multivariate models the contributions of these variables were rather small, and that their effects had relatively large correlations with the effects of other variables; `boundary=S` was correlated with `class=C` at r=−.50 and with `pre.pause` at r=−.22 in APS, and `pre.filler` was correlated with `class=C` at r=.37 and with `pre.pause` at r=.33 in SPS. This suggests that the estimation of the effects of these variables may not be reliable.[6]

These results suggest several important points regarding the function of prolongation. First, prolongation of CIMWPs does not necessarily reflect speakers' cognitive load in planning an utterance, since it is not influenced by the complexity of the clause to be construed. This stands in contrast to fillers, which are indicative of the cognitive load in planning a constituent being produced (Watanabe et al. 2006). Second, it is not clear that the prolonged CIMWPs have a trouble-announcing function, which fillers and repeated words sometimes have. Although they are associated with the presence of a succeeding pause, their relation to the neighboring fillers is not fully elucidated. Third, the positive correlation between the duration of the preceding pause and the PR rate suggests that prolongation of CIMWPs is likely to occur on a certain occasion where the preceding pause tends to be long. Considering that longer pauses tend to occur at a deep syntactic and discourse boundary (e.g., Hirschberg & Nakatani 1996), prolonged

[6] Note also that these variables were the last ones to be brought into the optimal models by the forward model selection procedure (see Table 2).

CIMWPs may be observed more frequently at a discourse boundary than at other places.

Another interesting observation that has yet to be presented is the overall distribution of fillers preceding and succeeding a CIMWP with respect to the word class of the CIMWP. The rates of preceding fillers differ considerably between when they precede a conjunction and when they precede a word of another class. The rate of preceding fillers is far lower before a conjunction than before the other word class (7.2% vs. 51.1% in APS and 2.4% vs. 34.2% in SPS). In contrast, the rate of succeeding fillers is relatively higher after a conjunction than after the other word class (27.8% vs. 14.6% in APS and 25.8% vs. 11.2% in SPS). These observations suggest that conjunctions at clause-initial position are more likely to precede a filler than to be preceded by a filler. Thus, the typical ordering of a conjunction and a filler at clause-initial position is 'a conjunction, followed by a filler,' but not the opposite order.

This ordering standard may help us to understand the results obtained in the current analysis. As discussed above, it seems that prolonged CIMWPs may have some relation with discourse structures. If we assume that prolonged conjunctions at clause-initial position have some discourse function, our ordering standard would naturally follow. Speakers may first produce a prolonged conjunction to carry out some discourse-related function, and then place a filler to manage the problem of the utterance being construed. The opposite order would be unnatural.

To verify this assumption, we should look more closely at a possible relation of prolonged CIMWPs to discourse-level phenomena. Fortunately, some portion of our data include information about discourse structures, which may be useful in searching for the discourse function of prolonged clause-initial conjunctions. This motivates our next analysis.

4. Analysis 2: Duration of the clause-initial conjunction *de*

4.1 Outline of the analysis

One of the drawbacks of the previous analysis is the lack of discourse factors, which may affect prolongation of CIMWPs. The current analysis takes a discourse factor into account by using the discourse structure annotation, which is provided for a part of the corpus. Another draw-back of the previous analysis is its reliance on prolongation tags in transcriptions, which were assigned based on the transcribers' intuition. More objective measurements can be obtained by using phonetic segmentation labels.

In utilizing actual durations of CIMWPs, we must be careful as many factors affect segmental duration. These include type of phoneme, word class, position in the word, and position in the sentence (on Japanese, see e.g., Kaiki & Sagisaka 1992, Campbell 1992). To avoid these influences, we focus on a particular subclass of CIMWPs, the clause-initial conjunction *de*. This is motivated by two reasons. First, occurrences of *de* are extremely frequent at clause initial position; among the 8029 CIMWPs analyzed in the previous section, *de* appears nearly 40% of the time (1415 out of 3716 instances in APS and 1697 out of 4313 instances in SPS). Second, the prolongation of *de* occupies the majority of prolonged CIMWPs in our data. Therefore, the investigation of *de* is suitable for furthering our analysis.

As in the previous analysis, we first examine the effects of each syntactic and acoustic feature separately. We then examine the relative contributions of the features using a multivariate statistical model. In the current analysis, we use a linear mixed-effects model, the simplest case of GLMMs for normally distributed dependent variables. Just as the models used in the previous section, we consider only a random intercept, representing individual differences, as random effects.

4.2 Method

4.2.1 Data selection

Of the 177 monologues in the Core data, 40 monologues have discourse structure annotation; 15 out of 70 speakers in APS and 25 out of 107 speakers in SPS. For these data, we extracted instances of the clause-initial conjunction *de*. To limit our attention to cases where the ordering standard of ‘a conjunction, followed by a filler’ was met, we excluded those cases where *de* was preceded by a filler; there were only 33 such cases. We obtained 833 instances of clause-initial *de*; 417 instances (27.8 per speaker) in APS and 416 instances (16.6 per speaker) in SPS.

4.2.2 Measurements

In this analysis, we used five independent variables: (i) the boundary type of the preceding clause unit, (ii) the complexity of the current clause unit, (iii) the duration of the preceding pause, (iv) the presence of a succeeding pause, and (v) the presence of a succeeding filler. Of the seven independent variables used in the previous analysis, word class, which was fixed to conjunction in the current analysis, and the presence of a preceding filler, whose possibility was excluded in the current analysis, had been removed. The boundary type of the preceding clause unit was extended to include ‘discourse’ boundary. The boundary type of the preceding clause unit was defined to be a discourse boundary when the current clause unit started a new discourse segment, regardless of the depth of the segment or the syntactic form of the preceding clause end.

In order to decide how many boundary types to distinguish, we obtained the distribution of clause-initial *de* with respect to the boundary types of the preceding clause unit. Fig. 11 shows the mean frequency of clause-initial *des* per speaker (bars) and the number of speakers who produced more than 4 instances of clause-initial *de* (dots) for each boundary type in APS and SPS.[7] Based on this result, we decided to distinguish only two boundary types: discourse boundary and non-discourse boundary. The criterion of 'more than 4 instances for both boundary types' left us 17 monologues (7 speakers in APS and 10 speakers in SPS) for further analysis.

Duration of the preceding pause was log-transformed by the formula log(x+α), where α was determined so that it maximized the fitness of the distribution of the transformed values to the normal distribution by the Shapiro-Wilk test of normality; α =.07 for APS and α=.31 for SPS were used.

The dependent variable was the duration of the clause-initial *de*, and was log-transformed by the formula log(x).

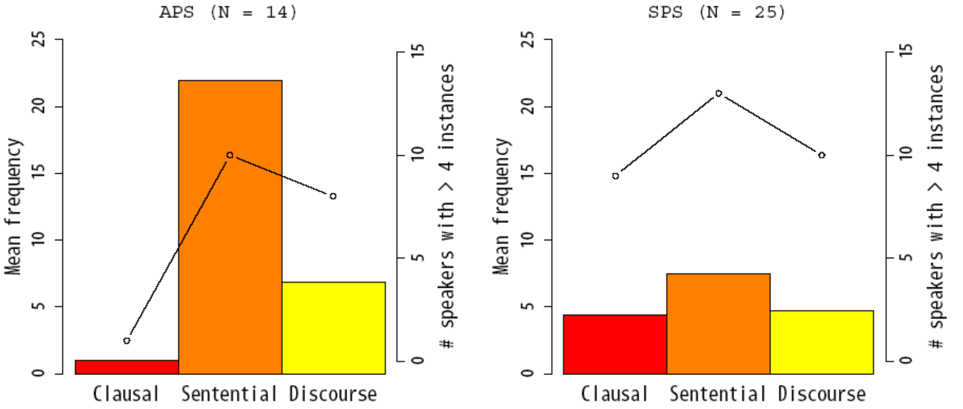


Figure 11: Distribution of clause-initial *de* with respect to boundary types. Bars represent the mean frequency of clause-initial *des* per speaker and dots represent the number of speakers who produced more than 4 instances.

4.3 Results

4.3.1 Effects of individual variables

To determine the effects of individual variables, we constructed a model with the main effect of each variable and an intercept as well as a random intercept, and compared it with the minimal (null) model that contained only an intercept as well as a random intercept. Table 4 shows the comparison between single main effect models and the

[7] The distributions for APS and SPS differed considerably, as was also the case in the previous analysis (cf. Fig. 2).

minimal model. Each row except for the first one corresponds to a single main effect model in which only the main effect of the variable labeled on the row is contained. The values of χ^2 and p are for the comparison with the minimal model by a likelihood ratio test.

The duration of the preceding pause and the presence of a succeeding pause and filler had significant effects in gaining the fitness of the model. The effect of clause complexity, in contrast, was not significant. Boundary type had a significant effect in SPS, but no significant effect in APS.

Table 5 shows the estimated coefficients for the single main effect models that fit the observations significantly better than the minimal model. All the estimated coefficients were positive, meaning that a discourse boundary (in SPS), the presence of a succeeding pause and filler, and a longer preceding pause all contribute to extending the duration of clause-initial *de*.

Table 4: Comparison between single main effect models and the minimal model.

APS					SPS				
Model	df	−2logLik	χ^2	p	Model	df	−2logLik	χ^2	p
Minimal	2	521.3			Minimal	2	603.4		
`boundary`	3	519.2	2.1	= .15	`boundary`	3	596.0	7.3	< .01
`complex`	3	520.8	.46	= .50	`complex`	3	602.1	1.3	= .26
`suc.pause`	3	493.6	27.7	< .001	`suc.pause`	3	560.2	43.2	< .001
`suc.filler`	3	474.6	46.7	< .001	`suc.filler`	3	573.9	28.5	< .001
`pre.pause`	3	501.2	20.1	< .001	`pre.pause`	3	584.6	18.7	< .001

Table 5: Estimated coefficients for single main effect models.

APS						SPS					
	Coef.	SE	df	t	p		Coef.	SE	df	t	p
						`boundary=D`	.23	.08	281	2.73	< .01
`suc.pause`	.42	.08	329	5.38	< .001	`suc.pause`	.57	.08	281	6.83	< .001
`suc.filler`	.41	.06	329	7.08	< .001	`suc.filler`	.52	.09	281	5.47	< .001
`pre.pause`	.16	.03	329	4.55	< .001	`pre.pause`	.42	.09	281	4.51	< .001

4.3.2 Multivariate model

To understand the relative contribution of the variables, we next applied a linear mixed-effect model, and compared the contributions of the variables by means of a model selection technique. Table 6 shows the model selection steps, for both the forward and the backward selections, for APS and SPS. For both speech types, the forward and the backward selections converged on the same optimal model. The optimal models for

APS and SPS also contained the same set of variables. These included the variables concerning the (log-transformed) duration of the preceding pause (`pre.pause`), the presence of a succeeding pause (`suc.pause`) and filler (`suc.filler`). The variables concerning boundary type (`boundary`) and clause complexity (`complex`) were excluded from the optimal models.

Table 7 shows the estimated coefficients for the optimal models. In both speech types, the presence of a succeeding pause or filler largely extends the duration of clause-initial *de* (Pause: 1.34 sec in APS and 1.52 sec in SPS; Filler: 1.36 sec in APS and 1.38 sec in SPS). The contributions of the two variables were almost comparable. Duration of the preceding pause also affected the duration of clause-initial *de* in a positive direction; the longer the preceding pause, the longer the initial *de* is. This effect, however, was very small in APS.

Table 6: Model selection steps. The upper half shows the steps for the forward selection and the bottom half shows the steps for the backward selection.

APS					**SPS**				
Model	df	−2logLik	χ^2	*p*	Model	df	−2logLik	χ^2	*p*
Minimal	2	521.3			Minimal	2	603.4		
`+suc.filler`	3	474.6	46.7	< .001	`+suc.pause`	3	560.2	43.2	< .001
`+suc.pause`	4	459.8	14.8	< .001	`+suc.filler`	4	547.5	12.7	< .001
`+pre.pause`	5	452.8	7.0	< .01	`+pre.pause`	5	538.0	9.5	< .005
Maximal	7	452.8			Maximal	7	537.2		
`-boundary`	6	452.8	.00	= .96	`-complex`	6	537.4	.17	= .68
`-complex`	5	452.8	.07	= .79	`-boundary`	5	538.0	.62	= .43

Table 7: Estimated coefficients for the optimal models.

APS						**SPS**					
	Coef.	SE	df	*t*	*p*		Coef.	SE	df	*t*	*p*
(Intercept)	−2.08	.09	327	−22.79	< .001	(Intercept)	−1.80	.10	279	−17.67	< .001
`suc.pause`	.29	.08	327	3.75	< .001	`suc.pause`	.42	.09	279	4.88	< .001
`suc.filler`	.31	.06	327	5.13	< .001	`suc.filler`	.32	.09	279	3.39	< .001
`pre.pause`	.09	.03	327	2.66	< .01	`pre.pause`	.28	.09	279	3.18	< .005
Random intercept: σ_1^2 = .05						Random intercept: σ_1^2 = .08					

4.4 Discussion

We have identified several factors that affect the duration of the clause-initial conjunction *de*:

- The presence of a succeeding pause or filler significantly increased the duration of clause-initial *de*.
- The duration of the preceding pause also had a significant effect on the duration of clause-initial *de*, although this effect was very small in APS; the duration of *de* was longer when the duration of the preceding pause was longer.
- There was no reliable evidence that clause complexity, measured by the number of words, affects the duration of clause-initial *de*.
- In the analysis of individual features, boundary type was shown to have a significant effect on the duration of *de* in SPS, but no significant effect was found in APS. The effect observed in SPS, however, disappeared when multiple variables were considered simultaneously.

It is widely known that segments at phrase-final position are generally lengthened (e.g., Klatt 1975, Takeda et al. 1989), a phenomenon known as *phrase-final lengthening*. Segments are more extended when there is a strong break between the current and the succeeding segments. The presence of a succeeding pause introduces very strong discontinuity and, thus, our result of the effect of succeeding pauses may display a general tendency brought about by phrase-final lengthening. In our case, however, the lengthened segment is a mono-moraic word appearing at clause-initial position, which is quite different from the syntactic and acoustic configuration in which phrase-final lengthening is generally observed. We may, rather, consider this a particular format that the speaker uses strategically: a prolonged *de*, followed by a pause. Such a format in the combined use of a prolonged segment and a succeeding pause is utilized when speakers repeat a word at the beginning of an utterance (Den 2007).

The result of the effect of succeeding fillers suggests that in the context where a filler is produced, the preceding *de* is also likely to be prolonged. Fillers, in general, tend to more often occur at a discourse boundary or at the beginning of a complex clause, where speakers' planning load is more severe (Swerts 1998, Watanabe et al. 2006). The current study, however, found no reliable evidence that boundary type or clause complexity affects the duration of clause-initial *de*. It can not be concluded that prolongation of *de* is attributable to speakers' cognitive load in planning the utterance to be construed.

Table 8: Summary of results. ○, △, and × mean that the effect was observed, partly observed, and not observed, respectively. APS or SPS in parentheses means that the effect was absent, or weak, in that speech type.

	Analysis 1	Analysis 2
`boundary`	△ (APS/SPS)	×
`complex`	×	×
`pre.pause`	○ (APS)	○ (APS)
`pre.filler`	△ (APS/SPS)	−
`suc.pause`	○	○
`suc.filler`	×	○

Duration of the preceding pause also contributed to modeling the duration of clause-initial *de*. It is known that pause durations between consecutive utterances are associated with discourse structure (Hirschberg & Nakatani 1996, Swerts & Geluykens 1994, Yoneyama et al. 2003). Pauses are longer at a discourse boundary than within a discourse segment. Thus, our result may indicate the existence of a discourse factor on the duration of *de*, although the effect of boundary type *per se* was not evident. It is possible that the boundary effect was buried under the more robust effects of pause duration and the presence of fillers, which by themselves may be affected by certain discourse factors. The current method cannot deal with such structured interaction among factors.

5. General discussion

The analyses presented in the previous two sections showed rather consistent results. Table 8 summarizes the results obtained in the current study. For each factor, ○, △, or × is indicated based on the results of the multivariate models constructed in each analysis. The three symbols indicate that the effect was observed, partly observed, and not observed, respectively. Clause complexity was consistently found to have no effect on the rate of prolongation at a CIMWP or on the duration of its particular form *de*. In contrast, the duration of the preceding pause and the presence of a succeeding pause had a consistent influence in the two analyses, although the effect of the preceding pause was relatively weak for APS in both analyses. The effect of boundary type was observed only in Analysis 1, whereas the effect of the succeeding filler was observed only in Analysis 2. Particularly, the boundary effect in Analysis 1 was observed only in APS and was very weak.

It is very interesting to note that despite a substantial difference in the nature of the two speech types, most observations were consistent between the two speech types.

APS and SPS are different from each other in many aspects including speaking style, spontaneity, and rate of disfluency (Maekawa et al. 2003). The difference is also evident in the current study as shown by the difference of the overall PR rate of CIMWPs in Fig. 3. Since speakers of APS usually plan the content of the discourse deliberately in advance of the presentation, the cognitive load in speech production may not be so heavy for speakers of APS, compared to speakers of SPS. This should have brought about some differences in the characteristics of prolongation between the two speech types. Nevertheless, the two speech types had in common many syntactic and acoustic factors related to prolongation.

As shown in the current study, speakers' cognitive load—at least in planning the utterance being construed—is not a decisive factor influencing prolongation of CIMWPs. This may be a reason why the major results for the two speech types are so similar. On the other hand, we do not know exactly what the decisive factor is. We suggest a possibility of some discourse factor that may affect the duration of pauses immediately preceding the clause. This possibility, however, should be further investigated by a more advanced method and larger data sets with discourse structure annotation.

In summary, we investigated a particular type of prolongation in Japanese, the prolongation of clause-initial mono-word phrases, and its subclass, the prolongation of the clause-initial conjunction *de*. We reported the effects of various syntactic and acoustic factors and their relative contributions. We suggest that such prolongation is not necessarily induced by speakers' cognitive load, but that some discourse factors may be involved. We believe this is a good starting point for a deep and precise understanding of the entire nature of the prolongation phenomena.

Yasuharu Den
Faculty of Letters
Chiba University
1-33 Yayoi-cho, Inage-ku
Chiba 263-8522, Japan
den@cogsci.L.chiba-u.ac.jp

Linguistic Patterns in Spontaneous Speech, 193-211
2009-1-025-009-000271-1

Spontaneous Mandarin Speech Recognition with Disfluencies Detected by Latent Prosodic Modeling (LPM)

Che-Kuang Lin[1], Shu-Chuan Tseng[2], and Lin-Shan Lee[1]
National Taiwan University[1]
Academia Sinica[2]

In this paper, a new approach for improved spontaneous Mandarin speech recognition using Latent Prosodic Modeling (LPM) for disfluency interruption point (IP) detection is presented. The basic idea is to detect the disfluency interruption points (IPs) prior to the recognition, and then to incorporate these information into the recognition process via the second pass rescoring. For accurate detection of disfluency interruption points (IPs), prosodic information from local to global, from observable to latent, were integrated using the proposed Latent Prosodic Modeling (LPM). A whole set of new features were first defined for each syllable boundary obtained in the first pass recognition by carefully considering the special characteristics of Mandarin Chinese, and the importance of each feature with respect to each disfluency type was analyzed. Then, a set of prosodic characters, prosodic terms, and prosodic documents were defined to be used in the Probabilistic Latent Semantic Analysis (PLSA), based on which the prosody can be modeled using a set of prosodic states representing various latent factors such as speakers, speaking rate, utterance modality, intonation behavior, etc. in terms of some probabilistic relationships with the observed prosodic features. Using all these different levels of information, the approach of incorporating the decision tree into the maximum entropy model training was developed to enhance the IP detection accuracy. Experimental results indicated that the proposed set of features and the IP detection approach based on Latent Prosodic Modeling (LPM) were very useful, and the obtained information about disfluency actually benefited the speech recognition performance.

Key words: spontaneous speech recognition, disfluency, prosody, latent modeling, Mandarin Chinese

1. Introduction

Disfluencies, as one of the primary sources of ill-formness in spontaneous speech, pose difficult but important problems for spontaneous speech processing. Substantial work has been reported in this area (Lickley 1996, Lendvai et al. 2003, Honal & Schultz 2005, Liu et al. 2005). While analyses regarding different aspects of disfluency phenomena have been conducted and much insight into the related acoustic and prosodic properties has been gained so far, we are trying to integrate all the valuable knowledge into a system for enhanced speech recognition. The structure of disfluencies is usually considered to be decomposed into three regions: the reparandum, an optional editing term, and the resumption. The disfluency interruption point is the right edge of the reparandum. The purpose of the research presented in this paper is to identify useful and important features in automatic detection of such disfluency interruption point (IP) in spontaneous Mandarin speech, and analyze how these features are helpful in speech recognition. The disfluencies considered here in this paper include the following four categories:

(i) Direct repetitions: the speaker repeats words in a way that can not be justified by grammatical rules. Many other cases of repetitions in Mandarin Chinese are perfectly legal syntactic constructions for emphasis purposes and so on, which should be excluded from this study.
(ii) Partial repetitions: only part of a word (including compound words) is repeated.
(iii) Overt repairs: the speaker modifies expressed words within an utterance.
(iv) Abandoned utterances: the speaker abandons an utterance and starts over.

Consider the following example in Mandarin Chinese:

是	進口	嗯	出口	嗎
shi4	**jin4kou3**	**EN**	**chu1kou3**	**ma1?**
is	import	[discourse particle]	export	[interrogative particle]

(*Do you import * uhn export products?*)

In this example, “uhn” is a filled pause and “export” is meant to correct “import”, which is an overt repair. Here ‘*’ denotes the right edge of the reparandum region, or the interruption point (IP) to be detected and utilized in recognition here. Consider another example:

因爲	因爲	他	有	健身	中心
yin1wei4	**yin1wei4**	**ta1**	**you3**	**jian4shen1**	**zhong1xin1**
because	because	it	has	fitness	center

(*Because * because it has a fitness center.*)

Here the speaker repeats the word “because” to restart the sentence.

Prosodic information in speech signals can be considered to be more or less orthogonal to MFCC features and therefore should be useful for many spoken language processing applications (Hirose & Minematsu 2004, Vergyri et al. 2003, Shriberg et al. 2000, Chen et al. 2003). However, very often such information was found useful in speech synthesis, but relatively difficult to use in speech recognition. The difficulties include, among many others, the fact that the prosody is usually speaker dependent (Chen et al. 2006), and that training corpora labeled with prosodic events usually require human efforts and are less available. In this paper, we try to develop a new framework of Latent Prosodic Model (LPM) for speech signals with a goal to at least handle parts of the above difficulties to a certain degree.

The concept of Latent Prosodic Modeling (LPM) is actually borrowed from the Probabilistic Latent Semantic Analysis (PLSA) very useful in the area of information retrieval (Hofmann 1999). In this approach, instead of directly counting the co-occurrence statistics between the document set $\{d_i\}$ and the term set $\{t_k\}$, a set of latent topics $\{z_l\}$ is created and the relationships between each document d_i and each term t_k are modeled by a probabilistic framework via these latent topics:

$$(1)\quad P(t_k \mid d_i) = \sum_{l=1}^{L} P(t_k \mid z_l) P(z_l \mid d_i)\ ,\forall i,k$$

where the probabilities were trained with EM algorithms by maximizing the total likelihood function:

$$(2)\quad L_T = \sum_{i=1}^{N} \sum_{k=1}^{N'} n(t_k, d_i) \log P(t_k \mid d_i)$$

and $n(t_k, d_i)$ denotes the frequency count of t_k in d_i, and N and N' are the total number of documents and terms respectively. In the Latent Prosodic Modeling (LPM) developed here, t_k, d_i, and z_l are to represent prosodic terms, prosodic documents, and the latent prosodic states respectively, as will be clear below.

Below, we first describe the corpus used in this research in §2 and then introduce the set of proposed prosodic features as well as IP detection models in §§3 & 4. Then in §5, we present the basic framework for LPM, while in §6, we describe the improved models for IP detection using LPM. Section 7 then gives the recognition approach incorporating the IP information. Analysis regarding the contribution of different features for detection of different types of disfluency IPs is presented in §8. The experimental results are then discussed in §9, and the concluding remarks finally made in §10.

2. Corpus used in the research

The corpus used in this research was taken from the Mandarin Conversational Dialogue Corpus (MCDC) (Tseng 2004, website http://mmc.sinica.edu.tw), collected from 2000 to 2001 by the Institute of Linguistics of Academia Sinica in Taipei, Taiwan. This corpus includes 30 digitized conversational dialogues with a total length of 27 hours. 8 dialogues out of the 30, with a total length of 8 hrs, produced by nine female and seven male speakers, were annotated by adopting a taxonomy scheme of four groups of spontaneous speech phenomena: disfluencies, sociolinguistic phenomena, particular vocalization, and unintelligible or non-speech sounds. Disfluencies here include breaks, word fragment, overt repairs, direct repetitions, abandoned utterances, discourse particles, and markers. In this paper, we only deal with direct repetitions, partial repetitions, overt repairs and abandoned utterances. The 8 hrs of annotated dialogues as mentioned above were used in this research. Due to the mono-syllabic structure of the Chinese language, i.e., in Mandarin Chinese every character has its own meaning and is pronounced as a monosyllable, while a word is composed of one to several characters (or syllables), every syllable boundary is considered as a possible interruption point (IP) candidate in this research. Table 1 summarizes the data used in the following experiments. Only 3.7% and 3.9% of the syllable boundaries are IPs.

Table 1: The summary of experiment data

	train (6.9hr)	test (1.3hr)
Number of IPs / non-IPs	3432/89891	673/16529
Chance of non-IPs	96.3%	96.1%

3. Prosodic features

As mentioned above, due to the mono-syllabic structure of Chinese language, every syllable boundary is considered as a possible interruption point (IP) candidate in this research. We therefore tried to define a whole set of prosodic features for each IP candidate, or each syllable boundary, and use them to detect the IPs. Many prosodic features have been proposed and proved useful for such purposes (Shriberg et al. 2000, Liu et al. 2003), and it has been found (Liu et al. 2005) that it is important to identify better features. Because this research is focused on IP detection, we tried to identify some IP specific features. Moreover, considering the special feature of Mandarin Chinese including the tonal language nature, some acoustic phenomena for Mandarin spontaneous speech may be quite different from those in English. Such consideration was reflected here by constructing a new set of features.

3.1 Pitch-related features

Pitch information is typically less robust and more difficult to use (Shriberg et al. 2000). Pitch contour stylization method has thus been used, and smoothing out the "micro-intonation" and tracking errors was found helpful for English (Shriberg et al. 2000, Liu et al. 2003). For a tonal language such as Mandarin Chinese, however, such "micro-intonation" apparently carries tone or lexical information, and thus should not be removed, although some approaches of pitch contour smoothing are certainly needed.

In this research, we used Principal Component Analysis (PCA) for syllable-wise pitch contour smoothing, instead of piece-wise linear stylization. For each syllable, the pitch contour was decimated or interpolated to become a vector with fixed dimension. PCA was then performed on such training vectors. By choosing the principal components with the largest eigenvalues, we projected the fixed dimension vectors onto the subspace spanned by the principal components to obtain the smoothed version of the pitch contours. Various pitch-related features were then extracted from these smoothed pitch contours, such as the pitch reset for boundaries being considered and so on. Quite several syllable-wise pitch-related features found useful in tone recognition were also used here, such as the average value of normalized pitch within the syllable, the average of absolute value of pitch variation within the syllable, the maximum difference of normalized pitch within the syllable and so on, all evaluated for the syllable before and after the boundary being considered. A total of 54 such pitch-related features were considered (Lin & Lee 2005, Lin et al. 2005).

3.2 Duration-related features

Duration features such as pause and phone duration features have been used to describe prosodic continuity and preboundary lengthening (Shriberg et al. 2000, Liu et al. 2003). By carefully examining the characteristics of IPs in our corpus, we hypothesized that deviation from the normal speaking rhythmic structure is an important cue to disfluency IP detection. For example, relatively sudden, sharp, discontinuous changes in speaking rate were consistently observed across IPs. We also hypothesized that certain ways of integration of pause and syllable duration fluctuations are important characteristics of the rhythmic structure of speech. Considering these observations, we derived the following set of duration-related features to try to detect IPs.

We first computed the average and standard deviation of syllable duration over several syllables before and after the boundary being considered. Then we calculated the ratio of the former to the latter. The possible ranges for evaluating the above statistics included one, two, three syllables as well as extending to the nearest pauses on both sides. Another group of duration-related features were generated by jointly considering the pause duration and the duration parameters of the syllables before or after the pause. The product of these two different duration parameters represented some integration of the two types of information. Alternatively, normalizing the syllable duration parameters by the duration of a nearby pause being considered may emphasize the fluctuations of these syllable duration parameters. Finally, a total of 38 such duration-related features were considered (Lin & Lee 2005, Lin et al. 2005).

4. Interruption point (IP) detection

The approaches used for IP detection are discussed in this section. The IP detection task is considered as a classification problem here in this research. For each syllable boundary, a decision between "non-IP" vs. various types of "IPs" was made. Because IPs were relatively rare events, we used ensemble sampling (Liu et al. 2004) on training data to equate the prior probabilities for different classes. This made the model trained more sensitive to any features that distinguish the classes.

4.1 Decision tree (DT) and maximum entropy (Maxent) model

In the first approach, we used decision trees to learn from data, and to make prediction while testing (Liu et al. 2003). The decision was then made according to the posterior probability of the leaf node where the test sample of the syllable boundary went to. In the second approach, we applied the maximum entropy model to make the decision (Berger et al. 1996). In this model, a feature is expressed by a binary feature function $f_i(x, y)$, in which x denotes the feature sets and y denotes the outcome. The final expression for p(y|x) in this model takes the following form:

$$(3) \quad p(y \mid x) = \frac{1}{Z(x)} \exp\left[\sum_i \lambda_i f_i(x, y) \right]$$

where $Z(x)$ is a normalization term.

The maximum entropy model is estimated by finding the parameters λ_i for each feature function $f_i(x, y)$ with the constraint that the expected values of the various feature functions match the empirical averages in the training data. In our experiments we used the L-BFGS parameter estimation method with Gaussian-prior smoothing to avoid overfitting.

4.2 Integration of DT and Maxent

Considering the decision trees and maximum entropy model mentioned above, we can find that each of them has some advantages and limitations in dealing with the problem here. Decision Trees can handle real-valued features directly while Maxent is working in a discrete style. On the other hand, by carefully designing the feature function, maximum entropy model may make finer decision on a certain feature parameter, while for the decision trees the training process only uses binary partitioning on feature parameters to split the data. When the problem is not linearly separable, it might not be possible for the decision trees to find a good partition without growing too deep. But too deep trees often lead to overfitting and thus degrade the performance on the testing set.

Based on all the above considerations, we developed a new approach to integrate the decision trees and maximum entropy model together (Lin & Lee 2005). In this approach, we used decision trees built with training data to derive the feature functions for the maximum entropy model, hoping to have the advantages of both. We first trained a set of decision trees by ensemble downsampling of the training data. Instead of growing the optimized tree by cross validation, we chose Bayesian criterion to grow the tree, resulting in a set of much deeper and bushy trees. Each leaf of all the trees was then used as a feature function in Maxent. In other words, a feature function was assigned “1” if and only if the sample being considered went to the corresponding tree leaf. Otherwise the feature function was 0-valued. By having each sample traversing down to the leaves, we had the feature function values all decided. The training procedure was then the same as the original maximum entropy model. While testing, the complete procedure was the same as that of training stage and the pre-trained trees were used again. This approach is referred to as the integrated maximum entropy (integrated maxent) model hereafter in this paper.

5. Latent prosodic modeling (LPM) for speech

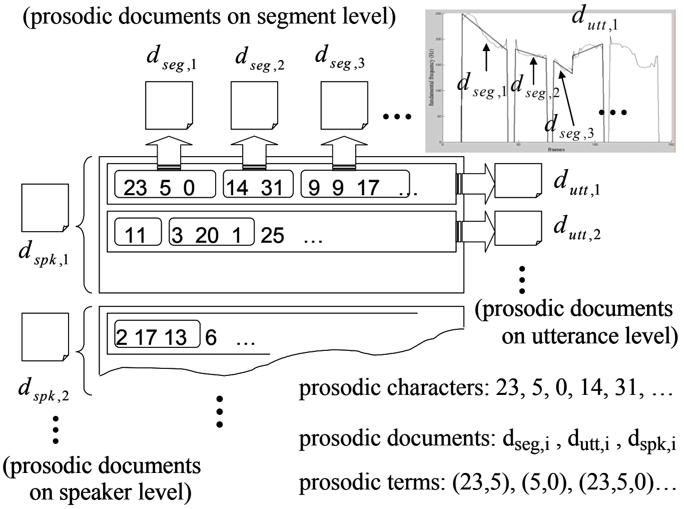
The dynamic behavior of speech prosody is affected by various latent factors, such as speakers, speaking rate, utterance modality, intonation behavior, etc., which leads to the significant variations in the observed prosodic features. The goal of LPM is to perform delicate analysis of the prosody by properly modeling the wide variety of prosodic features in terms of such latent factors, referred to as prosodic states here.

We defined prosodic terms and documents from speech signal as the unit of prosody analysis. Although these units are not defined based on any known theory of prosody, they served as units of analysis when not much the spoken content but the signal itself is available in a bottom-up recognition framework. The prosodic feature vectors can first be extracted for phones, syllables, words, phrases, etc. (for the present research, for each syllable boundary as mentioned in §3). Vector quantization (VQ) can then be used to label the feature vectors into discrete codewords, referred to as prosodic characters. The n-grams of these prosodic characters are then referred to as prosodic terms. The prosodic behavior of a certain part of the speech signal is then referred to as a prosodic document, composed of and characterized by the various prosodic terms included. The use of the term "document" is borrowed from PLSA and is used metaphorically here. All these are illustrated in Fig. 1, in which three levels of prosodic documents were considered in this paper: segments, utterances, and speakers. The segments are parts of an utterance obtained from the best fitting piece-wise linear function for the pitch contour (Shriberg et al. 2000).

For the set of each level of prosodic documents $\{d_i\}$ and the included prosodic terms $\{t_k\}$, we can then train a PLSA model as in equations (1) (2) by introducing a set of latent factors $\{z_l\}$, referred to as prosodic states here, and related all the prosodic terms t_k and prosodic documents d_i to the prosodic states z_l in terms of probabilistic distributions as shown in equation (1). This is the LPM proposed here in this paper. With such a model, the complicated behavior of the many prosodic features can be analyzed in terms of the latent prosodic states in some way. For instance, the similarity between any two prosodic documents d_i and d_j, $Sim_{LPM}(d_i, d_j)$, can be estimated by their probability distributions with respect to the various prosodic states, $P(z_l \mid d_i)$ and $P(z_l \mid d_j)$, with the expression below as one example:

$$(4) \qquad Sim_{LPM}(d_i, d_j) = \frac{\sum_l P(z_l \mid d_i) P(z_l \mid d_j)}{\sqrt{\sum_l [P(z_l \mid d_i)]^2} \sqrt{\sum_l [P(z_l \mid d_j)]^2}} \ .$$

Many other distance metrics can also be used, such as the Kullback-Leibler distance and Mahalanobis distance.



(prosodic documents on segment level)
$d_{seg,1}$
$d_{seg,2}$
$d_{seg,3}$
$d_{utt,1}$
$d_{seg,1}$
$d_{seg,2}$
$d_{seg,3}$
23 5 0
14 31
9 9 17 ...
$d_{utt,1}$
11
3 20 1
25 ...
$d_{utt,2}$
$d_{spk,1}$
(prosodic documents on utterance level)
2 17 13
6 ...
$d_{spk,2}$
prosodic characters: 23, 5, 0, 14, 31, …
prosodic documents: $d_{seg,i}$, $d_{utt,i}$, $d_{spk,i}$
(prosodic documents on speaker level)
prosodic terms: (23,5), (5,0), (23,5,0)…

Figure 1: Prosodic characters, terms and documents for latent prosodic modeling (LPM)

Although below we use this model for IP detection (Lin & Lee 2006), the above model can also be useful in many applications such as prosodic behavior classification, for example using the distance measure in equation (4). We may also realize delicate classification models that is adapted to, say, a specific speaker, a kind of utterance modality, or a particular intonation context, using other efficient classification algorithms (e.g. integrated maxent) but based on LPM in an unsupervised manner. As illustrated in Fig. 2, we can actively select the desired training set for a specific testing condition by LPM at each level of prosodic documents, the segments, utterances or speakers. Taking the speaker level for example, the speaker-type model based on a subset of training data produced by the group of speakers with similar prosodic properties may be obtained in this way.

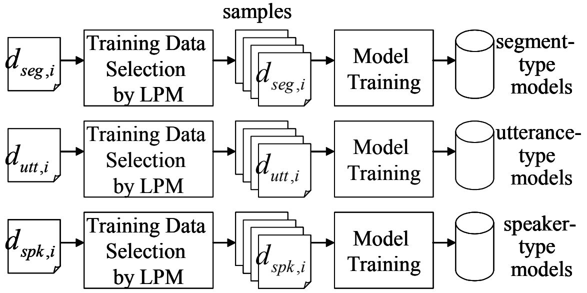


Figure 2: Training of segment-, utterance-, speaker-type models based on Latent Prosodic Modeling (LPM)

On the other hand, LPM can also be used in an alternative framework (Lin & Lee 2006) to learn the patterns for different classes of speech signals in a supervised manner, referred to as anchor modeling here. In this approach, the prosodic documents associated with each desired class were merged into a super-document representing the characteristics of this class, and LPM was then performed upon the set of super-documents. Thus the prosodic characteristics of each class anchor, in terms of the relationships with each prosodic state, can then be analyzed.

6. Interruption point (IP) detection in spontaneous Mandarin speech with LPM

We used LPM proposed here for IP detection for spontaneous Mandarin speech.

6.1 Integrated maximum entropy (integrated maxent) modeling based on LPM

The integrated maxent model mentioned above can be further improved by the LPM proposed here just as shown in Fig. 2. The prosodic documents in the training corpus were first classified by LPM based on the latent prosodic states, and then more delicate integrated maxent models based on the prosody of the segment types, utterance types and speaker types can be trained. As illustrated in Fig. 3, the classification scores obtained by the three delicated integrated maxent models based on segment types, utterance types and speaker types were then combined with the score by the integrated maxent model without LPM via a support vector machine (SVM) with a radial basis kernel using the LIBSVM tool (Chang & Lin 2004).

6.2 Anchor-based model with LPM

In this approach, we established with LPM a set of five anchors, each for one out of the four IP classes (overt repair, abandoned utterances, direct repetition, partial repetition) and the non-IP boundaries, to detect the disfluency IPs. As mentioned previously, prosodic documents in the training corpus associated with each of the above five classes were merged into five super-documents representing the prosodic characteristics of the four IP classes and non-IP, which produced a set of corresponding prosodic anchors after LPM. We similarly trained such anchor models on the three levels, i.e., for segment-, utterance- and speaker types, as in Fig. 2, and combined the scores using SVM as in Fig. 3.

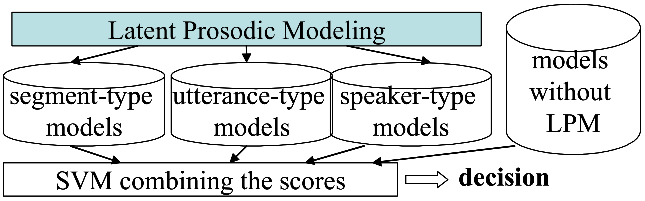


Figure 3: Integration of LPM-based classification models with SVM

6.3 LPM-based feature expansion for integrated maxent

In addition, the probabilities that each prosodic state z_l is related to the prosodic document d_i, $\{P(z_l \mid d_i), \forall l\}$, and the likelihood of the prosodic terms given the prosodic document, $\{\prod_{t_k \in d_i} P(t_k \mid d_i) = \prod_{t_k \in d_i} \sum_{l=1}^{L} P(t_k \mid z_l) P(z_l \mid d_i)\}$, obtained from LPM for prosodic documents at each level, can also be directly used as another set of features, together with other prosodic features for the integrated maxent models.

7. Speech recognition with IP detection

Here we present the way to incorporate the IP detection results into the speech recognition processes. The IP detection gave the probability for each syllable boundary to be an IP (very often zero) along a sequence of word hypotheses. For each utterance, we combined such information for each path in an n-best list, where the probability for each syllable boundary to be an IP was the weighted sum over all paths in the n-best list, using the total likelihood scores for the paths as the weights. This gave each syllable boundary a final probability to be an IP, which was then used in the following search process over the word graph.

We rescored the word graph based on the maximum a posterior (MAP) principle considering the prosodic information:

$$
\begin{aligned}
(5) \quad W^* &\equiv \arg\max_{W} P(W \mid X, F) \\
&= \arg\max_{W} P(W \mid F) P(X \mid W, F) \\
&\cong \arg\max_{W} P(W \mid F) P(X \mid W)
\end{aligned}
$$

where X and F are the acoustic and prosodic feature sequences respectively, the recognized word sequence W^* is the one which maximizes the posterior probability P(W|X,F), and the last expression was based on the assumption that X and F can be approximated as independent given the word sequence W. P(W|F) is modeled considering the probabilities for the different disfluency IP classes as follows:

$$(6)\quad P(W \mid F) = \prod_{n} P(w_n \mid w_{n-N+1}^{n-1}, F)$$
$$= \prod_{n} \sum_{c} P(c \mid w_{n-N+1}^{n-1}, F)^{\lambda} P(w_n \mid w_{n-N+1}^{n-1}, c)$$

where w_{n-N+1}^{n-1} is the N-1 words before the n-th word w_n, c is one out of the five IP classes including non-IP, λ is a weight parameter, and $P(c \mid w_{n-N+1}^{n-1}, F)$ is approximated using the probability obtained from IP detection, or $P(c \mid w_{n-N+1}^{n-1}, F) \cong P(c \mid F)$. The word n-grams crossing different classes of IP boundaries (i.e. $P(w_n \mid w_{n-N+1}^{n-1}, c)$) were evaluated from disfluency corpus separately, and then interpolated with the baseline language model.

8. Feature analysis for disfluency detection

8.1 Comparison between duration- and pitch- related features for different disfluency types

To get a further insight into the characteristics of various disfluency categories and the IP detection process, we tried to find the relation between the features used and the IP detection performance. A partial feature selection analysis was performed upon the full feature set mentioned earlier. In this approach, we excluded each single feature from the full set and then perform the complete IP detection process in each small experiment, to find out how much the IP detection performance was degraded due to the missing of this single feature. Here the performance is in terms of recall rate only. Because we grouped all the four types of disfluencies together into a single class due to the small size of the corpus, precision for each disfluency type was not obtainable, while recall was. The results discussed here were obtained from integrated maximum entropy model.

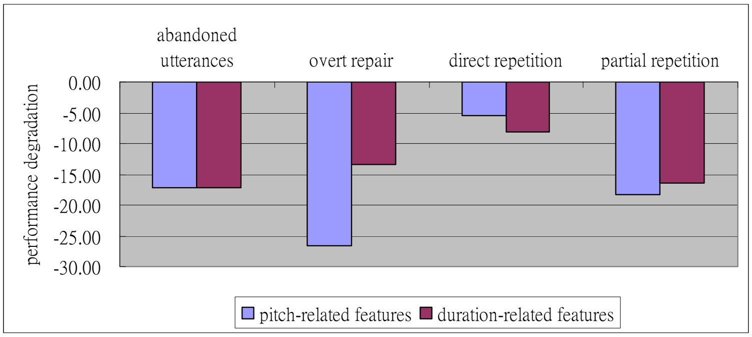


Figure 4: Performance degradation for the four disfluency types with respect to the two feature categories

First, to see how pitch-related and duration-related features contribute to the IP detection of different types of disfluencies, we compared the performance degradation for the four disfluency types being considered with respect to the two feature categories. In Fig. 4, we show the most serious performance degradation caused by removing one single feature from the two categories of either pitch-related or duration-related features. We can find that for overt repair and partial repetition, pitch-related features play relatively more important role for IP detection, and this is specially apparent for overt repair. This is in good consistency with the earlier findings (Tseng 2006a) that overt repairs are produced partly because the correction of the delivered information is required, and partly because the speaker changes his/her language planning. It is often true that when overt repairs are produced within utterances, the F0 level of the onset of the resumption part is approximately reset to that of the onset of the reparandum. In other words, the resumption part should fit seamlessly into the original utterance after removing the problematic items. Then the cleaned utterance should look like a natural utterance that obeys the normal F0 declination. In addition, intonation units have been defined and analyzed in Mandarin conversation (Tao 1996), which are unique characteristics in spoken language different from syntactic units. They are also found to be highly related to the language planning process. Moreover, it has been observed (Tseng 2006a) that almost all reparandum parts are themselves intonation units. The behavior of overt repair is just similar to that of a new intonation unit with respect to the preceding one. All these imply that overt repairs have a lot to do with the intonation units and thus pitch-related features. All these are consistent with the results here, i.e., the cues carried by pitch-related features provide important information for overt repair detection. On the other hand, we can also find that for direct repetition IP detection, the duration-related features are more important, and for abandoned utterances IP detection, both pitch-related and duration-related features have equally important impact.

8.2 Pitch-related features for IP detection of different disfluency types

The 13 features found to be the most important in IP detection for the four different types of disfluencies are represented by symbols (a) to (m) with their definitions as listed in Table 2, where the upper and lower halves are for pitch-related and duration-related features respectively. In Table 3, for each of the four categories of disfluencies, we list the symbols for the two pitch-related features causing the most serious recall rate degradation, or the two most important pitch-related features, together with the associated recall rate degradation, in the columns labeled as “pitch-related”. We can see that the average pitch value within a syllable, used in features represented by symbols (b) and (d) in Table 2, appears to be very important in three out of the four types of disfluencies, regardless of different smoothing methods used. This suggests that the level of pitch is a very good cue for disfluency IP detection, probably due to the tone information carried and the intonation unit property as mentioned earlier. Especially, the absence of this feature degrades the performance very severely on partial repetitions and abandoned utterances.

Direct repetition, on the other hand, is much less influenced. Moreover, the difference of maximum and minimum pitch values within a syllable, used in features represented by (e) and (f) in Table 2, is beneficial to IP detection of direct repetitions and partial repetitions. It has been found (Tseng 2006a) that as far as Mandarin Chinese is concerned, the overt repairs, direct repetitions, and partial repetitions tend to be shorter. The main reason is probably that in Mandarin Chinese there is no inflection and the word order can vary to a great extent, speakers can re-initiate at the morphological boundary immediately after some inappropriateness is sensed. Moreover, it was also found that simple direct repetition repeating only one syllable usually dominates (Tseng 2006a). With plenty of such mono-syllable repeats, the pair of (partially) repeated and re-initiated syllables very often exhibit highly similar pitch contours. With the tone information inside these contours, pitch level (features (b) and (d)) and range (features (e) and (f)) can thus capture the evidence of short direct repetition and partial repetition. Another important pitch-related feature in Table 3 is the difference of pitch value across boundaries (used in the feature represented by (a)). This feature somehow conveys to what degree the speaker resets the pitch at this boundary. The reset of pitch is often the evidence of starting a new intonation unit, which is probably also the beginning of a new planning unit. This may be the reason why this feature is very important in the detection of abandoned utterances and overt repair IPs.

Table 2: The definitions of features used in Table 3. Δ(z): the parameter z was evaluated for each syllable boundary, and Δ(z) is the difference of the parameter z for two neighboring syllable boundaries.

Pitch-related features	(a)	Δ(difference of pitch slope across boundary)
	(b)	Δ(average pitch value within a syllable), with pitch value obtained from raw f0 value
	(c)	averaged absolute value of pitch slope within a syllable, with pitch value obtained from linear approximation
	(d)	Δ(average pitch within a syllable), with pitch value obtained from PCA
	(e)	Δ(difference of maximum and minimum pitch value within a syllable), with pitch value obtained from raw f0 value
	(f)	Δ(difference of maximum and minimum pitch value within a syllable), with pitch value obtained from linear approximation

Duration-related features	(g)	Δ(ratio of the duration for the syllable before the boundary to the pause duration at the boundary)
	(h)	ratio of the duration for the syllable after the boundary to the pause duration at the boundary
	(i)	product of the duration for the syllable after the boundary with the pause duration at the boundary
	(j)	Δ(product of the duration for the syllable after the boundary with the pause duration at the boundary)
	(k)	Δ(ratio of the duration for the syllable after the boundary to the pause duration at the boundary)
	(l)	syllable duration parameter ratio across the boundary, with the duratoin parameter being the average over 3 neighboring syllables
	(m)	standard deviation of (product of the duration for the syllable before the boundary with the pause duration at the boundary)

Table 3: The recall rate degradation when excluding an pitch-related/duration-related feature for different types of disfluencies (with definitions of features listed in Table 2).

Disfluency Types	Most Important Features (recall degradation)		Second Important Features (recall degradation)	
	pitch-related	duration-related	pitch-related	duration-related
abandoned utterances	(a) (-17.25)	(g) (-17.25)	(b) (-14.97)	(h) (-14.97)
overt repairs	(c) (-26.67)	(i) (-13.33)	(a) (-20.00)	(j) (-13.33)
direct repetition	(d) (-5.40)	(k) (-8.10)	(e) (-5.40)	(l) (-8.10)
partial repetition	(b) (-18.21)	(h) (-16.33)	(f) (-18.21)	(m) (-16.33)

8.3 Duration-related features for IP detection of different disfluency types

Table 3 also listed similar analysis with respect to duration-related features, in which we list the two most important duration-related features, together with the associated recall rate degradation, for the four types of disfluencies, in the columns labeled as "duration-related". Although duration-related features are beneficial to direct repetition detection as mentioned above, they also help indicate IP of other types of disfluencies. First, jointly considering both the syllable duration and pause duration was shown to be useful across all kinds of disfluencies. Combining through ratio of syllable duration to pause duration (represented by (g), (h) and (k) in Table 2) is relevant to IP detection of abandoned utterances, direct and partial repetitions, while overt repairs and partial repetition benefit from the product of them (represented by (i), (j) and (m) in Table 2). The ratios may have normalized the syllable duration with respect to the breathing tempo of the speaker, if any, which was revealed by the pause duration fluctuation. The results showed that such features are actually useful. Moreover, a specific feature for direct repetition is the character duration ratio across boundary (represented by (l)), implying how the speaking rate was fluctuating. This showed that direct repetitions usually cause significant speaking rate deviation, and this is consistent with the observation obtained before (Tseng 2006a), in which it was concluded that the repeated words in the resumption are shorter than those in the reparandum part, because the direct repetition itself often provides no new information. Partial repetitions also exhibit similar properties to those of direct repetition. The contribution of standard deviation (represented by (m)) to partial repetition may thus be also due to the duration fluctuation related to partial repetition. Although the effect of standard deviation (feature represented by (m)) on direct repetition is not shown on Table 3, it indeed stands right behind (being the third important, not shown in the table), which supports the above argument.

9. Experimental results

9.1 IP detection with LPM-based models

Due to the limited quantity of the training data, we actually merged the four classes of IP into one and considered IP detection as a two-class classification problem in the experiments. For each syllable boundary, a decision between “non-IP” vs. “IP” was made with a probability. Fig. 5a compares IP detection accuracies obtained using the delicate utterance-type LPM-based models as shown in Fig. 2 to those using plain integrated maxent and anchor models, with the training data for the delicate model selected using hierarchical agglomerative clustering (HAC) and k-nearest-neighbor (kNN) approaches. We see that kNN-based approach is better and the delicate utterance-type LPM-based approach apparently improved the performance.

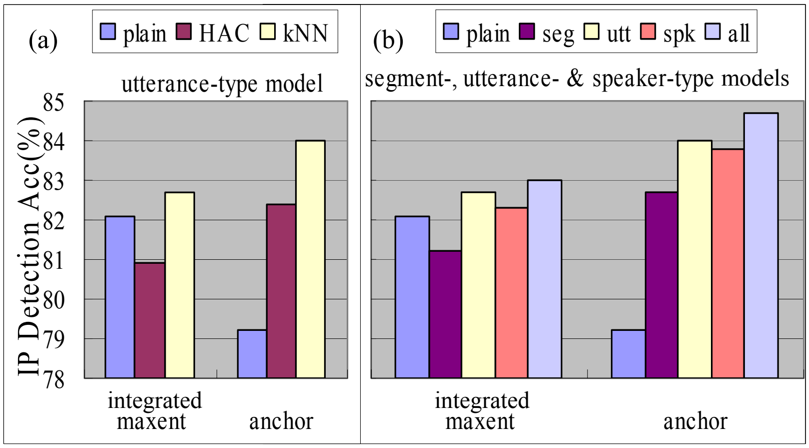


Figure 5: IP detection accuracy using LPM-based integrated maxent or anchor models, (a) with and without the utterance-type delicate models with training data selected by HAC- and kNN-based approaches. (b) with segment-, utterance-, and speaker-type delicate models with training data selected by kNN-based approach.

Fig. 5b then demonstrates the results when the delicate models of individual segment-, utterance- and speaker-types based on LPM were used, as well as all of them used together, all kNN-based. So the first and third bars in Fig. 5b for each case are the same as those in Fig. 5a. The improvements obtainable from LPM are obvious at different levels especially for the anchor model, and the use of all the three levels is clearly better. So the prosodic information from different levels is complementary. The relatively lower performance for the segment-type model may be due to the relatively poor segmentation by the pitch contours.

9.2 LPM-based feature expansion for integrated maxent

As mentioned in §3.3, LPM parameters $\{P(z_l \mid d_i)\}$ and $\{\prod P(t_k \mid d_i)\}$ can be used as extra features for the integrated maxent model. The results for such case are shown in Fig. 6, where the bar (a) is the same as the last bar for integrated maxent in Fig. 5b, and the bars (b)(c)(d) are the results when these features were added individually and together. We can see that the expanded features are indeed useful. The last bar (e) is the result when the finally enhanced integrated maxent model is combined with the anchor model by SVM, which eventually yielded the best result.

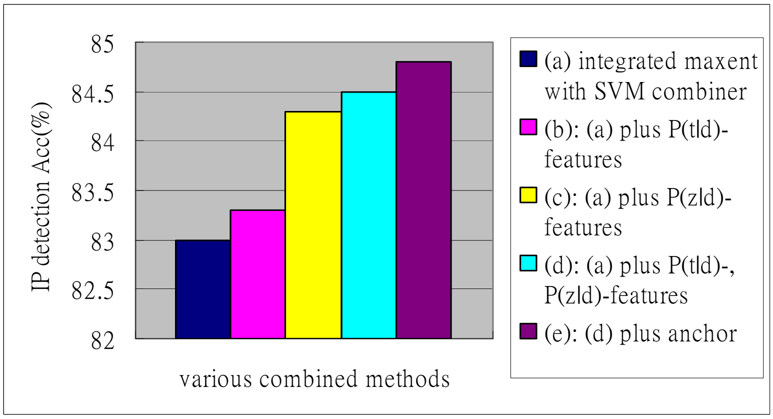


Figure 6: The performance of the integrated maxent model with expanded LPM-based features and finally integrated with the anchor model

9.3 Speech recognition results

The recognition experiments were performed with a lexicon of 50K entries, a trigram language model, and an intra-syllable right context dependent Initial/Final acoustic model set (a Mandarin syllable was decomposed into two parts: Initial and Final). Fig. 7 shows the character accuracy with IP detection results considered as a function of the weight parameters λ in equation (6), using the formula described in §7 in the rescoring process, compared to the baseline without the disfluency information. We see that the highest improvement achievable with the LPM-based IP probability is about 2% of character accuracy when λ is about 0.9.

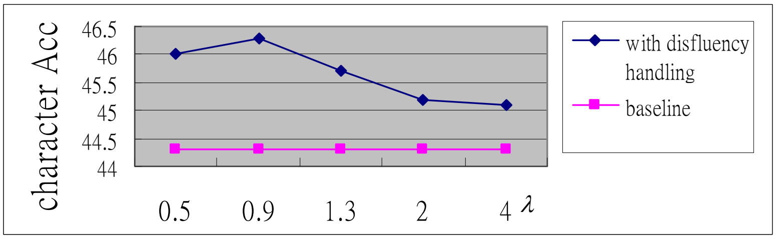


Figure 7: Character accuracy with disfluency IP detection.

10. Conclusions

We presented a new approach of modeling prosodic information in speech, LPM, for application in spontaneous Mandarin speech recognition with disfluency IP detection.

The LPM is a general approach for dealing with prosodic variation considering the latent factors not directly observable in speech signals. Experimental results showed improved performance when integrated maxent and anchor models were enhanced by LPM. The results also verified the benefit of embedding disfluency information in the recognizer.

Che-Kuang Lin
Graduate Institute of Communication Engineering
National Taiwan University
Taipei 106, Taiwan
kimchy@speech.ee.ntu.edu.tw

Shu-Chuan Tseng
tsengsc@gate.sinica.edu.tw

Lin-Shan Lee
lslee@gate.sinica.edu.tw

Linguistic Patterns in Spontaneous Speech, 213-239
2009-1-025-010-000272-1

Prosodic Similarities of Dialog Act Boundaries Across Speaking Styles*

Elizabeth Shriberg[1,2], Benoit Favre[2], James Fung[2],
Dilek Hakkani-Tür[2], and Sébastien Cuendet[2]

SRI International[1]
International Computer Science Institute[2]

1. Introduction

Spontaneous speech differs in a multitude of ways from read, formal, or laboratory speech (Maclay & Osgood 1959, Goldman-Eisler 1968, Levelt 1983, Biber 1988, Howell & Kadi-Hanifi 1991, Eskenazi 1993, Shriberg 1994, Swerts et al. 1996, Bruce 1995, Hirschberg 1995, Laan 1997). Although the labels "spontaneous" and "read" each reflect an underlying multidimensional space of different genres or "styles" (Eskenazi 1993), generally speaking, spontaneous speech exhibits greater segmental and suprasegmental variability than does read speech (Lieberman et al. 1985, Llisterri & Poch 1991, Wajskop et al. 1992, Kohler 1996, Greenberg 1999, Maekawa 2003, Tseng 2005, Benzeghiba et al. 2007). It is no surprise, then, that automatic speech processing techniques typically have more difficulty with spontaneous than with read speech (Weintraub et al. 1996, Greenberg 1999, McAllister et al. 1998, Riley et al. 1999, Ostendorf 2000, Binnenpoorte et al. 2004, Benzeghiba et al. 2007). For example, an early study in large-vocabulary speech recognition found degraded performance for spontaneous speech even when recording conditions, speaker, and word sequences were held constant (Weintraub et al. 1996). In the study, speakers read transcripts of what they had said in previous spontaneous conversations; the read versions were significantly easier to recognize than the spontaneous originals.

In this paper we focus on comparisons of prosody across speaking styles, specifically for the task of dialog act segmentation. Dialog act segmentation aims to segment the

* This research was supported by the Defense Advanced Research Projects Agency under contracts HI0011-06-C-0023 and NBCHD030010, and by IM2, a research network funded by the Swiss National Science Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies. We thank Harry Bratt for feature extraction software and Andreas Stolcke for helpful comments.

continuous speech stream into dialog act units, i.e., to find the boundaries of dialog acts such as statements, questions, and backchannels. The task is of particular importance for speech understanding applications. Dialog act segments are similar to sentence-level segments, which are required for semantic processing techniques, including machine translation, question answering, and information extraction. Most such applications are developed for text input and rely on the assumption that boundaries are marked overtly via punctuation or text formatting (Shriberg & Stolcke 2004, Mrozinski et al. 2006, Makhoul et al. 2005, Hakkani-Tür & Tür 2007).

When the application uses spoken language rather than text, the input is the stream of words produced by an automatic speech recognizer. Recognizer output typically lacks punctuation, and thus the locations of dialog act boundaries need to be recovered automatically. Automatic boundary annotation has been shown in various studies to aid automatic summarization, named entity extraction, machine translation, and part-of-speech tagging (Furui et al. 2004, Matusov et al. 2007, Hillard et al. 2006, Fügen & Kolss 2007, Rao et al. 2007). It has also been shown to aid human readability of the output of automatic speech recognition systems (Jones et al. 2005) and could be used for determining semantically and prosodically coherent boundaries for playback of speech to users in tasks involving audio search.

Studies of automatic dialog act or sentence segmentation have used lexical, syntactic, prosodic, speaker, and time-based features (Warnke et al. 1997, Shriberg et al. 2000, Kim & Woodland 2003, Liu et al. 2005, Ang et al. 2005, Liu et al. 2006, Kolar et al. 2006, Cuendet et al. 2007b, Batista et al. 2007, Dielmann & Renals 2006, Fügen & Kolss 2007, Matusov et al. 2007, Cuendet et al. 2007a). In many such studies, prosodic features have been shown to improve performance over lexical features alone and to have greater robustness than do lexical features to errors in speech recognition output. The complementarity of prosodic to lexical features for this task also lends itself well to procedures such as co-training, which can be useful when only small sets of boundary-labeled data are available (Guz et al. 2007).

To develop prosody features for automatic segmentation, one can look to the rich history of linguistic descriptions of such phenomena. The task of sentence or dialog act boundary detection corresponds most closely to the linguistics literature pertaining to major phrase boundaries. Prosodic properties of major phrase boundaries and related phenomena are discussed by a number of authors (Ladd 1980, Cutler & Ladd 1983, Vaissière 1983, Cruttenden 1986, Couper-Kuhlen 1986, Bolinger 1986, Bolinger 1989, Ladd 1996, Hirschberg 2002). General patterns for English (as well as for many other languages) include a pause at the boundary, preboundary pitch drop, pitch declination over the phrase, postboundary pitch reset, energy contours similar to pitch behavior, preboundary durational lengthening, and voice quality changes. Boundaries are also associated with particular boundary tones. Certain dialog acts are associated with specific patterns; for example, questions are often described as showing a preboundary intonational rise rather than a fall. Further insights, particularly in the area of boundaries at turn-relevant locations, are provided by work in conversation analysis, pragmatics, and discourse analysis (Sacks et al. 1974, Schegloff 1982, Atkinson & Heritage 1984, Couper-Kuhlen & Selting 1996).

Over the past decade, a line of research in computational processing has used a "direct modeling" approach aimed at capturing the phenomena from linguistic descriptions, such as those just mentioned, for use in automatic segmentation and other spoken-language processing tasks (Shriberg 2005). In the direct modeling approach (Shriberg & Stolcke 2004) no human annotation of prosodic events is required. Instead, features are extracted directly from the signal, and a classifier is trained to learn the relationship between the extracted features and the classes to be distinguished for the particular task at hand. Thus, instead of using phonological constructs (e.g., pitch accents or boundary tones) to find dialog act boundaries, the direct modeling approach uses sets of features designed to capture breaks in phrasing, such as pause information, local pitch slopes, or energy differences across candidate boundary locations. The approach generally uses a large set of prosodic features, some highly intercorrelated, and leaves it to a machine classifier to determine how to make best use of the available features. Because prosodic features do not depend on specific words, they offer the possibility of greater generalizability across different speech corpora than do lexical features.

A repeated finding, however, is that when automatic classifiers are used to predict sentence or dialog act boundaries from prosodic features, different sets of features are found to be useful for different corpora and speaking styles. Feature utility is determined by running machine learning experiments using subsets of features, or by determining feature importance when all features are available to the classifier, or via various feature selection approaches. Importantly, selected features vary depending on the data set, even if feature definitions, extraction methods, and classifiers are held constant (Shriberg et al. 2000, Liu et al. 2006, Cuendet et al. 2007a, Cuendet et al. 2007b).[1]

One explanation for the differences in prosodic feature usage across data sets is simply that inherent prosodic cues to dialog act boundaries differ across speaking styles. Given the literature on differences across speaking styles noted earlier, it would not be surprising if speakers adopted different strategies for marking dialog act boundaries prosodically, depending on the speaking context. For automatic processing, this

[1] We focus here on English language data. A separate but related controlled study of prosodic feature selection for three different languages has been recently reported (Fung et al. 2007). That study again finds differences in feature sets across languages, but no claims are made here about language-related prosodic differences.

hypothesis implies that dialog act segmentation approaches should use genre-specific models, thereby requiring some type of matched training data for the genre at hand.

An alternative hypothesis, however, is that there exist inherent prosodic similarities or invariants across styles, but that for reasons having to do with how experiments are conducted, such consistencies have not been immediately obvious. This possibility is of particular interest from an applied perspective. If features that help distinguish boundaries from nonboundaries are qualitatively similar for very different speaking styles, then it should be possible, in principle, to learn robust models that automatically detect dialog act boundaries across genres.

To explore the question of cross-genre prosodic feature similarities in boundary marking, we compare data from two very different speaking contexts—face-to-face meetings and read news broadcasts, as described in more detail in §2.1. The two contexts were chosen because they represent essentially opposite extremes on dimensions of naturalness and level of human-human interaction. The data sets also contain separate groups of speakers.

We examine prosodic features of word transitions for dialog act boundaries and nonboundaries, computed using matched procedures across corpora. We look at the *difference* between the two classes across corpora, as well as the distributions by class across corpora. We break down features by type, since it is conceivable that, for example, duration features may pattern one way but pitch another. In particular, pauses reflect both prosodic and discourse (turn-taking) factors, and the latter is certain to differ between conversational and read speech.

§2 describes the data, dialog acts, features, classifiers, and metrics used. §3 compares results of automatic boundary classification experiments for the two corpora, broken down by feature type. Inherent feature discrimination analyses are then presented in §4, for each of the prosodic feature types (pause, duration, pitch, energy). §§5 and 6 provide a general discussion and conclusions.

2. Method

2.1 Corpora and dialog acts

To represent a spontaneous speaking context at one end of the naturalness dimension, we examine data from the ICSI Meeting Recorder Dialog Act (MRDA) corpus (Shriberg et al. 2004, Janin et al. 2003). The ICSI Meeting corpus is a collection of 75 naturally occurring meetings, including simultaneous multichannel audio recordings and word-level orthographic transcriptions. This corpus has the advantage of being fully hand labeled for dialog acts and their boundaries (Dhillon et al. 2004). Participants knew each other, since they generally met regularly in their working environment. The meetings were not staged scenarios but rather actual meetings with goals related to the everyday research objectives of the participants. Meetings averaged about an hour in length, with a maximum of 103 and minimum of 17 minutes. We use a 73-meeting subset of this corpus that has been used in other studies of the meeting data (Ang et al. 2005, Kolar et al. 2006, Cuendet et al. 2007b), with the same split into 51 training, 11 held-out, and 11 test meetings. The held-out data is used for tuning of model combination and was selected to be roughly matched for distribution of meeting type and general statistics.

For purposes of this study, detailed dialog act annotations (Dhillon et al. 2004) are collapsed into the small set of orthogonal labels shown in Table 1, using a class mapping provided with the MRDA corpus. Each word transition is mapped to either a nonboundary (**n**) or a boundary (**s**, **q**, **b**, or **d**), where abbreviations are defined in the table.[2] Statements comprise the majority of the utterances containing propositional content. Questions include all forms, including yes-no, wh-questions, questions with declarative syntax, and tag questions. Backchannels such as “yeah” and “right” are typically only one or two words long and provide feedback that the listener is attending to what the foreground talker is saying, without taking the floor. Disrupted utterances include both speaker-initiated cut-offs (such as false starts) and cut-offs attributable to interruption from one or more other talkers. In analyses to follow, dialog acts in the meeting data are grouped into two classes, with statements, questions, backchannels, and disrupted utterances forming the “boundary” class. The disruption class shares characteristics with both boundaries and nonboundaries. Looking backward in time, it resembles a nonboundary. But looking forward, it is followed by the onset of a new dialog act and thus shares some characteristics with boundaries. For this work, we chose to put disruptions in the boundary class.

[2] There is a slight difference between the experiment and analysis sections with respect to the treatment of floor-grabbers and floor-holders (**f**, an infrequent class). In §3, such transitions are treated as boundaries, for historical reasons. Their treatment as boundaries versus nonboundaries makes little difference, however, in classifier experiments based on word recognition output, and we predict a similar lack of difference when using reference transcripts. In §4, they are treated as nonboundaries, arguably the more appropriate class, since by definition speakers intend to continue their utterance after producing these phenomena. We note also that dialog acts labeled as **z** (i.e., unlabelable because of inaudible words or other reasons) have been removed from both experiments and analyses.

Table 1: Dialog act classes and corresponding boundary and nonboundary classes. Candidate boundary locations in the examples are marked by "*"; corresponding preboundary words or word transitions are in bold face.

Abbreviation	Dialog Act	Example
Boundary		
s	statement	The new one is **better . ***
q	question	Is it almost **done ? ***
b	backchannel	**Uh-huh . * He's** done .
d	disruption	It's **my** – * **Thanks .**
Nonboundary		
n	nonboundary	The **new * one** is better .

For read data we examine data from the TDT4 English Broadcast News (BN) corpus (Strassel & Glenn 2003). The TDT4 corpus was collected by the Linguistic Data Consortium, and includes news stories from radio and television broadcasts, other electronic text, and web audio. In this study, we use a subset of read TDT4 English broadcast radio and television speech, mainly from professional news anchors. Orthographic transcriptions contain punctuation and include commercially produced transcripts for radio shows and closed captions for television programs. This style of read speech contains almost exclusively statements. The corpus is thus considered to have two dialog act boundary types: statement boundaries **s** (mapped from sentence-level punctuation marks) and nonboundaries **n** in all other locations.

Table 2 provides distributional details on the two data sets. Automatic classification experiments in §3 use training, held-out, and test sets. Feature distribution analyses in §4 use all data. Note, however, that unlike the case for automatic classification, which requires coverage of all word transitions in forced alignment output, analyses require that features be defined (not missing or undefined) for all tokens included.

As can be seen in Table 2, the different speaking styles differ significantly in mean sentence length, with sentences in meetings being only about half as long, on average, as those in broadcast news. The percentage of boundaries (relative to total boundaries) is thus higher in the spontaneous data. Meetings (and conversational speech in general) tend to contain syntactically simpler sentences and significant pronominalization. News speech is typically read from a transcript and more closely resembles written text. It contains, for example, appositions, center embeddings, and proper noun compounds, among other characteristics that contribute to longer sentences. Discourse phenomena also obviously differ across corpora, with meetings containing more turn exchanges, more incomplete sentences, and higher rates of short backchannels (such as "yeah" and "uh-huh") than speech in news broadcasts.

Table 2: Data set statistics. The BN corpus is not annotated for dialog act boundary subtypes; all boundaries are considered to be statement boundaries because classes marked by "–" have an estimated low occurrence rate.

	MRDA	BN
Automatic classification data sets		
Training set words	456,486	800,000
Test set words	87,576	82,644
Held-out set words	98,433	81,788
Vocabulary size (unique words)	11,894	21,004
Mean sentence length (words)	7.7	14.7
Distribution of dialog act boundaries	% of Total words	
Nonboundary (within dialog act)		
n nonboundary	86.55	93.20
Boundary (end of dialog act)		
s statement	8.62	6.80
q question	0.95	–
b backchannel	1.88	–
d disruption	1.99	–

2.2 Time alignments

Features are computed automatically based on time marks for words and phones. Since our interest is in inherent feature comparisons, we use forced alignments (rather than free word recognition) to avoid confounds attributable to differences in speech recognition performance rather than to the features themselves. To obtain time marks, audio waveforms were force aligned to reference orthographic transcripts using a state-of-the-art speech recognizer appropriate for each corpus (Zhu et al. 2005, Venkataraman et al. 2004). Because in this procedure words are constrained to reference words, accuracy is typically quite good. Errors can occur for specific examples, but since feature distributions are compared with other feature distributions using the same time-marking procedure, distances between the distributions overall should be largely unaffected. This consideration, coupled with the use of large amounts of data, makes it reasonable to assume that time-mark inconsistencies should show up as noise rather than bias results. The reference transcriptions of the BN data are quick transcriptions from closed captions and thus are expected to contain errors. We therefore use a flexible alignment procedure (Stolcke et al. 2006) that allows for the possibility of skipping or inserting words based on acoustic evidence.

2.3 Features

We associate each word transition with the word preceding it, as illustrated in Fig. 1. Features are extracted for each word transition, regardless of whether or not that word transition contains a pause. The features are based on any pause at the transition, on features of the word (or time window) preceding the transition, and on the word or time window following the transition. Thus, word-based features span at most two words (and any intervening pause). Longer spans could potentially yield gains, but they would also complicate matters because of the presence of short dialog acts: features that are more than one word away from a candidate boundary could pertain to a different dialog act.

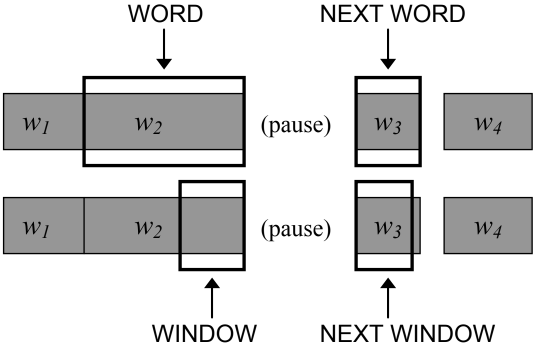


Figure 1: Illustration of feature extraction regions for the word-based (top) and window-based (bottom) features, shown for the word transition w_2, w_3, which happens to contain a pause.

2.3.1 Prosodic features

Prosodic features are extracted in the region of each word transition. The features were designed to capture breaks in temporal, intonational, and energy contours, based on the descriptive literature. Features were defined locally, since the segmentation into dialog act units is not known; one knows only that dialog act boundaries are constrained to occur at word transitions. And because of our interest in machine-based processing, features had to be extracted in a fully automatic manner, without reference to any hand labeling of prosody. Finally, for future robustness to automatic speech recognition errors, as well as for combination with lexical features, prosodic features were designed to be independent of word identities. That is, only a word's time marks, not its identity, were used for feature extraction.[3] Because of the large number of features, a general summary is provided here. Specific features of interest will be discussed in more detail in §4.

[3] Duration features, however, were normalized using information about phone-level duration statistics; strictly speaking, this uses phone identity (but not word identity) information.

Pause features included the duration of pauses as determined from recognizer forced alignment output. The pause model used by the recognizer was trained as an individual phone, which could occur optionally between words. Features included the pause duration (or 0, for no pause) at the transition, as well as that at the immediately preceding transition. Pause durations were not used for the following transition, because the presence of single-word dialog acts (such as backchannels) means that such locations may correspond to a different dialog act than that at the current word transition.

Duration features were intended to capture final lengthening before boundaries. Features included the duration of the last or maximum-duration (maximum after phone-based normalization) vowel or rhyme in the word (since lexical stress could be on other than the final syllable). We used both unnormalized and normalized versions of these features. Normalized versions were based on phone duration statistics compiled for the training data for the respective corpus.

Pitch features were based on frame-level output from a standard pitch tracker. We used an autocorrelation-based pitch tracker—the "get_f0" function in ESPS/Waves (ESPS 1993), with default parameter settings—to generate estimates of frame-level F0 (Talkin 1995). We then postprocessed the frame-level pitch to smooth out microintonation and pitch tracking errors, using median filtering followed by fitting using a piecewise linear model improved over that used in previous work (Shriberg et al. 2000, Sönmez et al. 1997). Features included the mean, maximum, minimum, first, or last value in the word or time windows shown in Fig. 1, as well as the value or sign (positive or negative) of fitted slopes in these regions. Time windows were either 200 or 500 milliseconds. Speaker-based pitch normalization used a "baseline" pitch calculated for each speaker from that talker's distribution of frame-level pitch values. The F0 distribution was modeled by three lognormal modes spaced log 2 apart in the log frequency domain. Locations of the modes were modeled with one tied parameter ($\mu - \log 2$, μ, $\mu + \log 2$), variances were scaled to be the same in the log domain, and mixture weights were estimated by an expectation maximization (EM) algorithm. Baseline pitch was estimated as the value occurring halfway between the middle and lower modes in the log domain, representing the lower end of the normal voicing mode for that speaker. The resulting measures were combined to create two types of features: those that look only at pretransition speech, and those that look at differences in pitch before and after the transition.

Energy features were based on frame-level RMS energy values from the "get_f0" function in ESPS/Waves (ESPS 1993), with default parameter settings, and were postprocessed using a piecewise linear model in a manner similar to that used for pitch regularization. Like the pitch features, energy features include mean, maximum, minimum, starting, and ending energy in the word or time windows shown in Fig. 1, as well as values and signs of fitted energy contours in these regions. Energy features were normalized based on the distribution of energy values from each speaker or channel and, like pitch features, include both pretransition features and difference features that compare energy values across the transition.

2.3.2 Lexical features

For purposes of comparison in automatic processing experiments, we also extract lexical features, which have been found to be helpful in previous work for similar tasks (Shriberg et al. 2000, Shriberg & Stolcke 2004, Liu et al. 2006, Cuendet et al. 2007b, Cuendet et al. 2007a). Lexical features are usually represented as *N*-grams of words. We represent lexical information using five *N*-gram features for each word transition, where "current" refers to the first word in the word transition:

- unigrams: {previous}, {current}, {next}
- bigrams: {current, next}
- trigram: {previous, current, next}

Because we use reference transcripts for prosodic features, for the reasons described earlier, we also use reference words for lexical features. Lexical feature results are thus optimistic compared to results using automatic speech recognition. Indeed, prior studies (Shriberg et al. 2000, Liu et al. 2006, Cuendet et al. 2007b) show that the degradation from errorful speech recognition output tends to be higher for lexical than for prosodic features for this type of task. This is particularly true for the meeting corpus, which shows a larger degradation than does BN when using recognized versus reference words.

2.3.3 Turn and overlap features

In experiments, a binary "turn" feature was included to capture locations of speaker change, rather than the more complex construct of a turn as defined in conversation analysis. The latter requires more sophisticated information than we can reliably extract automatically. For example, we do not have *a priori* knowledge of dialog act boundaries (they are, after all, what we seek to predict) or dialog act labels, so we cannot distinguish true interrupts from backchannels that do not disrupt another speaker's turn. The turn feature was necessarily computed differently for the two corpora. The meeting corpus records each speaker on an individual channel. Start and end times for talk from each speaker can thus be inferred from recognizer forced alignments on each of the individual recordings. In the broadcast news speech, however, a single channel contains multiple (generally nonoverlapping) talkers. Reference speaker information was not available for this data, so speaker change locations were estimated by an alignment between the output of an external diarization system (Wooters et al. 2004) and the reference words. Because the diarization system inserts a turn boundary at pauses greater than 500 milliseconds, the same treatment was applied to the meeting data for consistency in analyses.[4]

For analyses, particularly for conditioning of pause length distributions (§4.1), we also computed a more sophisticated feature based on overlapping speech in the region of the foreground speaker's candidate boundaries. The overlap feature asked how many other talkers (including 0) produced speech overlapping with the region extending from the onset of the foreground speaker's preboundary word through the end of the foreground speaker's postboundary word. We mapped a value of 1 or more overlapping speakers to "overlap" and 0 overlapping speakers to "no overlap."

2.4 Classifier

Dialog act segmentation can be seen as a binary classification problem, in which every word transition is to be labeled as either a dialog act boundary or a nonboundary. (More detailed models may distinguish among different types of dialog acts, such as statements versus questions.) For automatic classification experiments we use the AdaBoost algorithm (Schapire & Singer 2000), which combines weak multiple base classifiers to produce a strong classifier. In each iteration in the learning algorithm, a different distribution or weighting over the training examples is used, which gives more emphasis to examples that are often misclassified by the preceding weak classifiers. We use the BoosTexter tool (Schapire & Singer 2000); the weak learners are one-level decision trees.

[4] We note, however, that this treatment results in a higher rate of spurious turn marks in the meeting data than in news speech, because speakers in natural conversation often pause for that amount of time or longer within their turns. Since the pause feature is not affected by this treatment, but the turn feature is associated with increased false alarms, the effectiveness of the turn feature for dialog act boundary detection should be reduced for the meeting corpus (relative to what it would be were turns marked perfectly). Nevertheless, we will see later that the turn feature is still much more useful in meetings than in news speech for this task.

2.5 Metrics

We use two metrics in this study: (1) F-measure to evaluate automatic processing and (2) the Kolmogorov-Smirnov *D* statistic to quantify the difference between two feature distributions. F-measure, used in §3, is the harmonic mean of recall and precision. It is the metric traditionally used for various segmentation tasks. Note that it is an asymmetrical detection-based metric, dependent on the specific class of interest. Typically, this is the marked class (in our case, dialog act boundaries rather than nonboundaries). An F-measure of 1.0 indicates both perfect recall and perfect precision.

In the feature distribution analyses in §4, our goal is to discover which features provide best inherent separability between boundaries and nonboundaries. Since the features are not always normally distributed we use a nonparametric statistic. The Kolmogorov-Smirnov (K-S) test (Siegel & Castellan, Jr. 1988, Press et al. 1988) asks whether two data distributions differ significantly. We are interested in the value of the K-S *D* statistic, which measures the maximum difference between two cumulative distribution functions, and which can be compared directly across pairwise tests differing in sample size. The K-S statistic is

$$D = \max_{-\infty < x < \infty} |S_1(x) - S_2(x)|$$

where S_1 and S_2 are the two cumulative distribution functions to be compared: $S_i(x)$ is the percentage of the population in distribution *i* falling below *x*. It should be noted, however, that while larger *D* values reflect larger differences between two distributions, this does not guarantee that the two distributions can be separated by any specific classifier.

3. Automatic classification results

We first look at results from automatic classification using boosting with sets of features by feature type. Table 3 shows results in terms of F-measure. Chance performance is computed as follows. We assume that one has knowledge of the prior probability of a dialog act boundary in each corpus, since we know the average dialog act length in words. We compute this probability $p_t(s)$ for the training set and classify each word transition in the test set as a boundary with probability $p_t(s)$. The chance score is evaluated by computing the probability of each error and correct class and the ensuing value for the F-measure computation.

Table 3: F-measure results for automatic classification, by feature type(s) used in classifier. MRDA = meeting data, BN = TDT4 Broadcast News data.

Feature(s)	F-measure MRDA	F-measure BN
Chance	15.8	6.9
Duration	46.3	27.5
Energy	50.3	44.5
Pitch	53.8	50.5
Turn	63.9	36.7
Pause	69.5	63.7
All prosody	71.0	68.3
Lexical only	74.9	48.1
Combination (prosody + lexical)	82.0	76.9

As shown, in absolute F-measure, the meeting data has somewhat better performance than the news data on all feature types. It has particularly better performance for duration, turn, and lexical features. However, meetings also have a higher chance F-measure because of their shorter sentences. To adjust for the differences in chance performance, we should look at relative error reduction. Since the F-measure is a harmonic mean of two error types, one can compute the relative error reduction for a model with F-measure *F* and the associated chance performance *c* as

$$(1)\quad \delta = \frac{(100-c)-(100-F)}{100-c} = \frac{F-c}{100-c}$$

Results for relative error reduction are plotted in Fig. 2. A first observation about Fig. 2 is that despite the two very different speaking styles, recording conditions, speaker populations, and extremes of naturalness, relative error reduction is nearly identical for the corpora for energy, pitch, and pause features. Duration features are less useful than pitch or energy features, and pause features are the most useful individual feature type. The corpora differ in relative error reduction using duration, turn, and lexical features. Overall results when combining lexical and all prosodic features are similar.

Three of the feature types—duration, energy, and pitch—can be considered core prosodic features. It is surprising that energy and pitch features behave almost identically in degree of error reduction, given that read speech is typically considered more well-behaved prosodically. We will look more closely at these features in §4.

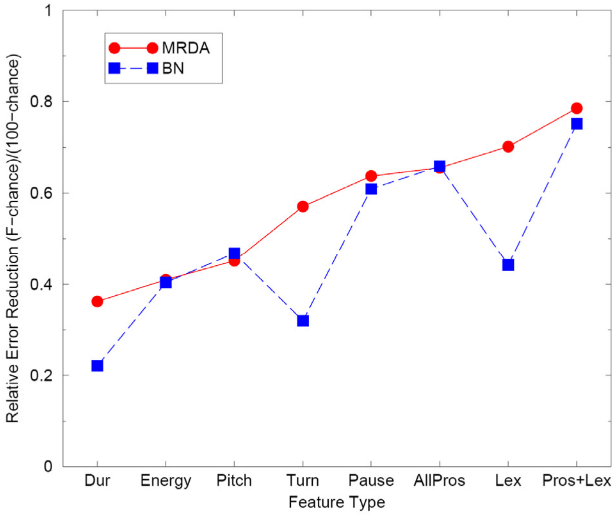


Figure 2: F-measure results by condition and features included in model. MRDA = ICSI meeting data; BN = TDT4 broadcast news data. Dur = duration features, AllPros = all prosodic features, Lex = lexical features only, Pros+Lex = combination of prosodic and lexical features.

Pause features reflect both prosodic patterns of the foreground speaker and speech activity from other talkers, particularly in meetings. It is interesting how close results are for the two speaking styles, particularly if one considers that pauses in meetings include time during which other speakers are talking, whereas pauses in news broadcasts typically do not. We examine the question of pausing further in §4.

As for the turn features (which are not prosodic *per se*), it is not surprising that they are more useful in meetings (in which there is more frequent alternation between speakers) than in read news broadcasts (in which anchors read many statements in sequence without a speaker change). It is also possible that there is some degradation for the news data from the use of turn labels from automatic diarization (since true speakers were not known), rather than from reference speaker information available via the separate recorded channels in MRDA.

Lexical features are more useful for boundary prediction in meetings than in news speech, a result consistent with previous work (Shriberg et al. 2000) that looked at spontaneous telephone speech rather than meetings. In the spontaneous speaking contexts, a small set of frequent words provides good cues to dialog act onsets. These include first person prounouns, certain fillers and discourse markers, and backchannels. In the case of backchannels, the dialog act consists of only the backchannel itself, so both the start and end of the dialog act are easy to identify. News speech, on the other hand, has far fewer of these elements, as well as significant noun compounding that can lead to phrasal ambiguity.

4. Analysis of feature distributions

The classifier experiments just described use sets of features and are affected by class priors. To understand inherent class separability by corpus and features, we need to look at individual feature distributions for boundaries and nonboundaries. We examine these distributions for the four prosodic feature types: pause, duration, pitch, and energy.

4.1 Pause features

As we just saw in Fig. 2, the most useful individual feature type for boundary classification in both corpora is pause duration. Here we examine more closely the behavior of the feature measuring pause duration at the word transition. Because a length of 0 should be interpreted as no pause, rather than as a very short pause, pauses are represented using a two-part feature: (1) presence of a pause (binary) and (2) length of the pause (if present). Furthermore, since the automatic word alignment approach used allows optional pauses between words, which can result in spurious short pauses, including those associated with stop gaps, we require a pause to have minimum duration of 50 milliseconds.

The pause feature is a rather special case, because in the meeting data it reflects both within-turn pauses and pauses that occur while a speaker does not have the floor (and presumably some other person is talking). That is, in the meeting data, since each speaker is recorded on a separate channel, pauses are simply those regions during which the foreground speaker is not talking. In the news broadcasts, speakers are recorded on a single channel, so pauses typically reflect within-speaker prosodic phrasing. Given that meetings involve more exchanges of speaker than do news broadcasts, we would expect the corpora to differ with respect to pause distributions.

Statistics on how often pauses occur at boundary versus nonboundary words are provided in Table 4.[5] Note that rates for boundaries and nonboundaries are logically independent. For the MRDA data, two versions of the statistics are computed. The “all” version counts pauses regardless of whether other talkers produce talk during the foreground speaker’s word transition. Thus, pauses in this group may be extremely long, since they count time elapsed during other talkers’ turns. In contrast, the “no overlap” version considers only those cases in which the foreground speaker is the only talker during the word transition in question. In BN, it is estimated that there is only one person speaking at a time (although there may be background speech or noise).

[5] The ratio of boundaries to total words can be computed and compared with those provided earlier. There are slight differences because the pause statistics require that the transition be defined. Undefined transitions occur when the preboundary and postboundary words are not from the same speaker, fail to align, or are at the edge of a recording.

Table 4: Rate of pauses for boundaries and nonboundaries. Pauses have a minimum duration of 50 ms. Percentage of (non)boundaries with pause is number of (non)boundaries with pause divided by total number of (non)boundaries. All = pauses regardless of other meeting participants' speech. No overlap = no overlapping speech from another participant during current speaker's word transition. BN is estimated to contain no such overlap.

	MRDA all	MRDA no overlap	BN no overlap (est.)
Total boundaries	103,839	38,917	88,608
Boundaries with pause	77,475	20,127	71,633
% Boundaries with pause	*74.6*	*51.7*	*80.8*
Total nonboundaries	666,372	543,844	1,214,394
Nonboundaries with pause	87,043	67,007	116,523
% Nonboundaries with pause	*13.1*	*12.3*	*9.6*

Two important observations can be noted from the table. First, in terms of overall rates of pauses for both boundaries and nonboundaries, the MRDA-all and BN conditions show rather similar results, with pauses in the 75-80% range for boundary words and around 10% for nonboundary words. This similarity means that from the perspective of individual channel recordings, meetings and news speech show similar pause rates, with boundaries about 7 to 8 times as likely to contain a pause as nonboundaries. The second observation is that when overlapped transitions in MRDA are removed, so that we consider only cases in which the current speaker is talking alone, there is a significant drop in the rate of pauses for boundaries—from about 75% to about 50%. That is, in meetings, only about half of nonoverlapped sentence boundaries contain a pause, a figure that may be surprising given canonical prosodic phrasing descriptions in linguistics. The fields of conversational analysis and discourse processing, however, may explain such cases via a phenomenon called "rush-through" or other mechanisms for turn retention (Schegloff 1982, Couper-Kuhlen & Selting 1996, Wennerstrom & Siegel 2003, Local & Walker 2004). Despite these statistics, pause duration is still the top feature in terms of the performance of automatic processing experiments, as we saw in Fig. 2. One explanation is that nonoverlapped transitions are actually the minority of boundaries in meetings. As can be construed from the table, only about 40% of boundaries are not overlapped.

To understand pause behavior when pauses do occur, we also look at the distribution of pause lengths at boundary versus nonboundary locations for the two data sets. In Fig. 3, distributions are normalized to unit area for comparison of inherent differences across different class sizes. Pause lengths are plotted on a log scale, since pause distributions are typically roughly lognormal across styles and languages (Campione & Véronis 2002).

Both corpora show a positive shift in pause length for boundaries as compared with nonboundaries. The BN data shows basically a shift, with otherwise roughly similar curves for boundaries and nonboundaries. The MRDA data, however, shows some interesting behaviors. When overlapped transitions are removed, the curve for MRDA boundaries looks quite similar to that for BN boundaries, after a positive shift. Thus, pause durations at boundaries are generally longer in meetings than in read speech but follow an otherwise similar distribution. When overlapped transitions are included, meeting boundaries show a striking positive tail, suggesting at least two underlying distributions, one of which must correspond to time during which other talkers are speaking. Further analysis by dialog act type reveals that the longer-duration pauses are associated mainly with boundaries after backchannels. That is, pauses after backchannels are long, because they correspond to the time during which another speaker has the floor. A final interesting observation concerns pause lengths at nonboundaries. In this case, the shape of the MRDA distribution (similar for overlapped and nonoverlapped transitions) differs from the shapes of the BN distributions and from the MRDA nonoverlapped boundary distribution. More specifically, it shows a broader and more positive range of values. We can hypothesize that the longer pauses in these nonboundary cases may correspond to hesitation pauses, which are rare in read news speech.

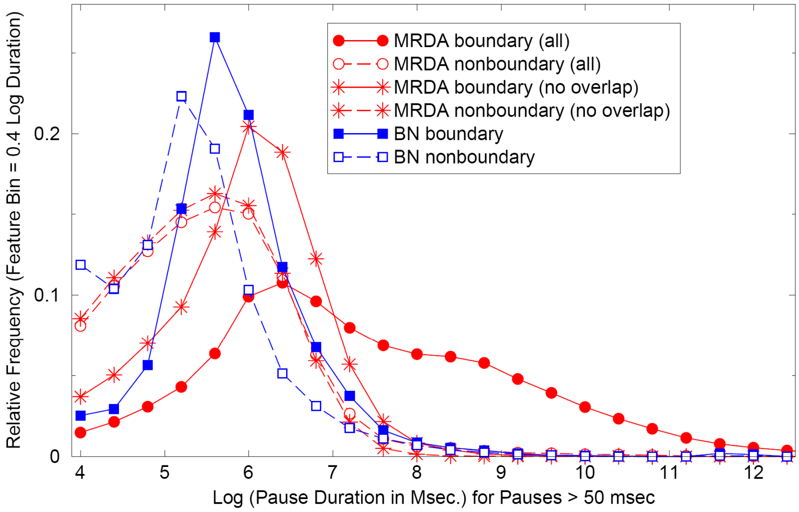


Figure 3: Distribution of pause lengths at word transitions that contain a pause (of at least 50 milliseconds), for both corpora. Length is plotted on a log scale; all distributions are normalized to unit area. MRDA is plotted in two ways: all transitions, and transitions during which no other meeting participants are speaking.

These findings on pauses show some similarities across corpora (rate of pause presence overall, shape of pause length distributions for nonoverlapped boundaries) but also differences (pause presence rates when overlaps are removed, shape of nonboundary distributions). Because pause features are typically the most powerful features for automatic classification, as seen in §3, the differences, as well as differences in corpus priors for boundaries and nonboundaries, are likely to affect results for all remaining features in machine learning experiments.

4.2 Duration features

We look first at features based on phone durations. Of the eight available features (based on rhymes versus vowels, maximum versus last syllables in words, and different normalizations) the best separation between boundary and nonboundary distributions is given by the feature RHYME_NORM_DUR_PH_MAX, which computes the mean of *Z*-score-normalized phone durations in each rhyme and selects the maximum value over all rhymes in the word. For both corpora, more separation is provided by rhyme-based features than by vowel-based features and by maximum-normalized-duration rhymes than by ending rhymes or longest rhymes in a word.

Fig. 4 shows distributions for RHYME NORM DUR PH MAX for boundary and nonboundary words in both corpora. In raw terms, news speech has slightly longer rhyme durations than meeting speech. Since durations in the figure are normalized based on the statistics of phones in the training data for each corpus, respectively, data for the majority class (i.e., nonboundaries) should look fairly similar across corpora. As shown, this is the case—although the meeting speech has a tighter distribution for nonboundaries than does the read data. The news speech shows a tendency for duration-increased nonboundaries at normalized durations between roughly 0 and 1. Listening analyses suggest that this behavior reflects a tendency for news anchors to produce frequent prominent syllables, including on nonboundary (and even function) words, perhaps to maintain listener attention.

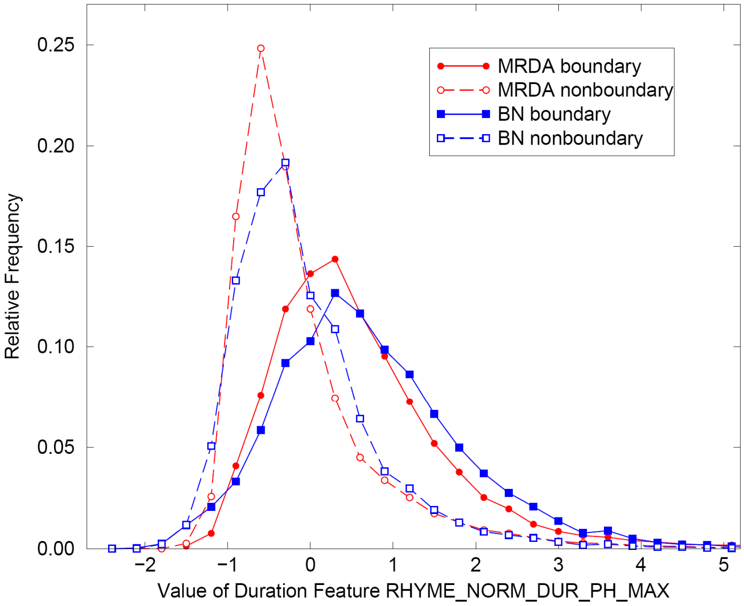


Figure 4: Distribution of the value of RHYME_NORM_DUR_PH_MAX, the maximum normalized rhyme duration in the word. Normalized rhyme duration is computed as the average of mean-and variance-normalized phone durations in the rhyme.

The most important and interesting observation from Fig. 4 is that the positive shift from nonboundary to boundary distributions is similar in both corpora. This is not simply an epiphenomenon of phone-duration normalization. Boundaries could have shown different variance versus shift patterns in the two corpora. Indeed, we might have expected spontaneous speech to show less preboundary lengthening and more variability in lengthening, but there is no evidence from the figure that this is the case. Instead, it appears that after adjusting for the difference in speaking rate via phone-based normalization, speakers use durational lengthening before boundaries to roughly the same degree in spontaneous speech as in read speech.

4.3 Pitch features

We examined inherent separation for 37 pitch features. Features were based on differences across the boundary, pitch relative to baseline estimated pitch, and pitch slopes based on lognormal fits. We look first at the top three for each corpus, based on the value of the K-S *D* statistic described in §2.5. This measures the difference between the distributions for the boundary versus nonboundary class within a corpus, by feature. Results are given in Table 5.

Table 5: Top three (out of 37 analyzed) pitch features for each corpus, based on K-S statistic for comparison of distributions for boundaries versus nonboundaries. Feature types: pitch diff = comparison of pitch level before and after boundary; pitch baseln = comparison of pitch level in word to estimated pitch baseline for speaker.

Corpus	Feature type	Feature	K-S *D*	Prob(*D*)
MRDA	pitch diff	WRD_F0K_DIFF_ENDBEG	0.303	0
	pitch diff	F0K_WRD_DIFF_LOLO_N	0.277	0
	pitch diff	F0K_20_20_WIN_DIFF_LOLO_N	0.265	0
BN	pitch diff	WRD_F0K_DIFF_ENDBEG	0.441	0
	pitch baseln	F0K_DIFF_LAST_KBASELN	0.433	0
	pitch baseln	F0K_LR_LAST_KBASELN	0.432	0

Rather remarkably, the feature that shows the most distributional differences between boundaries and nonboundaries is the same for both corpora: WRD_F0K_DIFF_ENDBEG. This feature is computed as the log ratio of the last good prestylized pitch in the last word before the boundary and the first good stylized pitch in the first word after the boundary. It is intended to capture pitch reset (from lower to higher pitch) in the case of boundaries, and thus should have a lower value for boundaries than for nonboundaries. Data is plotted in Fig. 5.

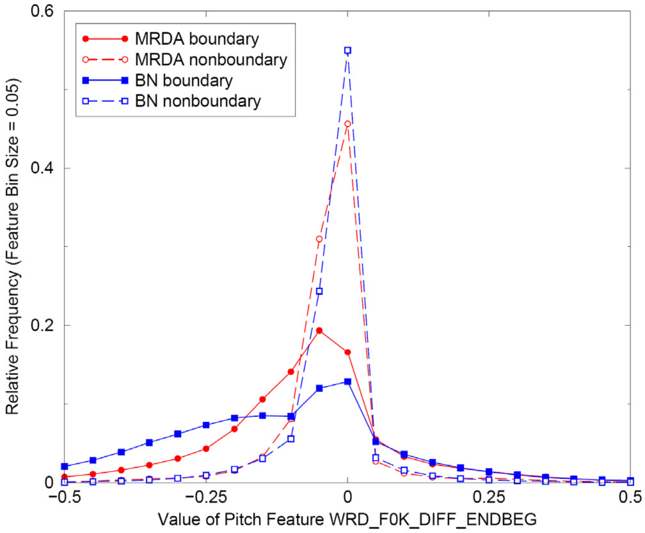


Figure 5: Distribution of the value of WRD F0K DIFF ENDBEG, the (natural) log ratio of the value of the last (fitted) pitch in the word to that of the first (fitted) pitch in the next word.

As predicted, values for the boundaries are lower than for nonboundaries. Four additional observations can be made from the figure. First, the nonboundary values are nearly identical for the two corpora; this is not a function of the feature computation, but rather reflects that nonboundary transitions in pitch are basically the same in read speech as in meeting speech—despite the different styles and also the different speaker populations in the two corpora. In general, this pattern of similar nonboundary distributions holds for many of the pitch features examined. Second, the boundary distribution for BN is more negative than that for MRDA, meaning the speakers in news broadcasts create larger pitch resets at dialog act boundaries than do meeting participants. Because of the increased value of baseline-related features in BN (discussed next), this is most likely attributable not only to higher starting pitch following a boundary but also to a drop to a lower pitch preceding the boundary. Third, both boundary distributions, and particularly the BN distribution, show a knee at roughly -0.1. The knee also occurs for some of the component dialog act boundaries and thus does not appear to be attributable to composition of different dialog-act-specific distributions. Although it requires further study, one possibility is that the knee reflects two categories of pitch distributions—one in which no pause is present (and thus reset is more constrained) and one in which a pause separates the pre- and post-transition words (and for which pitch jumps can be much larger). Finally, while the trends are similar across corpora, BN has overall higher discriminability. We will see why this is the case when we look at preboundary features, in Fig. 6.

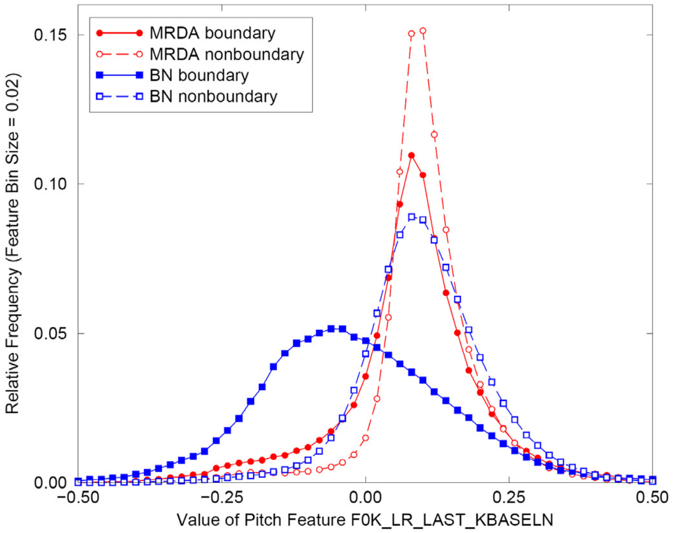


Figure 6: Distribution of the value of F0K LR LAST KBASELN, the (natural) log ratio of the value of the last (fitted) pitch in the word to that speaker's estimated baseline pitch value.

Looking back at Table 5, it can also be seen that after the shared best feature, the types of features in the two lists diverge. Features involving a speaker's baseline pitch are relatively more useful in news speech than in meeting speech for the boundary versus nonboundary distinction. The baseline features express the pitch range of the word in question by comparing it with a reference or baseline pitch for the speaker (Sönmez et al. 1997). This holds as a general trend when all 37 features are included: while both corpora have strong cross-boundary pitch cues, only read speech shows strong preboundary pitch cues.

This difference can be understood by looking at distributions for the preboundary pitch features. We choose the feature F0K_LR_LAST_KBASELN as an example, but results look quite similar for other features comparing the pitch of this word from other locations (e.g., mean, max, min in the word) with the baseline value. The feature F0K_LR_LAST_KBASELN is computed as the log ratio between the last good pitch value in the word preceding a boundary and the speaker's estimated baseline pitch.

As shown in Fig. 6, there is minimal separation between boundary and nonboundary words for the spontaneous speech, while read speech shows a negative shift for the boundary class. In meetings, data for both classes centers at about the same positive value as for the nonboundary class in news speech. These trends imply that before dialog act boundaries, speakers in read speech drop pitch to a value that is close to their baseline (on average, to slightly below it). In meetings, this preboundary pitch drop is much less common (only a small percentage of cases display more negative values than does the nonboundary class).

One reason for the lower prevalence of preboundary pitch drop in meetings may be that in spontaneous speech, pitch varies for reasons beyond phrasing, including paralinguistic factors such as affective state or emotion. Another reason may be that lexical information is a relatively stronger cue to dialog act boundaries in meetings than in news speech, as seen earlier. Phenomena such as first person pronouns, fillers, discourse markers, and backchannels all provide useful cues to dialog act onsets in meetings; such elements are much less common in news speech. Thus, there may be trading relationships between the lexical and prosodic cues. Finally, because dialog acts are, on average, shorter and less complex in meetings than in the news data, there may be less need to mark dialog act boundaries prosodically.

Preboundary pitch features are, however, among the best features in meetings for distinguishing different dialog acts—particularly for detecting questions. As can be seen in Fig. 7, questions show a shift to the right of the nonboundary distribution (rather than to the left, as in statement boundaries), reflecting the tendency of questions to end in higher pitch values.

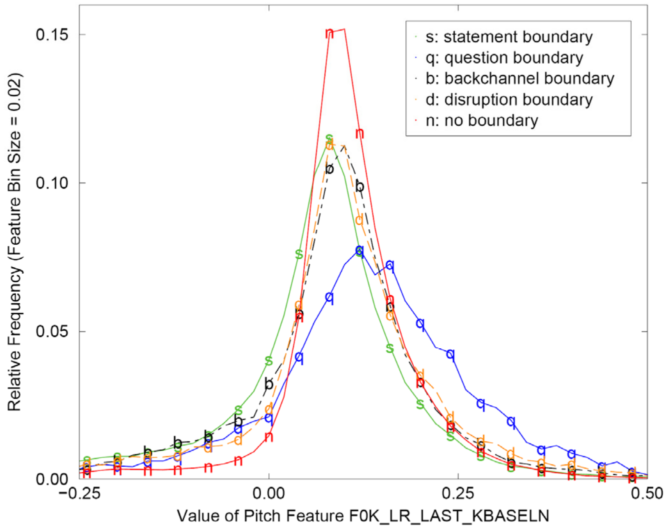


Figure 7: Dialog-act-specific distributions of the value of F0K_LR_LAST_KBASELN, the (natural) log ratio of the value of the last (fitted) pitch in the word to that speaker's estimated baseline pitch value, in the MRDA corpus. Dialog act labels are marked at every other bin.

4.4 Energy features

Out of 21 analyzed energy features, the same features appear in the top three for both corpora, as shown in Table 6. Unlike the case for pitch, good energy features require a comparison of pre- and post-boundary values in both corpora. Preboundary energy features (i.e., features using only one extraction point for feature computation) were weak in both the news and meeting data. This may reflect both the inherent nature of energy and the difficulty in normalizing energy effectively enough across different speakers and recordings to enable use of only a single extraction region.

Table 6: Top three energy features for each corpus, based on K-S statistic for comparison of distributions for boundaries versus nonboundaries. Feature types: energy diff = comparison of energy level before and after boundary; energy slope diff = difference in energy slope before and after boundary.

Corpus	Feature type	Feature	K-S *D*	Prob(*D*)
MRDA	energy diff	ENERGY_20_20_WIN_DIFF-HIHI_N	0.295	0
	energy diff	WRD_ENERGY_DIFF_ENDBEG	0.275	0
	energy slope diff	Slope_ENERGY_DIFF	0.272	0
BN	energy diff	WRD_ENERGY_DIFF_ENDBEG	0.398	0
	energy diff	ENERGY_20_20_WIN_DIFF_HIHI_N	0.306	0
	energy slope diff	Slope_ENERGY_DIFF	0.240	0

We look at the feature WRD_ENERGY_DIFF_ENDBEG, which has the best combined *D* value. Interestingly, this feature is the energy version of the pitch feature we saw earlier, which compared the pitch value at the end of the word in question with that at the start of the next word. For the present feature, fitted energy is used rather than fitted pitch. The distribution for boundaries and nonboundaries is shown in Fig. 8.

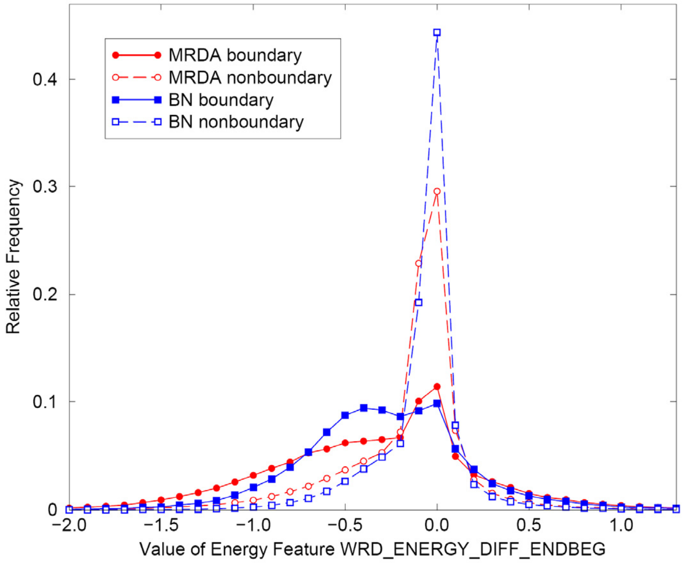


Figure 8: Distribution of the value of WRD_ENERGY_DIFF_ENDBEG, the log ratio of the last good fitted energy value in the word to that of the first good energy value in the next word.

The energy distributions look similar to those seen for the corresponding pitch feature. Knees in the distributions may reflect whether or not a pause is present at the word transition. As can be seen in Fig. 9, all dialog acts show a similar knee, except backchannels. This would make sense given the pause hypothesis, since backchannels have a high rate of following pauses.

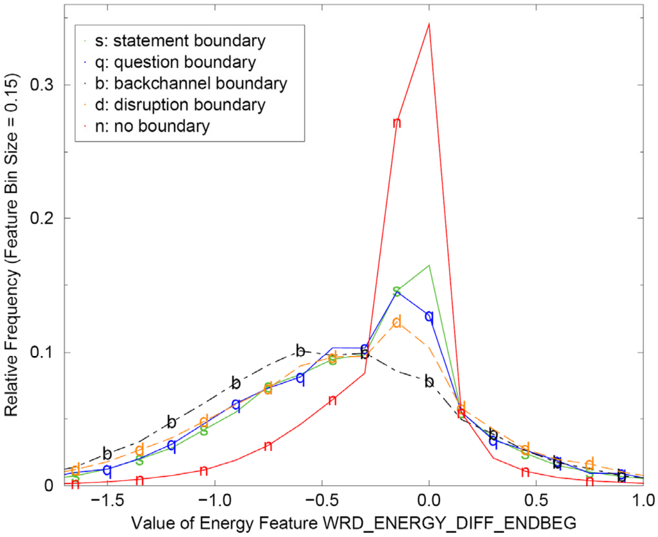


Figure 9: Dialog-act-specific distributions of the value of WRD_ENERGY_DIFF_ ENDBEG, the log ratio of the last good fitted energy value in the word to that of the first good energy value in the next word, in the MRDA corpus. Dialog act labels are marked at every other data point.

5. Summary and discussion

Results using feature groups in automatic classification experiments show that after adjusting for a measure of chance performance appropriate for F-measure analyses, meeting speech and news speech show similar dialog act segmentation performance for pitch, energy, and pause features. Perhaps counterintuitively at first, meetings show better performance for duration and lexical features. Lexical features are most likely more powerful in meetings because of the frequent occurrence of fillers, discourse markers, and first person pronouns at dialog act onsets. Turn features also differ, but in an expected manner given the different discourse contexts.

In analyses of feature distributions, it was seen that despite differences in turn-taking between the corpora, overall rates of the presence of a pause (of at least 50 milliseconds) were not that different when using very simple pause extraction definitions. Distributions of pause lengths given the presence of a pause differed, with longer pauses in MRDA, as expected, but the shapes and locations of the distributions suggest that a simple scaling could be used for adaptation. A small difference between the corpora in distribution shape for boundaries was explained by the distribution for backchannels, which show particularly long following pauses (consistent with their function of encouraging another speaker to continue talking).

Differences in the utility of duration features in the automatic experiments appear to be attributable, at least in part, to the tendency of news anchors to use longer durations for some nonboundaries, perhaps to keep the attention of listeners. When duration distributions are normalized for speaking rate, however, it appears that speakers in meetings use about the same amount of relative lengthening for boundaries as do speakers in news broadcasts.

Pitch and energy feature analyses showed remarkable similarities for the two corpora, both in terms of which features provided best inherent boundary/nonboundary separation within a corpus and in terms of the similarity of the feature distributions themselves. One difference between the two styles is that preboundary pitch drop appears to be more systematic in news speech than in meeting speech. Both styles, however, make good use of features that compare pitch across the transition, and these difference features also are the most robust for energy. Preboundary pitch is, however, important for distinguishing questions from other dialog acts in meeting speech (questions did not occur frequently enough in news speech to assess results).

The current study examined only two speaking styles, and clearly additional corpora and genres should be investigated to better understand prosodic consistencies for this task. Nevertheless, given the very different styles examined, results offer promise for improved approaches to cross-genre prosodic feature modeling for this task. Previous work in this domain has shown that if two speaking styles share characteristics, one can perform automatic adaptation from one style to another, to improve segmentation performance. An example has been shown recently for meeting data in a study of adaption using conversational telephone speech (Cuendet et al. 2006). Results such as those shown here could be used to help in selecting training data to better match characteristics of a test set. Measures of feature divergence could also be used more generally for feature selection.

But more importantly, we propose that feature distributions themselves could be shared. Classifiers will generalize to new data to the extent that feature distributions are invariant to genres, speakers, recordings, and other sources of variability. Features can possibly be made invariant by suitable normalizations, and normalization is facilitated by distributions that differ in simple ways—e.g., by simple scaling and shifts. A condition for effective normalization is that the parameters of the feature transformation can be estimated from data. For example, if the class-conditional distributions are unimodal and an optimal decision threshold is fixed relative to the overall distribution mean, it should be possible to estimate the decision threshold from the test data. An alternative approach would be to map the features from different domains to a common feature space, using techniques developed for channel compensation in speaker recognition (Reynolds 2003).

To benefit from such normalizations, one would also need to consider differences in class priors, which obviously affect the optimal classifier decision. Class priors can be gleaned from matched training data (using transcripts, with no prosodic feature extraction or modeling necessary). But in the absence of matched training data, this parameter could also be estimated in unsupervised fashion from test data. This could be achieved by running a preliminary classifier on a body of test data, using the outputs to update estimates of class priors, and reclassifying until estimates stabilize.

If features that differ across genres could be normalized so as to make their distributions similar for all genres, then classifiers could be trained on data irrespective of style and would generalize to new data mismatched for genre. An alternative approach could be to train the classifier to perform the normalization as part of the training process. In principle, this could be achieved by giving the data source as an additional input feature, so that the classifier could learn to adjust feature values to be usable for the same task across different genres.

6. Conclusions

Feature selection experiments using machine learning approaches yield different prosodic feature sets for dialog act segmentation, depending on the corpus. Yet we have shown herein that inherent prosodic feature distributions are remarkably similar across styles. This suggests, rather surprisingly, that prosodic marking of dialog boundaries in spontaneous speech is much like that in read speech. Differences in feature selection are likely a consequence of differences in class priors and in pause thresholds across corpora, which then affect usage of remaining features in automatic classification approaches. Clearly, other features, and especially other genres of speech, need to be investigated before definitive conclusions about the genre specificity of prosodic patterns can be drawn. In future work it will be worth investigating techniques for feature normalization and prior calibration tailored to the distributions studied here, with the goal of improving the robustness and generalizability of automatic sentence segmentation algorithms across speaking styles.

Elizabeth Shriberg
Speech Technology & Research Laboratory
SRI International
333 Ravenswood Avenue
Menlo Park, CA 94025, USA
ees@speech.sri.com

Linguistic Patterns in Spontaneous Speech, 241-261
2009-1-025-011-000273-1

Exploring Silence Application and Politeness Strategies in Interpersonal Business Communication[*]

Annie Wenhui Yang
Guangdong University of Foreign Studies

Much research on linguistic politeness remains at the level of the speech act, which is hardly surprising at all given overt picture of conversation analysis (CA). In face-to-face spontaneous interaction, silence, as a communicative act, has traditionally been ignored from the perspective of linguistics. My aim in this study is to develop a more ethnography-based, discourse-level approach to both silence and linguistic politeness and their interrelation with interpersonal relationships, and to address the paucity of research on the function of silence and its interpreting diversities. Based on the data of naturally-occurring trade negotiations, the findings have shown that (communicative) silence, like verbal speech act, can make people to do things, and is associated with politeness principles. Its meaning and function are determined by interlocutors' psychological motives, intentions, their assumptions of the imposition of face threatening acts (FTA) that the silence embeds under different interpersonal relations. The study further promotes Brown & Levinson's (1987) politeness theory on silence and Sifianou's (1997) proposal on silence manifestation of politeness superstrategies. In addition, it provides an insight into the developmental and cultural account of the traditional perceptions about Chinese silence through contextualized analysis, reflecting the complexity of both silence and politeness, and their complex relation with social and institutional contexts.

Key words: silence, politeness strategies, FTA, imposition, interpersonal relationship

1. Introduction

Conversation discourse involves verbal presences and absences. Silence as the verbal absences, either intentional or unintentional, being meaningful is undoubted. However, silence and its roles are not always well understood because interactants often fail to notice the imbalances of verbal speech and silence, except perhaps via a vague sense of discomfort, then silence gets attended and becomes meaningful. In Sino-Western business communication, the extensive use of silence in conversation by Chinese was

[*] I would like to express my appreciation to two reviewers from LPSS and LPSS participants for their comments on this paper. I am the person who is solely responsible for all the remaining omissions, errors, and shortcomings.

found negatively by Westerners to be exotic, mysterious, ambiguous, sinister, menacing and frightening. Therefore, in order to achieve a comprehensive and well-balanced understanding of cross-cultural business communication, we cannot afford to neglect the silence and its roles occurring therein in the negotiation text. In addition, a good understanding of silence can obviously reduce the misunderstanding and suspicion among the business negotiators so as to smooth the communication.

Despite its important features in conversation analysis (CA) and communication, silence has traditionally been ignored or less attended within linguistics. Although conversation has been proved to be one of the most fruitful areas of linguistic study, the great majority of research in this field tends to concentrate on the more tangible verbal aspect (such as speech acts), which is hardly surprising at all given overt picture of conversation analysis and communication. In conversation, it is important in terms of not only its physical appearance, but also its structures, meanings and functions, which jointly contribute to the interpretation and understanding of the conversation operation. Owing to its "complex dimensions and structure" (Saville-Troike 1985:4), Barnlund (1989) argues that silence is one of the most elusive of all communicative behaviors to describe and measure. The capacity of silence to provoke divers interpretations and ambiguities makes it "one of the highest forms of communication and yet one of the greatest sources of misunderstanding" (Liu 2002:39).

In this study, I am primarily concerned with the propositional silence (or named as "communicative silence"[1] by Saville-Troike 1985:6) which carries meaning and illocutionary force in business context, and it is performed by actors who vocalize nothing and may or may not use any contextual or textual clues to make the silence addressee to do things. This form of silence is usually produced consciously, which promotes (sometimes fails to promote) interaction in different ways, reflecting a variety of both positive and negative attitudes and face values. Given silence is treated as a discursive act rather than natural effect of the absence of expression, and based on Brown & Levinson's (1987) analysis of politeness phenomena on face threatening acts (FTA) and methodological framework of CA, I will discuss: how silence is interpreted, employed, and negotiated as the politeness strategy by Chinese business people? Are there any differences when silences are applied as politeness strategies by business people with different interpersonal relationships?

[1] The previous terms I used for the dimension of silence were "intentional silence" and "unintentional silence" which were more ambiguous and hard to classify and distinguish. In this paper, I adopt Saville-Troike's (1985) term "communicative silence" to refer to those consciously made silences which are interpreted by silence addressees as propositional ones with illocutionary force and perlocutionary effects.

2. Research background

Business communication is task-oriented and participants have to accommodate their conflicting interests into a mutually acceptable settlement, at the same time to create cooperativeness and competitiveness in the negotiating situation. The cooperative side is that both of the negotiators try to locate meaningful utterance to serve their understandings; and the competitive side is that both of the negotiators try their every effort to catch as many verbal and nonverbal units as possible to serve their negotiation aims and process. In business communication, silence as the most commonly used nonverbal element is largely used as an additional interpretative framework which allows people to overcome the shortcomings of verbal communication. Its rules, rites and usage are generally cultural-bounded although it may contribute to a psychological unease that makes communication more difficult.

2.1 Dimension of conversational silence

“Silencing takes place where there is discourse” (Thiesmeyer 2003:1). It has been discussed from many social perspectives (e.g. Dauenhauer 1980, Dendrinos & Ribeiro-Pedro 1997, Houston & Kramarae 1991, Jaworski 1993, 1997, Lakoff 1995, Mendoza-Denton 1995, Rotman 1987). Commonly, silence is inherently viewed ambiguous because of its symbolic nature, which complicates the situation even more (Saville-Troike 1985). Silence takes various forms, such as unnoticed cessation of sound, pausing, ellipsis, and hesitation within the stream of speech making up a speaker’s turn and between speakers’ turn. Silence is not opposite of speech act, but a communicative expression working with speech and forming a continuum of discursive interaction. As O’Connell & Kowal (1983) observe, silence is characterized by its “multi-determinism” in that its occurrence is determined by a multiplicity of factors, such as physical (e.g. breathing), psychological (e.g. embarrassment, weariness, anxiety, confusion), linguistic (e.g. syntactic complexity, temporary unavailability of lexical items), stylistic (e.g. emphasis), and interactive (e.g. interruption) ones and many other contextual factors. Thiesmeyer (2003:2) argues that silence can be the result of personal choice, but silencing clearly involves choices made by other people as well as by the potential speaker. To perform silence, the addressor needs to consider social norms and judgments, because it offers the chance to see how discursive actions operate within the social field. She proposes that to understand silence, people should look at not only the imposition of one discourse or another, but also the social and discursive boundaries among imposition, compliance and self-silencing.

Silence is composed of complex dimensions. As Saville-Troike (1985:4-7) sees it, it can be distinguished into two types. One is “the absence of sound when no communication is going on"; another one is "the part of communication". The first one carries meanings but not propositional content, such as pauses and hesitations, that occur within and between turns of talking—"the prosodic dimension of silence". These are non-propositional silences which may be volitional or nonvolitional, and may convey a wide variety of meanings. They are "connotative but not denotative". Their meanings are nonetheless symbolic and conventional with various patterns of use and norms of interpretation. The second one is "a communicative act" which is entirely dependent on "adjacent vocalization for interpretation, carrying its own illocutionary force". This type of silence can be one of the forms of speech act may take, filling many of the same functions and discourse slots—and should be considered along with the production of sentence tokens as a basic formational unit of linguistic communication. It can be used as questioning, requesting, promising, denying acts and many other speech acts as well as to carry out various kinds of ritual interaction. Thus, the second type of silence can be analyzed as having both illocutionary force and perlocutionary effect. Within this study, I only discuss the second type and its functions in the phenomenon of linguistic politeness.

2.2 Silence and politeness

Taking into account of many works and research on silence study, researchers seem to show less interest in the inter-relationship between silence and politeness. To my knowledge, this issue has barely been considered, even the most extensive and influential studies on politeness phenomena by Leech (1983) and Brown & Levinson (1987) who give silence very limited attention. Leech (1983:142) considers silence as "a special case of the maxims of agreement and sympathy since the topics chosen in phatic communication tend to be uncontroversial and sympathetic towards the addressee". Silence is largely viewed as an undesirable conversational practice which should be shunned in some way or other. Brown & Levinson (1987) provide a better seminal basis for the exploration of the relationship between silence and linguistic politeness although they only give a simple scant touch on the issue of the politeness encoded in silence. In their framework, silence is classified as the fifth category of "Don't do the FTA", which is ranked as the most polite behavior, whilst the first one is the most direct act because it does no face redressive work. The second and third are redressive and indirect to redress addressee's positive face to be close and negative face to be distant. Off record is even more indirect because the addressor does not do the impositive work explicitly but leaves clues to the addressee for interpreting the intended meanings. Strategies one to four are mainly concerned with verbal expressions, while strategy five and ellipsis in strategy four are related to nonverbal expressions—silence.

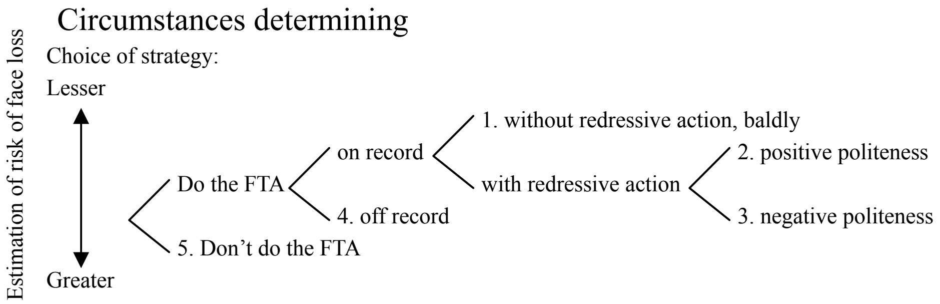


Figure 1: Possible strategies for realizing FTAs, source Brown & Levinson (1987:60)

Despite its importance in politeness study, Brown and Levinson's fifth superstrategy has not been given sufficient attention in politeness research, who simply rank it as the most polite strategy. But it could be problematic because Brown and Levinson's theory relies heavily on speech act theory and based on speaker-oriented analysis. The strategy of "Don't do the FTA" might be a vague option because silence might be an initiator of an exchange, or an indication of agreeing to redress the addressee's face want, or a face-threatening disagreeing or denying responses. Its function can be either acceptance or rejection, or it can be dispreferred seconds or redressive act to make the situation less threatening for both S and H (Sifianou 1997). In a very similar vein, Jaworski's (1993) claims that silence can be associated with the fourth off-record politeness superstrategy as non-conventional indirectness, yet Sifianou (1997:72-74) further points out that in fact "silence can be applied to perform many other politeness strategies". For instance, silence can be used as "a positive politeness strategy when it functions as a sign of solidarity and rapport", while it can also be "a negative politeness strategy if it functions as a distancing tactic". Besides, silence has "a positive value in avoiding imposition and can be the least polite form because it places high inferential demands on the addressee". She (ibid.: 71) claims that "overt polite behavior reflects underlying motives, but it is not always easy to detect what the exact motives are". They may reflect the interaction of politeness motivations and psychological factors, such as embarrassment and shyness. Although their findings have advanced Brown and Levinson's politeness theory, they still leave gaps, for instance, how silence achieve these superstrategies in naturally-occurring talks, and how social factors impact on the interlocutors' perceptions and applications of silence in interaction and its relation with politeness, especially under different interpersonal relationships. Furthermore, their findings are generally based on their intuition rather than empirical study. Thus, data-driven research is essential and necessary to be integrated to let us have a thorough and scientific understanding of silence and its politeness functions in social interactions.

3. Research method

As the major aim of this study is to examine the function of silence and its inter-relation with politeness in the discourse of institutional trade negotiations, various recorded trade negotiations happened at national trade fairs, factories, and companies among mainland Chinese trade representatives (2001-2006) are used for analysis. The corpus mainly consists of face-to-face conversations (1136 dialogues, 236 hours), involving 284 individual subjects from 25 provinces and cities. Although the cultural differences do exist geographically, they are comparatively insignificant since the data is only limited to Mandarin, so the sub-cultural differences are not considered in this study, neither are gender and individual differences. The transcribed data set for this study comprises 15 conversations (4:03’29”) equally distributed to three interpersonal business relations (business stranger, friend and partner), i.e. five conversations for each interpersonal relationship with similar length (15 minutes per dialogue), based on the conventions of CA which in large part developed by Jefferson (1984). Some additional data from the corpus is applied when there is a need. All subjects involved in the transcribed data hold college degree or above, with at least two years of institutional trade negotiating experience:

Business Stranger (hereafter BS): the business people who meet the first time, although sometimes they might have heard of each other or been referred to by other intermediary. Under this relationship, the communication is viewed as an initial interaction, and possibly the beginning of business.

Business Friend (hereafter BF): the business people who have met more than once and are known to each other, but without any business contract to bind their business obligations. Under this relationship, negotiators are communicating with each other with a goal to achieve contracted relation, aiming at the establishment of long standing and bonded institutional and cooperative duties. The relationship is under developing but is not matured for business.

Business Partner (hereafter BP): the business people who have met many times and have a (either written or oral) business contract to bind their institutional obligations. They have to frequently work together to continue and maintain their business continuum. Under this relationship, the interpersonal relations are stably and regularly established with contracted buying and selling duties.

In this study, interviews were conducted with the participants in order to get insight into the silence addressors’ and addressees’ perceptions on silences happened in the interactions subsequent to the recording. They provide both the author and the negotiation participants the chances to talk about the silence itself and its implicatures and functions regarding the politeness. The interviews proved to be very enlightening and rewarding for the analysis and interpretation of the silence discourse data.

Although it is important to assess the silence, whether propositional or non-propositional, many non-propositional silences may be interpreted by other interactants as impositive and face threatening ones according to interviews. Thus while I use the silence "strategy" for convenience in this study, the concept includes behavior which may be non-propositional yet could possibly be interpreted as propositional and impositive. Those being viewed non-propositional ones are not counted although some or even many of them may convey motives, meanings and functions. The classification of silence is based on its illocutionary force and perlocutionary effect that can be traced from participants' latte utterances. Double-check, to a large extent, is conveyed in interviews to confirm the accuracy of counting on silence.[2]

4. Silence as politeness strategies in trade negotiations

In my paradigmatic case, I only focus on "communicative" propositional silence in dyadic interaction of question-answer adjacency sequence where silence is either the response or unsaid speech act: the first or the second member of the adjacency pair. Based on Brown and Levinson's politeness frame and Sifianou (1997)'s findings, I classify the silences into five categories.

Bald on record: is the non-conformity with the Grice's Maxims.

Negative politeness: is to estrange the interactants with respect and deference.

Positive politeness: is to close the interactants with solidarity and involvement.

Off record: is done with a number of defensible interpretations and the decoding of the exact message is left to addressee.

Don't do the FTA: is an avoidance of doing any acts that embedded with face threatening phenomena.

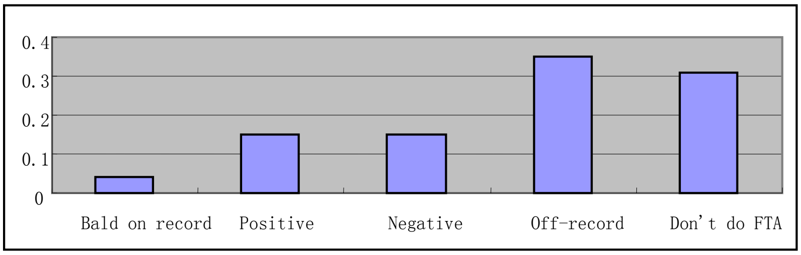


Figure 2: General distribution of silence as politeness superstrategies

[2] During the interviews, variations do happen regarding the interpretations, because some non-propositional silences addressed by silence addressors might have been counted because they are viewed and interpreted propositionally and impositively by the addressees in their later utterances. In this case, the perlocutionary effect of the silence is more emphasized in counting the propositional silence.

From the transcribed data, the most favored superstrategy is off-record (35%), followed by “Don’t do the FTA” (31%). Positive and negative are equally favored (15%). The least favored strategy is bald on record (4%), although it is widely applied by Chinese negotiators across three interpersonal relations in the additional data.

4.1 Bald-on-record strategy

According to Brown & Levinson (1987), the reason of doing bald-on-record is that S wants to do the FTA with maximum efficiency more than to satisfy H’s face to any degree. Concerning the talk exchange, it is usually the expected replier who uses silence as the violation of the maxims of quality, relevance and manner, who does not want to meet any of these conditions and be maxim-wise, but uses it as a strategy to impose FTA on the silence addressee. At the meantime, this silence also forges the first member of the second “question-answer” adjacency pair as a non-speech act with strong illocutionary force because it can actually make people to act by reflecting its perlocutionary effect. Silence (usually is stereotyped as indirectness) in some communicative context can be interpreted as rude, direct and face threatening act, such as unsaid directive act. As a communicating pattern, it is not often found in my transcribed data since the interactions with this type of silence mainly happens between BSs in short conversations or occasionally between BFs’ or BPs’ hours of talks, however, they present different features. The following two examples are under BS relation.

(1) (additional data*[3])

1 A1：你好！請問有什麼感興趣的嗎？
Ni hao! Qing wen you sheme gan xingqu de ma?
You good! Please ask have what feel interested ma-PRT (PRT= particle)?
Hello, may I ask, are there anything you are interested in?
2 B1：……(8.0) (walking away)
3 A1：慢走。
Man zou.
Slow go.
Goodbye.

[3] These sorts of interactions are usually shortly ended without any continuation between BSs, so they are not counted to be transcribed as one of the 15 conversations which last averagely 15 minutes per dialogue. So to explain the variety of silence, the additional data selected from the recorded corpus is used.

(2) (additional data)

1 A2：先生，有什麼可以幫忙的嗎？
Xiansheng, you sheme keyi bangmang de ma?
Sir, have what can help ma-PRT?
Sir, what can I do for you?
2 B2：… (5.0) 隨便看看。
Suibian kankan.
… (5.0) voluntarily look look.
… (5.0) I just want to have a look.
3 A2：好，您慢看。
Hao, nin man kan.
Good, you slow look
OK, take your time.

Silence in turn 2 in the two examples obviously functions as both a "response" to As' question and a directive "request" to As: "Don't bother me while I am looking at the products, or I don't want to reply you now". In the examples, As' greetings and inquiries are indeed polite, which are furnished with deference strategies, such as using interrogative sentence, politeness marker "請 please", and particle tail "嗎 ma-PRT" to address Bs' negative faces to be respected. However, Bs respond As with silences, followed by either gesture "walking away" or an utterance of Maxim violation (quality) "I just want to have a look". Without being frustrated by B1's silence and the potential face threatening imposed on him, A1 seems do not feel like to pick up the turn but take B1's silence as an unsaid reply and request, so he does not give any further utterances until A1 sees B1 walking away. The silence in Example (1) is impositive because B1 uses it as a non-redressed directive sign: don't bother me. Although B2 gives a reply finally after pausing in Example (2), but her reply also violates the Maxim of Quality as suggested by Grice (1975), which supports her intention of "don't want to reply you now". B2's intention is understood by A2 who gives explicit commissive utterance "take your time". In the two interactions, Bs do not honor As' politeness in the questions and give a reciprocal politeness payback: be verbally relevant and cooperative, but use silence to stop B's further inquiries. Presumably, this kind of silence is perceived as a bald reply: "I don't want to speak to / tell you" and projected as bald request "don't bother me" or "stop talking to me" and as such constitute the cognitive connotation, which is inextricably associated with specific patterns (e.g. the way of speaking: question => silence/request => reply) and motives (no face redress is necessary since they do not know each other). This sort of silence can terminate the interaction if no one wants to pick up the turn or continue the conversation. In this case, the bald silence inhabits multiple functions that not only may constitute the reply in the referential content, but also, given that the same reply, may contain another embedded meaning which is adjacent as a directive non-speech act.

Silence is applied bald-on-recordly, because such non-redressive act occurs when there is no need to redress the facework and the participants do not fear or care the retaliation or non-cooperation from the addressees. This form of silence has the features of imperative and directive speech acts which has clear meanings within the confines of the context, which is relevant to the participants' cognition of their relation (being distant) and the potential of business (being low).

For acquaintances or intimates, such communicative silence as the second member of "question-answer" adjacency pair is conventionally viewed improper, rude and unacceptable. Given the chances that it might cause irritation from the addressees, BF or BP negotiators prefer tracing the silence with repairs by the post-justification if they find the imposed silence propositional, meaningful and directive.

(3)

1 B： 別叫苦連天的，我又不是不付錢。
Bie jiao ku lian tian de, wo you bu shi bu fu qian.
Don't call bitterness together with sky de-PRT, I again not is not pay money
Don't pour out endless grievances. Surely, I will pay you.

2 A： 哼…
Heng
Hen…
Hen…

3 B： (3.0) 你什麼意思？
Ni sheme yisi?
(3.0) You what meaning?
(3.0) What do you mean?

4 A： 喔，沒什麼特別的意思，只是覺得…
Wo, mei sheme tebie de yisi, zhishi juede…
Wo, no what special meanings, only is feel…
Wo, nothing particular, I just feel…

5 B： (4.0) 覺得啥？
Juede sha?
(4.0) Feel what?
(4.0) Feel what?

6　A：　我的意思是你別老壓我們的價了。我覺得你每次都不停地壓我們的價，我們都快做不下去了。
Wo de yisi shi ni bie lao ya women de jia le. Wo juede ni mei ci dou bu ting di ya women de jia, women dou kuai zuo bu xia qu le.
My meaning is you don't always press our price le-PRT, I feel you every time all not stopping press our price, we all quick do no down go le-PTY
I mean you shouldn't always cut down our price. I think you are decreasing our quotation continuously, and we cannot carry on the business anymore.

7　B：　不會吧。
Bu hui ba.
No like ba-PRT.
I don't think it will.

In this extract, A and B are BPs but their business relation was primarily developed from BS relation, which means that their shared knowledge is largely limited to business. A's elliptical utterances in turn 2 after "Hen…" and in turn 4 "I just feel…" are frustrating and ominous to B because incompletion (ellipsis) in Chinese communication usually functions as propositional element by which leaves the floor to the interlocutors. If its interpretation can be figured out by shared knowledge, such silence may function off-recordly. However, if the shared knowledge cannot help to interpret the meaning, then such ellipsis may function as an unfinished directive act to direct the addressee to pick up the turn and carry on the speech, such as inquiring and requesting explanation and clarification for the ellipses. In this case, it is impositive since addressee is made to pick up the turn to speak. In this extract, A's silent motive thus is to arouse B's interest and to get his/her expected information by B's repair turns and A's post-justification. Repair of silence and ellipsis or verbal replacement of elided information is certainly face threatening because B can interpret it meaningful but is unable to make sense of the elliptical utterances. Besides, B's failing to have the shared knowledge with A to reconstruct the elided information is also viewed face losing and face threatening. No matter the ellipsis is intentionally made or its meaning is desperately expected or requested, face threatening is done respectively rather than to satisfy the addressee's face by verbal explicitness. Hence, the realization of silence (ellipsis) can not only make people to do things with maximum efficiency just like direct speech act by being impositive and doing FTA, but also convey addressor's intentions and motives that it's the addressee who should be responsible for the initiation of the repairs.

Therefore, in sequences of interaction in which silence of bald-on-record involves linking answer and a directive act in such a way that participants may come to the intended matters or meanings, which can be connected with the participant's either commissive act or repairs initiation for post-justification. The impositive and directive silence is distinct that it constitutes or contributes to the terminating or preceding the talk in which different types of relationship are negotiated.

4.2 Positive politeness strategy

Positive politeness is a redressive action directed to the addressee's positive face, his perennial desire that his wants should be thought of as desirable. This action "is widened to the appreciation of alter's wants in general or to the expression of similarity between ego's and alter's wants" (Brown & Levinson 1987:101). Tannen (1985) argues that silence of positive politeness reflects rapport and understanding without putting one's meaning on record. It is achieved through greater understanding of shared perspective, experience and intimacy, which according to her, is the positive value of silence stemming from the existence of something positive underlying. This is very much like what Brown and Levinson proposed linguistic realization of "ellipsis", indicating solidarity and common ground.

(4)
1 A： 怎麼樣？這幾年的收入翻了幾番吧？
Zenmeyang? Zhe ji nian de shouru fan le ji fan ba?
How like? These years' income doubled several times ba-PRT?
How are you doing? Your income must have been increased several times these years, right?
2 B： 哪裡呀，比去年還差呢。
Nali ya, bi qunian hai cha ne.
Where ya-PRT, compare to last year even bad ne-PRT.
Certainly not, it gets even worse comparing to last year.
3 A： 是嗎… (7.0) 怎麼了？又不管你借錢，急什麼？
Shi ma… Zenme le? You bu guan ni jie qian, ji sheme?
Is ma-PRT… (7.0) How le-PRT? Too not manage you borrow money, hurry what?
Really?… (7.0) Don't worry, I will not borrow money from you.

In Example (4), A and B are BFs and personal friends, who wish to start business based on their years' friendship. After B's response on A's conventional small talk on "income" in turn 2, A pauses for 7 seconds after positively polite comment "Really?" to indicate her care and consideration for B "gets even worse comparing to last year". Before finding the right concoction to serve both A's positive face of being close/considerate and B's positive face wants of being approval, A remains silent for a while to signal her understanding so as to avoid threatening to the addressee's positive face by uttering explicit but sympathetic words. Besides, this pause functions as a propositional device before another utterance: preparing the topic by presenting A' further involvement with B, A teases and jokes on B with "Don't worry, I will not borrow money from you", which is positively polite for "solidarity reasons" (Brown & Levinson 1987). Hence, the rapport for further talking is maintained and improved by the imposed silence for understanding and sympathetic expression.

Silence as a positive politeness strategy is contextualized and is measured against what is expected (rapport and understanding) in that context. Participants sometimes use verbal utterances after the silence to summarize or to elaborate the topic, or his/her understanding and consideration. Its function is to show one's desire to reinforce solidarity by emphasizing the positive self- or other-image that both addressor and addressee want to claim.

4.3 Negative politeness strategy

Negative politeness is "redressive action addressed to the addressee's negative face: his want to have his freedom of action unhindered and his attention unimpeded" (Brown & Levinson 1987:129). Tannen (1985:98) claims that when "silence reflects recognition and respect for the addressee's negative face needs not have his or her freedom of action interfered with, it is a manifestation of negative politeness". Through silence, addressors can protect themselves from potential intrusions if their interlocutors have performed a self-face-threatening act, such as making a *faux pas*. While it is positively polite to tease or joke about this (for solidarity reasons), it is negatively polite to ignore this action and remain silent (Brown & Levinson 1987, Jaworski 1993, Sifianou 1997). Here is an example:

(5)
1 A： 我這次只是想儘快擺平收支，彌補上幾次的虧空罷了。別怪我。
Wo zhei ci zhi shi xiang jin kuai baiping shouzhi, mibu shang ji ci de kueikong ba le. Bie guai wo.
I this time only is want soon balance income and expense, compensate last several times' losses ba-le-PRT. Don't blame me.
This time, I just want to balance our income and expenses and wish to compensate our last loss. Don't blame me.
2 B： 看你說哪兒啦？兄弟嗎，我們別計較那麼多。
Kan ni shuo naer la? Xiongdi ma, women bie jijiao name duo.
Look you say where la-PRT? Brother ma-PRT, we don't care for that much.
Certainly I will not. As brothers, we shouldn't calculate too much the cost and benefit.

3 A： … (10.0) 那＝
Na
… (10.0) Then=
… (10.0) Then=
4 B： ＝行了，我信你。500 台，75 元。
Xing le, wo xin ni. Wubai tai, qishiwu yuan.
OK le-PRT, I trust you. 500 sets, 75 yuan.
OK, I trust you. We will order 500 (sets) at 75 yuan per set.

In Example (5), although A and B are business partner, B, as the supplier, provided less support to A in the previous deals, which caused great loss to A. A complains about B's helplessness and intends to take great discount from B this time so that A's loss could be compensated. In turn 1, A tries to explain why the discount is needed, and ends his explanation with a negation request "don't blame me". B replies it with "certainly I will not" together with another request "As brothers, we shouldn't calculate too much the cost and benefits". The later request is impositive and face threatening because "we" here actually means "you", which makes it more complaint-alike. In this request, B performs a visible FTA in his utterance which implies A has asked too much discount from B and puts B at the edge of no profits. This is not the conducting way of business "brothers", which indeed should be blamed. To reply, A chooses a long silence in turn 3 as the response to B's blame to avoid saying anything negative—not confronting B, although A may have every reason and motive to argue back because A has suffered loss earlier in the previous deal because of B's helplessness, and that was not the way of doing business with brothers either. To redress B's negative face (the need not to be imposed on), A uses intended and meaningful silence as the defensiveness and expresses his reluctance to confront with B. Thus, silence applied here is to ignore A's impositivness, but has a positive value as a way of serving negative politeness (as suggested by Tannen 1985)—not imposing back on others.

Silence as a negative politeness strategy is contextualized too and measured against what is not expected (e.g. to reduce or eliminate confrontation) in that context. Participants normally do not repair this type of silence with justifications or explanations, but simply leave it untouched. Its meaning and function are self-evident and any elaboration would be viewed as redundancy or challenge, which, contrary to face redress, is impositive and face threatening to both addressor and addressee. Thus, the avoidance of speaking can function as a face-saver and negative politeness strategy to avoid verbal and explicit confrontation imposed on the addressee.

4.4 Off-record politeness strategy

Off-record strategy is essentially “indirect use of language to avoid the responsibility for doing the FTA but leave it up to the addressee to decide how to interpret it” (Brown & Levinson 1987:211). In this case, the meaning of silence is indeed ambiguous since the interpretation can be very dynamic, depending on various social and psychological factors and perspectives. Silence itself, to large degree, might be closely associated with off-record politeness since positive and negative politeness is usually enacted through or without the elaboration of redressive action. Sifianou (1997) proposes that silences of off-recordness in some contexts can give the addressee the opportunity to make unrequested offers; which is one of the motivations of off-record indirectness.

(6)

1 B： 王小姐是哪裡人呢？好像也不是廣州人。
Wang xiaojie shi nali ren ne? Haoxiang ye bu shi Guangzhou ren.
Wang Miss is where people ne-PRT? Seem either not to be Guangzhouness.
Where is Miss Wang from? You seem to be not Guangzhouness either.

2 A： 你猜猜看。
Ni cai cai kan.
You guess guess see.
(You) guess where I am from.

3 B： (2.0) (pause 1) 江浙一帶的？(5.0) (pause 2) 江西的？
Jiangzhe yidai de? Jiangxi de?
(2.0) (pause 1) Jiangzhe a area de-PRT? (5.0) (pause 2) Jiangxi de-PRT?
(2.0) (pause 1) You are from Jiangsu and Zhejiang area? (5.0) (pause 2) Or from Jiangxi province?

4 A： (4.0) (pause 3) 都不對。(6.0) (pause 4) 猜不著？＝
Dou bu dui. Cai bu zhao?
(4.0) (pause 3) All not right. (6.0) (pause 4) Guess no zhe-PRT?=
(4.0) (pause 3) No. neither one is correct. (6.0) (pause 4) Cannot get it?=

5 B： ＝猜不著，你沒什麼口音。
Cai bu zhao, ni mei sheme kouyin.
=Guess not zhe-PRT, you no what accent.
=I can’t get it since you don’t have accent (while you speak).

6 A： 我是湖南人，長沙人。
Wo shi Hunan ren, Changsha ren.
I am Hunanese, Changshanese.
I am from Hunan, Changsha.

In the above excerpt, A and B, who are BSs, meet the first time and they convey a small talk on birthplace in the middle of negotiation, intending to close their interpersonal relation. Whilst B's first silence in turn 3 (pause 1) may be interpreted as a hesitation that he is thinking and needs time to figure out the proper answer, his second silence (pause 2) seems to have different function. According to the classic turn-taking rule by Sacks et al. (1974), A is required to give response somehow after B's first guess to affirm it or to negate it. But A uses silence as the response and leaves its interpretation open: (a) the guess is incorrect; (b) go on guessing; (c) I don't want make you feel embarrassed by saying "No" directly; (d) you can ask tips from me if you want and so on. However, since A doesn't pick up her turn to respond, the length of A's silence seems to be long enough to entitle B to restore his turn and give his second guess because B interprets A's silence as response (a) and request act (b). From different angle, this silence (pause 2) can also be interpreted that B uses silence to (a) wait for the reply from A; (b) give signal "Am I correct?" (c) I have to think again since A gives no affirmation to the guess. Such non-vocalization period in fact involves both participants' motives and intentions, which are dynamic and open for interpretations. However, it can also be presumed that the silence (pause 2) may indicate the process of silence negotiation and the result of the negotiation as well, it suggests that B admits his first guess is not correct and he has to come up with a new one; and A doesn't want to embarrass B by verbal negation. Therefore, such silence can be interpreted as A's negating act and B's commissive act to continue the game, or it can be viewed as an admission of his failure in guess, which, in contrary, causes an overt loss of face on B himself for admitting failure and A's consideration of not being impositive. The silence in combination with the above indicators can be interpreted as an "off-record" politeness strategy imposed by both sides. In this case, the silence of off-record can be interpreted dynamically, embedded with various interactive aims and motives.

In fact, silence as off-record politeness can be found extensively in my study of business communication in China. Its meaning sometimes can be easily obtained if the shared information or common knowledge prevails, but if the relation is distant or shared knowledge is limited, the silences are tricky and evasive (as shown in Example (6)). Interlocutors have to infer the implicature of the intended FTAs through the cooperation and exchange of unsaid denial, request or commissive acts. Nevertheless, when this type of silence occurs, the business negotiators sometimes provide clues and hints as the presuppositions to guide their interlocutors to the "correct" interpretation.

4.5 Don't do the FTA strategy

Silence conveying little FTA is a basic assumption of Brown and Levinson's politeness superstrategies. As the fifth superstrategy, silence of “Don’t do the FTA”, which according to the predictions of the Brown and Levinson’s theory, encodes the highest degree of politeness. However, things can be problematic if concerning different acts with different facework. For instance, acts like compliments and offers, which not only encode face-threatening aspects for the addressee’s negative face because they may make their addressees incur debts, but also embed with face-saving or even face-enhancing aspects. For the previous one, silence is a more polite option than the production of such act, but for the last one, substituting them with silence cannot be the most polite option. Concerning either the positive or negative face wants of the addressee, the avoidance of expressions of disagreement, disapproval, contempt or any similar act is certainly an effective politeness strategy. Remaining silent in order to avoid the threat to the addressee’s face seems to confirm the high politeness value of the “Don’t do the FTA” strategy (Sifianou 1997). In Chinese trade negotiations, “Don’t do the FTA” seems to be only applied when Chinese negotiators refrain from certain expressions of disagreement, criticism, blame and complaints to their interlocutors. Here is an example:

(7)

1 A： 你可把我坑苦了知道啵，我跟你買得上一批貨 (1.0) 有些不對路，不，是很不對路。
Ni ke ba wo keng ku le zhidao bo, wo gen ni mai de shang yi pi huo you xie bu dui lu, bu, shi hen bu dui lu.
You but grasp me hole bitter le-PRT know bo-PRT, I and you bought last one lot goods (1.0) have some no right road, no, is very not right road.
You know you have made great trouble for me. The products I bought from you (1.0) are not of good quality. No, they are actually very poor.

2 B： 嘿嘿... (5.0) 是嗎？不對路？
Heihei Shi ma? Bu dui lu?
Hehe… (5.0) Is ma-PRT? Not right road?
Hehe … (5.0) Really? Poor quality?

3 A： 嘁…
Qi
qi…
qi… (exclamatory)

4 B： (4.0) 是嗎？說說 (2.0) 怎麼回事＝
Shi ma? Shuo shuo zenme hui shi
(4.0) Is ma-PRT? Speak speak (2.0) how round matter.=
(4.0) Really? Tell me (2.0) what the matter is.=

In example (7), after long discussion about the discount, A complains B's products for their poor quality so that it can frame his base for requesting the discount. Facing A's complaint of the poor quality of the products, B in turn 2 grins embarrassedly to save his self (embarrassing) face, then remains silent for 5 seconds but see A has no intention to pick up the turn, B then has no choice but restores his turn again to repeat A's complaint "Really? Poor?" to redress A's positive face, showing his understanding and willingness to continue the talk. A replies it with a complaining exclamatory "qi" in turn 3 without any further words, forming his unsaid complaint to avoid verbal confrontation to B. The silence after exclamatory "qi" shows that A does not want to further complain and challenge A with explicit utterance, which might make A feel worse. Too much explicitness in the complaint may cause serious damage to H's face which might affect their relation and further negotiation, since his complaining message has been literally transferred to B in the first turn. Silence of non-resistance, non-confrontation, or unwillingness of performing face-threatening disagreement and complaints thus could be considered as the chief construction of Chinese politeness strategy of "Don't do the FTA".

This type of silence of "Don't do the FTA" is only applicable in limited face-threatening contexts as unsaid but intended FTAs, which generally are viewed as highly and heavily face-threatening for being explicit, such as disagreeing, criticizing, complaining, and showing angry. The politeness orientation of being silent and "Don't do the FTA" appear to assume the maintenance of the facework that both of the participants hold and a rapport atmosphere through great restraint.

5. Silence application in trade negotiations

The corpus for this study consists of 124 instances of silence in the transcribed data (the minimum length is 1 second), which have been contextually made, noticed or interpreted by silence addressees across three interpersonal relations.

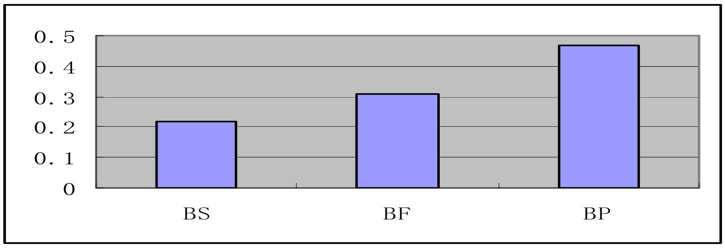


Figure 3: General distribution of communicative silences under three interpersonal relations

There are some apparent evidences that Chinese business people use more communicative silences in communication as the interpersonal relationship develops. The frequency of silences may predict that BS (27 instances, 22%) is more fluent or verbal orientated than BF (39 instances, 31%) in communication, while BP (58 instances, 47%) is the most hesitant and elliptical group in business interaction. Although such conclusion might not be accurate enough since, as I argued in the above, BSs do apply many communicative silences (such as bald-on-record silence) to ignore the face want their interlocutors hold in many short conversations. But concerning the longer interpersonal communication, the preference of silence applications is distinct.

Though the application of silence as politeness strategies to realize various unsaid FTAs is unarguable now, precautions should be taken that we should not consider these superstrategies of business negotiators as a fixed attribute. Instead, we should consider how politeness functions and how the meanings of silence are negotiated under various interpersonal relations.

Table 1: Silence application and politeness superstrategies in Chinese trade negotiations

Category	BS group	%	BF group	%	BP group	%
Total frequencies	27	100%	39	100%	58	100%
Bald on record	0	0%	1	3%	4	7%
Positive	1	4%	5	13%	12	21%
Negative	7	26%	8	21%	4	7%
Off record	3	11%	10	26%	30	52%
Don't' do FTA	16	59%	15	38%	8	14%

Despite silence of bald-on-record is less identified in the transcribed data, it does not mean it seldom happens in trade negotiations as I mentioned in the above. In fact, it usually takes place across all interpersonal relations, especially in short meetings between BSs who find the potential of getting down to business is low, and between BF and BP relations who don't want to redress H's positive and negative face want, although sometimes this sort of silence may be supported by paralinguistic sources (e.g. exclamatory). For the other notable superstrategies preferred by BSs, silence of "Don't do the FTA" (59%) is the most widely used by them, followed by negative strategy (26%) and off-record strategy. The reason might be that although they tend to consider less their interlocutor's face want, and face threatening seems to be irrelevant for them, they are sensible enough if they find the intended FTA too impositive (e.g. criticizing) which might bring unwelcome effects either for institution, personal image or future business. Hence, they prefer the strategy of "Don't do the FTA" to avoid the unnecessary confrontations by remaining silent or apply negative strategies without repairs, elaborations or justifications and off-record strategy by referring the common knowledge.

BFs are the most complicated group, who want to balance the face wants of both addressor and addressee, their various complicated intentions of wanting to do the FTA to get business done, and many other considerations of their delicate personal and business relations. Although silence of "Don't do the FTA" (38%) is the most favored strategy, other superstrategies, such as off-record (26%), negative (21%) and positive (13%) silences seem to be preferred by BFs too to observe their various face need consideration and interactive motives. The feature of being cluster in indirectness may constitute a better explanation on BFs' politeness orientations by using silence: redressive but multidimensional and polyhedral.

As the relation develops to BP stage, negotiators have shared more information and knowledge either personal or institutional about each other. Silence is mainly associated with off-recordness (52%) by leaving the interpretations open to addressees, which can be largely inferred from their shared knowledge and interacting experience, and positive politeness (21%) to show one's understanding and involvement. Thus, we can presume that the politeness orientation of silence by BPs chiefly exemplify their relationship backgrounds and accumulated information, which are more predominant in helping to interpret the silence application and its motivations.

To a large extent, the interpersonal context and business relationships play important roles in making and comprehending the silence. The important factors in this regard are relationship management, perception on the intended FTAs and interlocutor's multiple intentions and motives which construct the primary reasons for negotiators' being silent in the first place. The silence addressee's recognition on the relationship and addressor's psychological motives prompts the addressees to look for the ulterior meanings and functions of the contextually applied silence.

6. Conclusion

This research has achieved two goals: descriptive and theoretical. The former is realized through a discussion of the rich and dynamic uses of silence as communicative strategies by Chinese business people. At discourse level, the analysis reveals that silence should not be restricted to the superstrategy of "Don't do the FTA". The production of silence applied in Chinese business community in fact can perform all the politeness strategies identified by Brown and Levinson contextually and textually. Silence can be very direct and impositive if it functions as a directive act to make the addressee to do things; silence can be extremely polite if it functions as "doing nothing impositive", although it might not always be viewed polite, for example, there is a need to compliment and praise the other interactants. Silence can be applied by addressor to distance her/himself from others to redress the addressee's negative face for respect, whilst silence can be applied too to affiliate to their interlocutors through redressing addressee's positive face want for understanding and creating rapport. From different perspectives, either the contribution of the addressor/addressee or the dynamics of evolving interpersonal relationships, the interpretations of silence vary. In addition, the business negotiators with different interpersonal relationships have different preferences of silence, different perceptions on silence, its meanings and its interrelation with politeness.

Theoretically, the findings of this study further promote Brown and Levinson's politeness theory and its relation with silent behavior in communication and Sifianou's (1997) proposal on silence manifestation of politeness superstrategies. This study offers a new perspective in the developmental and cultural account of the traditional assumptions about silence through contextual analysis, reflecting the complexities of both silence and politeness, the complex social relations between the interlocutors and their different perceptions on the imposition and facework embedded within the silence. It behooves me to argue that, like verbal speech act, silence can make people to do things, imply meanings and functions contextually. It is a multifaceted linguistic construct, serving different face wants, cognition, and social conventions for politeness and relationship negotiation.

To conclude, Chinese business people, who use silence as a strategic token in task-oriented business communications, do have expectations when it is appropriate to talk and when it is appropriate to be silent in conversation under various social relations. In simultaneous settings, the prototypical form of language use—fluent speech without naturally occurred or intentionally made silences, gaps, or pauses is rare. The silence engagement with different kinds of interpersonal relations and politeness strategies is a crucial aspect of the art and practice of business negotiation, which potentially fracture the prevailing cultural and communicating ethos. Nevertheless, the silence discourse conveyed here suggests that the intricate blend and interplay of politeness and silence guide Chinese silence discourse, which may be realized with different intentions and motives. These features, in turn, allow room for interpretations or post-justifications.

Annie Wenhui Yang
School of English for International Business
Guangdong University of Foreign Studies
Guangzhou, Baiyun Dadao Bei 2#, 510420, China
annieywh@yahoo.com

Linguistic Patterns in Spontaneous Speech, 263-274
2009-1-025-012-000274-1

Recognizing Local Dialogue Structures and Dialogue Acts[*]

Kenji Takano and Akira Shimazu
Japan Advanced Institute of Science and Technology

One of the important abilities that an intelligent agent must have in order to appropriately communicate with human in natural language is to recognize a dialogue act conveyed with a human utterance. There are basically two methods to recognize dialogue acts: the plan recognition method, and the linguistic-based method. Especially as dialogue corpora increase nowadays, linguistic-based methods such as the n-gram model become popular. For dialogue agents to understand dialogue, it is desirable to recognize a chunk of dialogue act which constitutes a subdialogue structure. However, the popular n-gram model is weak in treating subdialogue structures.

This paper presents a method which recognizes local dialogue structures and dialogue acts of utterances by analyzing local dialogue structures. We define a local dialogue structure based on utterance units (Dousaka & Shimazu 1996) and grounding (Traum 1994) (Traum & Nakatani 1999). Based on the analysis of discourse structure in the corpus, we defined rewriting rules which capture patterns of local dialogue structures, and whose terminal symbols are dialogue acts. Using such rules, we can analyze dialogue acts of utterances and local dialogue structures. Our preliminary experiment shows the effectiveness of the method.

Key words: dialogue understanding, dialogue act, dialogue structure

1. Introduction

There is much research aiming to build natural language dialogue interfaces which can be used easily by anyone. Although such research has been briskly carried out in recent years, present systems are still poor, compared with dialogues between humans. In order to build a computer system that can interact with a human in a friendly way in natural language, various problems must be solved. Recognition of dialogue acts is one among them. A dialogue act is the method performed by a speaker, and which conveys the speaker's intention to a hearer.

There are basically two methods to recognize dialogue acts: plan recognition and linguistic-based method. Plan recognition uses some plan operators in order to infer the aim of the speaker and the intention behind the utterance. Linguistic-based method

[*] This research is supported by KAKENHI of MEXT.

indicates surface utterance intention using syntactic information, pattern of grammar and otherwise. As dialogue corpora increase nowadays, linguistic-based methods such as n-gram model are receiving much attention. Though the n-gram model is popular, it is unable to treat subdialogue structures, whose recognition is required for understanding dialogue (Litman & Allen 1990).

Assuming relationship between dialogue acts and discourse structures, we are analyzing a spoken dialogue corpus to find relations between dialogue acts and structures (Shimazu, Taguchi & Kawamori 2000) (Takano & Shimazu 2004). In route guidance dialogues, we found a sequence of dialogue acts with a pattern which corresponds to exchange units (Coulthard & Brazil 1981) or discourse units (Traum 1994) (Traum & Nakatani 1999) approximately. Though a discourse structure of a dialogue as a whole is ambiguous, a local dialogue structure is not generally ambiguous. Accordingly, a local dialogue structure is used to determine its constituent dialogue acts.

This paper presents a method which recognizes dialogue acts of utterances by analyzing local dialogue structures. We define a local dialogue structure based on utterance units (Dousaka & Shimazu 1996) and grounding (Traum 1994, Traum & Nakatani 1999). Based on the analysis of discourse structure in the corpus, we defined rewriting rules which capture patterns of subdialogues and whose terminals are dialogue acts. Using such rules, we can analyze local dialogue structures, and accordingly, dialogue acts of utterances. Our preliminary experiment shows the effectiveness of the method. In this paper, we focus on investigating to what extent dialogue acts are recognized using only linguistic information, and we do not use prosodic information.

The next section describes the previous research regarding recognition of dialogue acts. In §3, we explain the utterance unit, the dialogue act tag set, local dialogue structure and local dialogue structure rules. They are employed to recognize local dialogue structures and dialogue acts. Section 4 describes experiments.

2. Recognition of dialogue acts

In order to process a dialogue smoothly, it is indispensable to understand a dialogue act. For example, “hai” in Japanese is an affirmative answer act to a yes/no question, or a response as a backchannel utterance. If a system misrecognizes this act, it misunderstands the intention of the speaker, and the dialogue will become strange.

Litman and Allen showed a method for recognizing dialogue acts of each utterance as recognition of a discourse intention using plan recognition (Litman & Allen 1990). They dealt with the dialogue of a traveler and a station employee in a train station. They showed the use of surface linguistic cues.

Nagata and Morimoto presented a trigram model as a method of recognizing speech acts (Nagata & Morimoto 1994). Their model achieved 39.7% prediction accuracy for the top candidate, and 61.7% for the top three candidates.

As a method using a n-gram model, there are Reithinger and Maier's method (Reithinger & Maier 1997). They used VERBMOBIL corpus. The best results were obtained by using linear interpolation of uni- and bigrams. Their method recognized dialogue acts correctly 65.18%.

Samuel et al. showed automatic tagging of dialogue acts using Transformation-Based Learning (TBL) (Samuel, Carberry & Vijay-Shanker 1998). They used the same training set as Reithinger. The accuracy of the TBL model is 71.22%, and the amount of training data is far smaller than the N-gram model of Reithinger and Maier.

Koh and Shirai presented a dialogue act tagging tool (Koh & Shirai 2005). The tool makes a decision tree using C4.5 algorithm, and uses mainly surface linguistic information and dialogue acts just before an utterance. They applied the decision tree to route guidance dialogues by telephone. The best accuracy of their experimental results was 72.8%.

3. Model based on local dialogue structure

Our method recognizes local dialogue structures and dialogue acts using local dialogue structure information and surface linguistic information. For the surface information, we use the dialogue act tagging tool (Koh & Shirai 2005).

The corpus which we used is the transcribed text from route guidance dialogues in telephone. It is similar to that used by (Koh & Shirai 2005). We divided each dialogue into utterance units, and tagged them with dialogue act and local dialogue structure tags as described below.

3.1 Utterance unit

In dialogues, a speaker does not always tell necessary information in one sentence, unlike sentences in texts; rather he divides information into phrases or clauses to convey a message. Such a phrase or a clause is called an utterance unit (Dousaka & Shimazu 1996). We assume that a dialogue system processes an utterance unit as a basic unit, and understands each utterance unit incrementally. When we divide an utterance into utterance units as defined by (Dousaka & Shimazu 1996), each unit generally corresponds to a dialogue act, and we adopt this unit in order to deal with dialog acts.

In analysis of transcriptions of spoken dialogues, we divide an utterance into utterance units based on the following conditions, in accordance with (Dousaka & Shimazu 1996): (i) a clause is an utterance unit, (ii) an interjection is an utterance unit, (iii) a filler shows the start of an utterance unit, and (iv) an repair utterance shows the start of an utterance unit.

In this paper, we do not focus on monologues, however treat only dialogues. There are cases in which an utterance unit may be interrupted by utterances of a dialogue partner, and an utterance unit is distinctly divided into multiple units. We call such an utterance unit a pseudo-utterance unit, and treat each divided utterance as an utterance unit.

3.2 Dialogue act tag set

We attached dialogue act tags to utterance units based on the standard dialogue act tag set (Araki et al. 1999). Table 1 and 2 show the tag set which we used. An utterance with two functions, Reply and Initiation, we expressed by "[Reply tag] / [Initiation tag]."

We deleted "Idiomatic utterances of start/end tag" that did not be used, and added two new tags, "Backchannel" and "%", to the tag set. When an utterance like "Hai (yes)" is replied to an initiation utterance, it may be aimed at making dialogue smooth, rather than functioning as a dialogue act. As in the case of a filler, a backchannel is not treated as a dialogue act in some papers. However, we tag such an utterance with "Backchannel" as a special case, since in some cases it is difficult to distinguish backchannels from agreement utterances which are represented in the same words. "%" is also a special tag. It is used when an utterance is divided into multiple utterance units, and an utterance unit among them is only a particle or a pronoun, which does not function as a dialogue act.

Table 1: Standard dialogue act tag set (translated from Araki et al. 1999) and added tags 1

Tag name	Definition
Initiation (function to start new exchange.)	
Suggestion	The demand of an act to a hearer. A hearer is not required returning yes, no, or other response.
Request	The demand of an act to a hearer. A hearer must return yes, no, or other response.
Proposal	The proposal of the act performed by both speaker and hearer. A hearer is not required returning yes, no, or other response.
Invitation	The proposal of the act performed by both speaker and hearer. A hearer must return yes, no, or other response.
Confirmation	The question whose speaker has had a prediction of hearer's response by a context or certain knowledge.
Yes/No-question	The question whose speaker has not had a prediction of hearer's response. A hearer answers by Yes or No.
WH-question	The question whose speaker has not had a prediction of hearer's response. It requires a certain value or expression as a response.
Promise	The proposal of a speaker's act.
Hope	A speaker describes a target state.
Informative	Speaker states his knowledge, his opinion, and what he considers fact.
Other statement	Expression of gratitude, an apology, etc.
Other forward-looking function	Adjustment of a dialogue, etc.

Table 2: Standard dialogue act tag set (translated from Araki et al. 1999) and added tags 2

Tag name	Definition
Response (function to respond to initiation.)	
Affirmative/Acceptance	An answer which affirms propositional content of a Yes/No-question. An answer which shows that a demand has been accepted, in response to a request or invitation.
Negative/Refusal	An answer which negates propositional content of a Yes/No-question. An answer which shows that a demand has been refused, in response to a request or invitation.
WH-answer	Utterance which responds to WH-question.
Suspension	A hearer does not return a reply directly, or hearer's answer means he will return a response in the future, when ought to return some kind of response.
Backchannel	Reply which mainly adjusts a dialogue to a speaker's utterance.
Other reply	Utterance which hearer to return a response clearly when duty to return a certain response.
Follow-up (function to close an exchange.)	
Agreement	Utterance which means the purpose of an exchange was attained, following a reply.
Others	
%	Only a particle or a pronoun utterance, which does not function as a dialogue act.
Filler	The utterance which fills an utterance interval, such as "well", "mm", etc. This is not treated as a dialogue act.

We divided the transcriptions into utterance units, and tagged dialogue acts using the tag set. This work was done by 15 persons. Each dialogue was tagged by at least three persons, trained in advance. Table 3 shows three persons' coincidence rate of segmentation and labeling of dialogue acts in one dialogue corpus.

Table 3: Consistency of dividing utterance units and tagging dialogue acts

Dividing utterance units	Tagging dialogue acts
90.1%	74.9%

3.3 Local dialogue structure

A local structure of a dialogue is generally determined, even if the structure of a dialogue as a whole is ambiguous. Specifically, when information to be conveyed is grounded by dialogue participants, we generally regard a sequence of utterances corresponding to the grounding process as a local dialogue structure. A local dialogue structure generally consists of multiple utterances corresponding to several exchange units (Allen 1983). Figure 1 shows an example of a local dialogue structure. The portion enclosed in "——" is a local dialogue structure and "1→2:WH-question" immediately below "——" is the name of the local structure. Each utterance is expressed with an utterance number, Speaker ID and the contents of utterance, and a dialogue act is bitted for each line. "1→2:WH-question" means that the local dialogue structure is a WH-question from dialogue participant 1 to participant 2, and its succeeding utterance sequences consist of an answer, confirmation and affirmation/acceptance. This local dialogue structure continues until the goal of the first initiation utterance 3-7 is satisfied. There are multiple dialogue acts in one local dialogue structure. We name a local dialogue structure the name of the dialogue act which the speaker desires most among dialogue acts which constitute of the local dialogue structures. From this point of view, the local structure is defined as "WH-question from 1 to 2." In the route guidance dialogues, we refer to the guide as 1, and the person who received guidance as 2. In Figure 1, D is a guide, and K a person who receives guidance. Accordingly, we express this structure as "1→2:WH-question."

———

1→2:WH-question
3-7
D: soko-no moyori-no eki-tte iunowa? (Where is the closest station there?)
WH-question

4-1
K: eeto, (well...)
Filler

4-2
K: Mutuai. (Mutuai.)
WH-answer

5-1
D: Odakyu-sen-no Mutuai? (Mutuai is Odakyu Line's station, isn't it?)
Confirmation

5-2
K: hai. (Yes.)
Affirmation/Acceptance

———

Figure 1: Example of a local dialogue structure

3.4 Local dialogue structure rules

We define rewriting rules that correspond to patterns of local dialogue structures. The left-hand side of a rewriting rule is assigned a local dialogue structure label. In figure 3, "⟨1→2:Informative⟩" and "⟨1:Informative part⟩" is left-hand side. An element of the right-hand side is a dialogue act or a local dialogue structure label. A form of a local dialogue structure rule is generally a style of context-free grammars. Recursive rules are obtained to express a repeated pattern of dialogue act sequences. Figure 2 shows some of local dialogue rules.

⟨1→2:Suggestion⟩ ⇒ ⟨1:Suggestion part⟩ ⟨2:Affirmative/Acceptance part⟩
⟨1→2:Request⟩ ⇒ ⟨1:Request part⟩ ⟨2:Affirmative/Acceptance part⟩
⟨1→2:Yes/No-question⟩ ⇒ ⟨1:Yes/No-question part⟩ ⟨2:Affirmative/Acceptance part⟩
⟨1→2:Yes/No-question⟩ ⇒ ⟨1:Yes/No-question part⟩ ⟨2:Negative/Refusal part⟩
⟨1→2:Informative⟩ ⇒ ⟨1:Informative part⟩
⟨1→2:Informative⟩ ⇒ ⟨1:Informative part⟩ ⟨2→1:Confirmation⟩
⟨2→1:Confirmation⟩ ⇒ ⟨2:Confirmation part⟩ ⟨1:Affirmative/Acceptance part⟩
⟨2→1:Confirmation⟩ ⇒ ⟨2:Confirmation part⟩ ⟨1:Affirmative/Acceptance part⟩
⟨2:Agreement part⟩
⟨1:Informative part⟩ ⇒ 1:Informative
⟨1:Affirmative/Acceptance part⟩⇒ 1:Affirmative/Acceptance part

Figure 2: Example of a local dialogue structure rules

We create local dialogue structure rules from annotated dialogue transcriptions, extracting a local dialogue structure label like “1→2:WH-question”, and a sequence of dialogue acts which constitutes the local dialogue structure. This makes a rewriting rule whose left-hand side is the local dialogue structure label, and whose right-hand side consists of the sequence of dialogue acts.

3.5 Extending a set of rules

Extracted rules from the annotated dialogue transcriptions are generally insufficient for recognizing diverse dialogues not used for extracting rules, even if they are similar route guidance dialogues. Some local structures may not be captured by rules extracted from the annotated dialogue transcriptions. Thus, we add appropriate rules to the original rule set. We added new rules as follows:

- An utterance unit with a dialogue act may be generally divided by backchannels (aizuchi). In such a case, we regard the utterance as being composed of all the divided utterance units, and each divided utterance unit as having the same dialogue act as the original utterance unit. To handle such phenomena, we add a new rule which consists of a sequence of the same dialogue act and backchannel. In figure 3, rule of (2), (3), (4) and (5) are made from (1) by the above-mentioned method.

- Since a few Follow-up utterances follow a Response utterance, we add a new rule whose right-hand side includes Follow-up utterances. In figure 3, rule of (6) and (7) are made by this method.
- The source rule (1) is removed because the local dialogue structure accepted by rule (1) is included by the added rules.

Source:
⟨1→2:Informative⟩ ⇒ 1:Informative 1:Informative 2:Backchannel ··· (1)

Addition:
⟨1→2:Informative⟩ ⇒ ⟨1:Informative part⟩ ··· (2)
⟨1:Informative part⟩ ⇒ 1:Informative ··· (3)
⟨1:Informative part⟩ ⇒ 1:Informative 2:Backchannel ··· (4)
⟨1:Informative part⟩ ⇒ ⟨1:Informative part⟩ 1:Informative 2:Backchannel ··· (5)
⟨1→2:Informative⟩ ⇒ ⟨1:Informative part⟩ 2:Agreement ··· (6)
⟨1→2:Informative⟩ ⇒ ⟨1:Informative part⟩ 2:Agreement 1:Agreement ··· (7)

Figure 3: Examples of added local dialogue structure rules

3.6 Recognizing local dialogue structures and dialogue acts

We propose a method for recognizing local dialogue structures and dialogue acts using local dialogue structure rules. Our method is roughly described as follows: First, candidate dialogue acts corresponding to an utterance unit are inferred from the surface linguistic information of the utterance. Then, local dialogue structures are analyzed using local dialogue structure rules and accordingly dialogue acts are determined.

We also propose a hybrid method, in addition to the hierarchical aspects shown above. For the hybrid aspect, when a local dialogue structure cannot be obtained, we use the previous utterance dialogue act to infer candidate dialogue acts as described below in the experiment.

Figure 4 shows our method for recognizing local dialogue structures and dialogue acts. For simplicity, we explain the method using an example of two utterances. The first utterance is "*Odakyu-sen no Mutuai?* (Mutuai is a station of Odakyu Line, isn't it?)" by Speaker 1. The second utterance is "*Hai.* (Yes.)" by Speaker 2. For the first utterance, the decision tree infers either "Yes/No-question" or "Confirmation" as the dialogue act, and the probabilities of the candidates are 0.3 and 0.7. We presuppose that two local dialogue structures could be applicable here. One is "⟨1→2:Y/N-question⟩," whose first constituent is "Yes/No-question." Another local dialogue structure rule is "⟨1→2:Confirmation⟩," whose first constituent is "Confirmation." Therefore, we can use the two local dialogue structure rules "⟨1→2:Y/N-question⟩" and "⟨1→2:Confirmation⟩." The next constituent of both of the local dialogue structure rules are "Affirmation/Acceptance." Next, candidates of dialogue acts for the second utterance are "Backchannel" and "Affirmation/ Acceptance." Both of the local dialogue structure rules predict "Affirmation/ Acceptance." Accordingly, both of the rules are acceptable. On the other hand, applying "⟨1→2:Confirmation⟩" has higher probability than "⟨1→2:Y/N-question⟩." Therefore, the local dialogue structure is analyzed as "1→2:Confirmation," the dialogue act of the first utterance "Confirmation" and the second utterance "Affirmation/Acceptance."

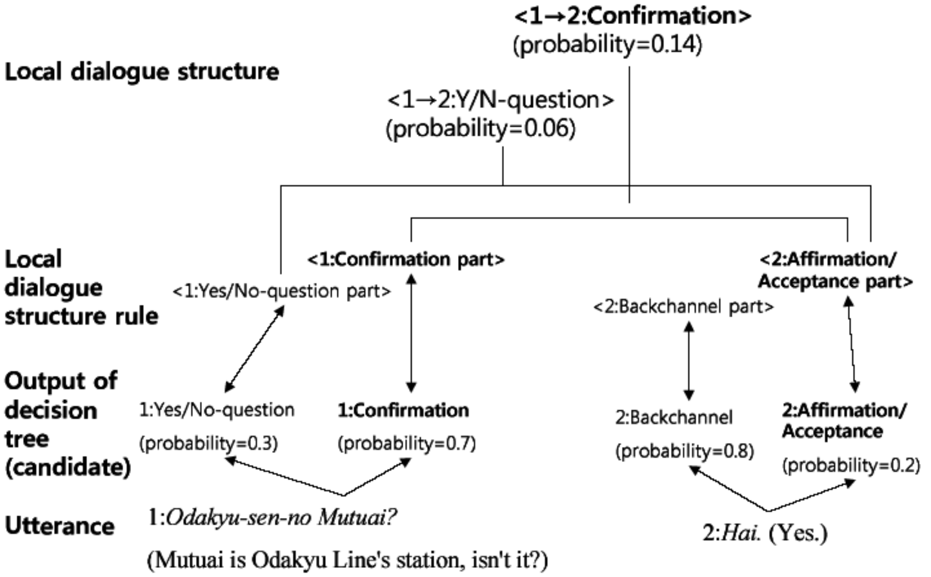


Figure 4: Recognition of dialogue act and local dialogue structures

4. Experiments and evaluation

We carried out the following experiments. First, we made a decision tree that seeks for candidate dialogue acts from the surface linguistic information of an utterance. The decision tree was made using the C4.5 algorithm. We used morphemes of the last phrase (Bunsetu) of an utterance. Nine dialogues (1,776 Utterance units) were used for making the decision tree.

We also extracted dialogue act sequences of local dialogue structures from the nine dialogues, and made local dialogue structure rules. The number of extracted local dialogue structure rules is 269. The number of local dialogue structure rules is 411.

We applied this method to the other three transcribed dialogues (564 utterances units) not used for training C4.5 and extracting local dialogue structure rules. We assumed that our experiment input is each utterance unit and each local dialogue structure of the three transcribed dialogues.

Table 4 shows an experiment result. In the route guidance dialogue, the most frequent utterances by a guide are informative. Backchannel utterances are the most frequent utterances by a person who receives guidance information. Accordingly, we use informative and backchannel as "Baseline". "Linguistic info" shows the case when using the highest probability dialogue act output by the C4.5 decision tree, based solely on surface linguistic information mainly part of speech. "+Previous DA" shows the case when using the highest probability dialogue act output by the C4.5 decision tree using dialogue act information just before an utterance unit, in addition to surface linguistic information. "Structure" shows the case when applying local dialogue structure rules and using the C4.5 decision tree without previous dialogue act information. "Combined" shows the combination of "Structure" and "+Previous DA," where "Structure" is used when local dialogue structure information can be used, and "+Previous DA" is used when local dialogue structure cannot be used. For "Structure", the system did not output dialogue acts in all utterance units, and the recall are shown in parentheses.

The precision of "Structure" (Table 4) is higher than the others and that of "Combined" is lower than "Structure" though both uses the local dialogue structure rules. Precision of "Structure" is counted only when a local dialogue structure is recognized. We can say that the precision of the dialogue act recognition is higher than the others when a local dialogue structure governing dialogue acts is recognized.

Table 5 shows structure recognition's recall, precision and F-measure in this experiment;

$$\text{Recognized structure} = \frac{\text{the number of dialogue structures system recognized.}}{\text{the number of relevant dialogue structures.}}$$

$$\text{Recognized relevant structure} = \frac{\text{the number of relevant dialogue structures system recognized.}}{\text{the number of relevant dialogue structures.}}$$

$$\text{Recall} = \frac{\text{the number of relevant dialogue structures recognized which is a best probability.}}{\text{the number of relevant dialogue structures.}}$$

$$\text{Precision} = \frac{\text{the number of relevant dialogue structures recognized which is a best probability.}}{\text{the number of system recognized structures.}}$$

$$\text{F - measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

"Recognized Structure" means only that system can recognize local dialogue structure without right or wrong. "Recognized Relevant Structure" means that system can recognize relevant local dialogue structure without ranking. “Recall” and “Precision” is considered ranking.

All the methods except “Structure” output candidate dialogue acts for all the utterance units (Table 4). On the other hand, for local dialogue structure recognition, the number of local dialogue structure rules were insufficient because of the shortage of analyzed corpus, and the method “Structure” did not cover all the local dialogue structures of the dialogues we experimented. The recognition failures were caused by misrecognizing about 30% of dialogue acts which constitute the local dialogue structure.

Our results is lower than in (Koh & Shirai 2005). We consider a few reasons for this: The corpus we used is the same domain as (Koh & Shirai 2005), however different kinds of dialogues. (Koh & Shirai 2005) used cue words more than we did. The reason that the accuracy of recognizing local dialogue structures is not so high, is that the size of the corpus is insufficient, and accordingly the number of rules is insufficient.

Table 4: Recall of dialogue acts recognition

Dialogue number	Baseline	Linguistic info	+Previous DA	Structure (Precision)	Combined
A04	42.70%	49.44%	60.67%	46.07% (63.08%)	60.67%
A06	38.10%	51.58%	62.96%	55.56% (88.24%)	66.14%
A21	34.52%	37.06%	44.67%	51.27% (59.06%)	57.87%
Average	38.44%	46.03%	56.10%	50.97% (70.13%)	61.56%

Table 5: Local dialogue structures recognition

Dialogue number	Recognized structure	Recognized relevant structure	Recall	Precision	F-measure
A04	81.82%	36.36%	15.15%	18.52%	0.1667
A06	91.67%	58.33%	27.78%	30.30%	0.2899
A21	89.19%	43.42%	27.03%	30.30%	0.2857
Average	87.56%	45.98%	23.58%	26.88%	0.2512

5. Conclusion

We proposed a method for recognizing local dialogue structures and dialogue acts based on local dialogue structures. Due to the limited size of the corpus, recognition accuracy of dialogue acts using only the structures is lower than when using the previous dialogue act. The method combining the structure and the previous dialogue act shows a better result than the method using previous dialogue act.

We will increase data and experiment to see the effect of this method, and apply the method not only to route guidance dialogues however to dialogues of different domains as well.

Kenji Takano
School of Information Science
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi,
Ishikawa 923-1292 Japan
k-takano@jaist.ac.jp

Akira Shimazu
School of Information Science
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi,
Ishikawa 923-1292 Japan
shimazu@jaist.ac.jp

References

Abercrombie, David. 1965. *Studies in Phonetics and Linguistics.* London: Oxford University Press.

Adda-Decker, Martine, Benoit Habert, Claude Barras, Giles Adda, Philippe Boula de Mareüil, and Patrick Paroubek. 2003. A disfluency study for cleaning spontaneous speech automatic transcripts and improving speech language models. *Proceedings of ISCA Tutorial and Research Workshop on Disfluency in Spontaneous Speech* (*DiSS'03*), 67-70. Göteberg, Sweden.

Allen, James F. 1983. Recognizing intentions from natural language utterances. *Computational Models of Discourse*, ed. by Michael Brady & Robert C. Berwick, 107-166. Cambridge: MIT Press.

Anderson, Anne H., Miles Bader, Ellen G. Bard, and Elizabeth Boyle. 1991. The HCRC Map Task Corpus. *Language and Speech* 34.4:351-366.

Ang, Jeremy, Yang Liu, and Elizabeth Shriberg. 2005. Automatic dialog act segmentation and classification in multiparty meetings. *Proceedings of ICASSP-2005*, 1061-1064. Philadelphia, Pennsylvania, USA.

Anon. 2005. E-MELD 2005 Ontology FAQ. http://emeld.org/workshop/2005/ontology.html.

Anon. 2006. OLAC: Open Language Archives Community. http://www.language-archives.org/2006 04 12. (last consulted 2006-10-23)

Araki, Masahiro, Toshihiko Ito, Tomoko Kumagai, and Masato Ishizaki. 1999. Proposal of a standard utterance-unit tagging scheme. *Journal of the Japanese Society for Artificial Intelligence* 14.2:251-259. (In Japanese)

Aristar, Anthony. 2005. Phonetics Ontology. http://emeld.org/workshop/2005/ontology-html/phonetics. (last consulted 2006-10-22)

Atkinson, John Maxwell, and John Heritage. (eds.) 1984. *Structures of Social Action: Studies in Conversation Analysis.* Cambridge: Cambridge University Press.

Atterer, Michaela, and Dwight Robert Ladd. 2004. On the phonetics and phonology of 'segmental anchoring' of F0: evidence from German. *Journal of Phonetics* 32.2: 177-197.

Auer, Peter. 2001. 'Hoch ansetzende' Intonationskonturen in der Hamburger Regionalvarietät ['High-starting' intonation contours in the Hamburg regional variety]. *Neue Wege der Intonationsforschung*, ed. by Jürgen Erich Schmidt, 126-165. Hildesheim: Olms.

Aylett, Matthew P. 2003. Disfluency and speech recognition profile factors. *Proceedings of ISCA Tutorial and Research Workshop on Disfluency in Spontaneous Speech* (*DiSS'03*), 51-54. Göteberg, Sweden.

Barnlund, Dean C. 1989. *Communicative Styles of Japanese and Americans: Images and Realities.* Belmont: Wadsworth.

Baron, Don, Elizabeth Shriberg, and Andreas Stolcke. 2002. Automatic punctuation and disfluency detection in multi-party meetings using prosodic and lexical cues. *Proceedings of ICSLP-2002*, 949-952. Denver, Colorado, USA.

Bates, Rebecca A., and Mari Ostendorf. 2002. Modeling pronunciation variation in conversational speech using prosody. *Proceedings of ISCA Tutorial and Research Workshop on Pronunciation Modeling and Lexical Access* (*PMLA-2002*), 42-47. Estes Park, Colorado, USA.

Bates, Rebecca A., Mari Ostendorf, and Richard A. Wright. 2007. Symbolic phonetic features for modeling of pronunciation variation. *Speech Communication* 49.2: 83-97.

Batista, Fernando, Diamantino Caseiro, Nuno Mamede, and Isabel Trancoso. 2007. Recovering punctuation marks for automatic speech recognition. *Proceedings of INTERSPEECH-2007*, 2153-2156. Antwerp, Belgium.

Beattie, Geoffrey, Anne Cutler, and Mark Pearson. 1982. Why is Mrs. Thatcher interrupted so often? *Nature* 300.23:744-747.

Beckman, Mary E. 1989. Timing models for prosody and cross-word coarticulation in connected speech. *Proceedings of the Second DARPA Speech Recognition Workshop*, 12-21. San Francisco: Morgan Kaufmann.

Beckman, Mary E., and Jan Edwards. 1990. Lengthenings and shortenings and the nature of prosodic constituency. *Between the Grammar and Physics of Speech*, ed. by John Kingston & Mary E. Beckman, 152-178. Papers in Laboratory Phonology 1. Cambridge & New York: Cambridge University Press.

Beckman, Mary E., and Gayle Ayers Elam. 1994. Guidelines for ToBI labeling. Manuscript. Columbus: Ohio State University.

Beckman, Mary E., and Julia Hirschberg. 1994. The ToBI annotation conventions. Manuscript. Columbus: Ohio State University; New York: Columbia University.

Benzeghiba, Mohamed, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens. 2007. Automatic speech recognition and speech variability: a review. *Speech Communication* 49.10-11:763-786.

Berger, Adam L., Stephen A. Della Pietra, and Vincent J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics* 22.1: 39-72.

Biber, Douglas. 1986. Spoken and written textual dimensions in English: resolving the contradictory findings. *Language* 62.2:384-414.

Biber, Douglas. 1988. *Variation Across Speech and Writing.* Cambridge & New York: Cambridge University Press.

Biber, Douglas, and Edward Finegan. 1994. Register and social dialect variation: an integrated approach. *Sociolinguistic Perspectives on Register*, ed. by Douglas Biber & Edward Finegan, 315-347. New York & Oxford: Oxford University Press.

Biber, Douglas, and Edward Finegan. 2001. Register variation and social dialect variation: the register axiom. *Style and Sociolinguistic Variation*, ed. by Penelope Eckert & John R. Rickford, 235-267. Cambridge & New York: Cambridge University Press.

Bickley, Corine. 1982. Acoustic analysis and perception of breathy vowels. *Speech Communication Group Working Papers* 1:71-81. Cambridge: Research Laboratory of Electronics, MIT.

Bierwisch, Manfred. 1966. Regeln für die Intonation deutscher Sätze. Untersuchungen über Akzent und Intonation im Deutschen [Rules for the intonation of German sentences. Investigations on accent and intonation in German]. *Studia Grammatica* 7:99-199.

Binnenpoorte, Diana, Catia Cucchiarini, Helmer Strik, and Lou Boves. 2004. Improving automatic phonetic transcription of spontaneous speech through variant-based pronunciation variation modeling. *Proceedings LREC-2004*, 2981-2984. Lisbon, Portugal.

Bitar, Nabil, and Carol Espy-Wilson. 1996. A knowledge-based signal representation for speech recognition. *Proceedings of ICASSP-1996*, 29-32. Atlanta, Georgia, USA.

Boersma, Paul. 1993. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampling sound. *Proceedings of the Institute of Phonetic Sciences 17*, 97-110. Amsterdam: University of Amsterdam.

Boersma, Paul, and David Weenink. 2005[2006]. Praat: doing phonetics by computer. http://www.fon.hum.uva.nl/praat 6.7.2006.

Bolinger, Dwight. 1986. *Intonation and Its Parts: Melody in Spoken English.* Stanford: Stanford University Press.

Bolinger, Dwight. 1989. *Intonation and Its Uses: Melody in Grammar and Discourse.* Stanford: Stanford University Press.

Borys, Sarah. 2003. The importance of prosodic factors in phoneme modeling with applications to speech recognition. *Proceedings of HLT-NAACL 2003 Student Research Workshop*, Vol. 3, 7-12. Edmonton, Canada.

Bremer, Otto. 1893. *Deutsche Phonetik* [*German Phonetics*]. Leipzig: Breitkopf & Härtel.

Browman, Catherine P., and Louis Goldstein. 1992. Articulatory phonology: an overview. *Phonetica* 49.3-4:155-180.

Brown, Gillian, Karen L. Currie, and Joanne Kenworthy. 1980. *Questions of Intonation.* Baltimore: University Park Press.

Brown, Penelope, and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage.* Cambridge & New York: Cambridge University Press.

Bruce, Gosta. 1995. Modelling Swedish intonation for read and spontaneous speech. *Proceedings of ICPhS-1995*, Vol. 2, 28-35. Stockholm, Sweden.

Bühler, Karl. 1934. *Sprachtheorie: Die Darstellungsfunktion der Sprache*. Jena: Gustav Fischer.

Byrd, Dani, and Elliott Saltzman. 2003. The elastic phrase: modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics* 31.2:149-180.

Campbell, Nick. 1992. Segmental elasticity and timing in Japanese speech. *Speech Perception, Production and Linguistic Structure*, ed. by Yoh'ichi Tohkura, Erik Vatikiotis-Bateson & Yoshinori Sagisaka, 403-418. Tokyo: Ohmsha.

Campione, Estelle, and Jean Véronis. 2002. A large-scale multilingual study of silent pause duration. *Proceedings of Speech Prosody 2002*, 199-202. Aix-en-Provence, France.

Cao, Jianfen, and Ian Maddieson. 1989. An exploration of phonation types in Wu dialects of Chinese. *UCLA Working Papers in Phonetics* 72:139-160.

Carnap, Rudolf. 1958. *Introduction to Symbolic Logic and Its Applications*. New York: Dover Publications.

Carson-Berndsen, Julie. 1998. *Time Map Phonology: Finite State Models and Event Logics in Speech Recognition.* Dordrecht & Boston: Kluwer Academic Publishers.

Chafe, Wallace L. 1982. Integration and involvement in speaking, writing, and oral literature. *Spoken and Written Language: Exploring Orality and Literacy*, ed. by Deborah Tannen, 35-54. Norwood: Ablex.

Chafe, Wallace L. 1985. Linguistic differences produced by differences between speaking and writing. *Literacy, Language and Learning: The Nature and Consequences of Reading and Writing*, ed. by David Olson, Nancy Torrance & Angela Hildyard, 105-123. Cambridge & New York: Cambridge University Press.

Chafe, Wallace L. 1994. *Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing.* Chicago: University of Chicago Press.

Chafe, Wallace L., and Jane Danielewicz. 1987. Properties of spoken and written language. *Comprehending Oral and Written Language*, ed. by Rosalind Horowitz & S. Jay Samuels, 83-113. New York: Academic Press.

Chang, Chih-Chung, and Chih-Jen Lin. 2004. LibSVM: A Library for Support Vector Machine. System documentation. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chang, Shuangyu, Steven Greenberg, and Mirjam Wester. 2001. An elitist approach to articulatory-acoustic feature classification. *Proceedings of EUROSPEECH-2001*, 1725-1728. Aalborg, Denmark.

Chao, Yuen Ren. 1968. *A Grammar of Spoken Chinese.* Berkeley: University of California Press.

Chen, Ken, and Mark Hasegawa-Johnson. 2003. Improving the robustness of prosody dependent language modeling based on prosody syntax cross-correlation. *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding* (*ASRU'03*), 435-440. St. Thomas, Virgin Islands, USA.

Chen, Ken, Mark Hasegawa-Johnson, and Jennifer Cole. 2003. Prosody dependent speech recognition with explicit duration modeling at intonational phrase boundaries. *Proceedings of EUROSPEECH-2003*, 393-396. Geneva, Switzerland.

Chen, Ken, Mark Hasegawa-Johnson, Aaron Cohen, Sarah Borys, Sung-Suk Kim, Jennifer Cole, and Jeung-Yoon Choi. 2006. Prosody dependent speech recognition on radio news. *IEEE Transactions on Speech and Audio Processing* 14.1:232-245.

Chen, Marilyn. 2000. Nasal landmark detection. *Proceedings of ICSLP-2000*, 636-639. Beijing, China.

Chen, Zi-He, Zhi-Ren Zeng, Yuan-Fu Liao, and Yau-Tarng Juang. 2006. Probabilistic latent prosody analysis for robust speaker verification. *Proceedings of ICASSP-2006*, 105-108. Toulouse, France.

Cheng, Robert L. 1985. Sub-syllabic morphemes in Taiwanese. *Journal of Chinese Linguistics* 13.1:12-43.

Cheng, Shou-Yi, Dian-Song Wu, and Tyne Liang. 2006. A Corpus-based developed rhetorical parser on Chinese texts. *Proceedings of ROCLING XVIII*, 15-28. Hsinchu, Taiwan. (In Chinese)

Cheng, Xianghui, and Xiaolin Tian. 1992. *Modern Chinese.* Taipei: Bookman Press. (In Chinese)

Chiu, Bonnie. 1995. An object clitic projection in Mandarin Chinese. *Journal of East Asian Linguistics* 4.2:77-117.

Chomsky, Noam, and Morris Halle. 1968. *The Sound Pattern of English.* New York: Harper and Row.

Chu, Chauncey C. 1987. *Historical Syntax—Theory and Application to Chinese.* Taipei: Crane.

Chu, Chauncey C. 2002. Relevance theory, discourse markers and the Mandarin utterance-final particles A/YA. *Journal of the Chinese Teachers Association* 37.1: 1-42.

Chu, Stephen, and Thomas S. Huang. 2000. Bimodal speech recognition using coupled Hidden Markov Models. *Proceedings of ICSLP-2000*, 747-750. Beijing, China.

Chung, Raung-fu. 1997. Syllable contraction in Chinese. *Chinese Languages and Linguistics*, Vol. 3: *Morphology and Lexicon*, ed. by Feng-fu Tsao & H. Samuel Wang, 199-235. Taipei: Institute of History and Philology, Academia Sinica.

CKIP. 1995. Sinica Balanced Corpus. Taipei: Academia Sinica. (In Chinese)

Clark, Herbert H., and Thomas Wasow. 1998. Repeating words in spontaneous speech. *Cognitive Psychology* 37.3:201-242.

Clark, Herbert H., and Jean E. Fox Tree. 2002. Using *uh* and *um* in spontaneous speaking. *Cognition* 84.1:73-111.

Cole, Jennifer, Mark Hasegawa-Johnson, Chilin Shih, Heejin Kim, Eun-Kyung Lee, Hsin-yi Lu, Yoonsook Mo, and Tae-Jin Yoon. 2005. Prosodic parallelism as a cue to repetition and error correction repair disfluency. *Proceedings of ISCA Tutorial and Research Workshop on Disfluency in Spontaneous Speech* (*DiSS'05*), 53-58. Aix-en-Provence, France.

Cole, Jennifer, Heejin Kim, Hansook Choi, and Mark Hasegawa-Johnson. 2007. Prosodic effects on acoustic cues to stop voicing and place of articulation: evidence from radio news speech. *Journal of Phonetics* 35.2:180-209.

Collins, Peter C. 1991. *Cleft and Pseudo-cleft Constructions in English.* London & New York: Routledge.

Cook, Mark, and Mansur Lalljee. 1973. Uncertainty in first encounters. *Journal of Personality and Social Psychology* 26.1:137-141.

Corley, Martin, and Robert Hartsuiker. 2003. Hesitation in speech can... um... help a listener understand. *Proceedings of the 25th Annual Meeting of the Cognitive Science Society*, ed. by Richard Alterman & David Kirsh, 276-281. Boston: Cognitive Science Society.

Coulthard, Malcolm, and David Brazil. 1981. Exchange structure. *Studies in Discourse Analysis*, ed. by Malcolm Coulthard & Martin Montgomery, 82-106. London & Boston: Routledge & Kegan Paul.

Couper-Kuhlen, Elizabeth. 1986. *An Introduction to English Prosody.* Tübingen: Niemeyer.

Couper-Kuhlen, Elizabeth, and Cecilia E. Ford. 2004. *Sound Patterns in Interaction: Cross-linguistic Studies from Conversation.* Amsterdam & Philadelphia: John Benjamins.

Couper-Kuhlen, Elizabeth, and Margret Selting. 1996. Towards an interactional perspective on prosody and a prosodic perspective on interaction. *Prosody in Conversation: Interactional Studies*, ed. by Elizabeth Couper-Kuhlen & Margret Selting, 11-56. Cambridge & New York: Cambridge University Press.

Couper-Kuhlen, Elizabeth, and Margret Selting. (eds.) 1996. *Prosody in Conversation: Interactional Studies.* Cambridge & New York: Cambridge University Press.

Croft, William. 1995. Intonation units and grammatical structure. *Linguistics* 33.5:839-882.

Cruttenden, Alan. 1986[1997]. *Intonation.* Cambridge & New York: Cambridge University Press.

Crystal, David. 1969. *Prosodic Systems and Intonation in English.* London: Cambridge University Press.

Cuendet, Sébastien, Dilek Hakkani-Tür, and Gokhan Tur. 2006. Model adaptation for sentence unit segmentation from speech. *Proceedings of SLT-2006.* Aruba.

Cuendet, Sébastien, Dilek Hakkani-Tür, Elizabeth Shriberg, James Fung, and Benoit Favre. 2007a. Cross-genre feature comparisons for spoken sentence segmentation. *Proceedings of ICSC-2007*, 265-271. Irvine, California, USA.

Cuendet, Sébastien, Elizabeth Shriberg, Benoit Favre, James Fung, and Dilek Hakkani-Tür. 2007b. An analysis of sentence segmentation features for broadcast news, broadcast conversations, and meetings. *Proceedings of SIGIR Workshop on Searching Conversational Spontaneous Speech*, 37-43. Amsterdam, Netherlands.

Cutler, Anne, and Dwight Robert Ladd. (eds.) 1983. *Prosody: Models and Measurements.* Berlin & New York: Springer-Verlag.

Dankovičová, Jana. 1997. The domain of articulation rate variation in Czech. *Journal of Phonetics* 25.3:287-312.

Dauenhauer, Bernard P. 1980. *Silence: The Phenomenon and Its Ontological Significance.* Bloomington: Indiana University Press.

Delattre, Pierre C., Alvin M. Liberman, and Franklin S. Cooper. 1955. Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America* 27.4:769-773.

Den, Yasuharu. 2001. Are word repetitions really intended by the speaker? *Proceedings of ISCA Tutorial and Research Workshop on Disfluency in Spontaneous Speech* (*DiSS'01*), 25-28. Edinburgh, Scotland, UK.

Den, Yasuharu. 2003. Some strategies in prolonging speech segments in spontaneous Japanese. *Proceedings of ISCA Tutorial and Research Workshop on Disfluency in Spontaneous Speech* (*DiSS'03*), 87-90. Göteborg, Sweden.

Den, Yasuharu. 2007. Hatsuwa-bôtô-fukin-de-no goku-no kurikaeshi-no kinô [The function of repeated words around an utterance-initial position]. *Jikan-no Naka-no Bun-to Hatsuwa* [*Sentences and Utterances in Time*], ed. by Shuya Kushida, Toshiyuki Sadanobu & Yasuharu Den, 103-133. Tokyo: Hitsuji Shobo. (In Japanese)

Den, Yasuharu, and Herbert H. Clark. 2000. Word repetitions in Japanese spontaneous speech. *Proceedings of ICSLP-2000*, 58-61. Beijing, China.

Dendrinos, Bessie, and Emilia Ribeiro-Pedro. 1997. Giving street directions: the silent role of women. *Silence: Interdisciplinary Perspective*, ed. by Adam Jaworski, 215-238. Berlin & New York: Mouton de Gruyter.

Dhillon, Rajdip, Sonali Bhagat, Hannah Carvey, and Elizabeth Shriberg. 2004. Meeting Recorder Project: Dialog Act Labeling Guide. Berkeley: International Computer Science Institute.

Dielmann, Alfred, and Steve Renals. 2006. Multistream recognition of dialogue acts in meetings. *Lecture Notes in Computer Science* 4299:178-189.

Dilley, Laura, Stefanie Shattuck-Hufnagel, and Mari Ostendorf. 1996. Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics* 24.4: 423-444.

Doddington, George, Andres Corrada, Aravind Ganapathiraju, Vaibhava Goel, Barbara Wheatley, Katrin Kirchhoff, Mark Ordowski, and Joe Picone. 1997. Syllable-Based Speech Processing. Johns Hopkins Center for Language and Speech Processing.

Dousaka, Kohji, and Akira Shimazu. 1996. A computational model of incremental utterance production in task-oriented dialogues. *Proceedings of COLING-1996*, 304-309. Copenhagen, Denmark.

Du Bois, John W., Stephen Schuetze-Coburn, Susanna Cumming, and Danae Paolino. 1993. Outline of discourse transcription. *Talking Data: Transcription and Coding in Discourse Research*, ed. by Jane A. Edwards & Martin D. Lampert, 45-89. Hillsdale: Lawrence Erlbaum Associates.

Eklund, Robert. 2001. Prolongations: a dark horse in the disfluency stable. *Proceedings of ISCA Tutorial and Research Workshop on Disfluency in Spontaneous Speech* (*DiSS'01*), 5-8. Edinburgh, Scotland, UK.

Epstein, Melissa A. 2002. *Voice Quality and Prosody in English*. Los Angeles: UCLA dissertation.

Eskenazi, Maxine. 1993. Trends in speaking style research. *Proceedings of EURO-SPEECH-1993*, 501-509. Berlin, Germany.

Esling, John H. 1978. The identification of features of voice quality in social groups. *Journal of the International Phonetic Association* 7:18-23.

ESPS. 1993. ESPS Version 5.0 Programs Manual. Washington, D.C., USA.

Espy-Wilson, Carol. 1994. A feature-based semi-vowel recognition system. *Journal of the Acoustical Society of America* 96.1:65-72.

Fagyal, Susan. 1995. Subject: 6.607 Sum: Spontaneous Speech. LINGUIST LIST vol-6-607 Tue 25 Apr 1995 ISSN: 1068-4875.

Fant, Gunnar. 1960. *Acoustic Theory of Speech Production.* The Hague: Mouton.

Fant, Gunnar. 1997. The voice source in connected speech. *Speech Communication* 22.2-3:125-139.

Fant, Gunnar, and Anita Kruckenberg. 1989. Preliminaries to the study of Swedish prose reading and reading style. *STL-QPSR* 30.2:1-83.

Farrar, Scott, and Terence Langendoen. 2003. A linguistic ontology for the semantic web. *GLOT International* 7.3:97-100.

Ferrer, Luciana, Elizabeth Shriberg, and Andreas Stolcke. 2002. Is the speaker done yet? Faster and more accurate end-of-utterance detection using prosody in human-computer dialog. *Proceedings of ICSLP-2002*, 2061-2064. Denver, Colorado, USA.

Féry, Caroline. 1993. *German Intonational Patterns.* Tübingen: Niemeyer.

Fischer, Kerstin. 2000. *From Cognitive Semantics to Lexical Pragmatics: The Functional Polysemy of Discourse Particles.* Berlin & New York: Mouton de Gruyter.

Fischer-Jørgensen, Eli. 1967. Phonetic analysis of breathy (murmured) vowels. *Indian Linguistics* 28:71-139.

Fiscus, Jonathan. 1997. A post-processing system to yield reduced word error rates: recognizer output voting error reduction (ROVER). *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding* (*ASRU'97*), 347-354. Santa Barbara, California, USA.

Fiscus, Jonathan, William M. Fisher, Mark Przybocki, and David S. Pallett. 2000. The 2000 NIST evaluation of conversational speech recognition over the telephone. *Proceedings of Speech Transcription Workshop 2000*, P5:1-9. College Park: University of Maryland.

Floridi, Luciano. 2003. *The Blackwell Guide to the Philosophy of Computing and Information*. Oxford: Blackwell.

Fougeron, Cecile, and Patricia A. Keating. 1997. Articulatory strengthening at edges of prosodic domains. *Journal of the Acoustical Society of America* 101.6:3728-3740.

Fowler, Carol A., and Jonathan Housum. 1987. Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language* 26.5:489-504.

Fox Tree, Jean E. 2000. Coordinating spontaneous talk. *Aspects of Language Production*, ed. by Linda Wheeldon, 375-406. Hove & Philadelphia: Psychology Press.

Fox Tree, Jean E. 2002. Interpreting pauses and ums at turn exchanges. *Discourse Processing* 34.1:37-55.

Fügen, Christian, and Muntsin Kolss. 2007. The influence of utterance chunking on machine translation performance. *Proceedings of INTERSPEECH-2007*, 2837-2840. Antwerp, Belgium.

Fung, James G., Dilek Hakkani-Tür, Mathew Magimai-Doss, Elizabeth Shriberg, Sébastien Cuendet, and Nikki Mirghafori. 2007. Cross-linguistic analysis of prosodic features for sentence segmentation. *Proceedings of INTERSPEECH-2007*, 2585-2588. Antwerp, Belgium.

Furui, Sadaoki, Tomonori Kikuchi, Yousuke Shinnaka, and Chiori Hori. 2004. Speech-to-text and speech-to-speech summarization of spontaneous speech. *IEEE Transactions on Speech and Audio Processing* 12.4:401-408.

Ganapathiraju, Aravind, Jonathan Hamaker, Joseph Picone, Mark Ordowski, and George R. Doddington. 2001. Syllable-based large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing* 9.4:358-366.

Gerratt, Bruce, and Jody Kreiman. 2001. Towards a taxonomy of nonmodal phonation. *Journal of Phonetics* 29.4:365-381.

Gibbon, Dafydd. 1976. *Perspectives of Intonation Analysis*. Bern: Peter Lang.

Gibbon, Dafydd. 1992. Prosody, time types and linguistic design factors in spoken language system architectures. *KONVENS '92*, ed. by Günther Görz, 90-99. Berlin: Springer.

Gibbon, Dafydd. 1998. German intonation. *Intonation Systems: A Survey of Twenty Languages*, ed. by Daniel Hirst & Albert di Cristo, 78-95. Cambridge & New York: Cambridge University Press.

Gibbon, Dafydd. 2002. Prosodic information in an integrated lexicon. *Proceedings of Speech Prosody 2002*, 335-338. Aix-en-Provence, France.

Gibbon, Dafydd. 2005. Prerequisites for a multimodal semantics of gestures and prosody. *Proceedings of IWCS-6.* Tilburg, Netherlands.

Gibbon, Dafydd. 2006. Time types and time trees: prosodic mining and alignment of temporally annotated data. *Methods in Empirical Prosody Research*, ed. by Stefan Sudhoff et al., 281-309. Berlin & New York: Walter de Gruyter.

Gibbon, Dafydd, Roger Moore, and Richard Winski. (eds.) 1997. *Handbook of Standards and Resources for Spoken Language Systems.* Berlin & New York: Mouton de Gruyter.

Gibbon, Dafydd, Inge Mertins, and Roger Moore. (eds.) 2000. *Handbook of Multimodal and Spoken Dialogue Systems: Resources, Terminology and Product Evaluation.* Boston: Kluwer Academic Publishers.

Gibbon, Dafydd, Ulrike Gut, Benjamin Hell, Karin Looks, Alexandra Thies, and Thorsten Trippel. 2003. A computational model of arm gestures in conversation. *Proceedings of EUROSPEECH-2003*, 813-816. Geneva, Switzerland.

Gilles, Peter. 2005. *Regionale Prosodie im Deutschen* [*Regional Prosody in German*]. Berlin: Walter de Gruyter.

Gobl, Christer. 2003. *The Voice Source in Speech Communication: Production and Perception Experiments Involving Inverse Filtering and Synthesis*. Stockholm: KTH dissertation.

Godfrey, John J., Edward C. Holliman, and Jane McDaniel. 1992. SWITCHBOARD: telephone speech corpus for research and development. *Proceedings of ICASSP-1992*, 517-520. San Francisco, California, USA.

Goldman-Eisler, Frieda. 1968. *Psycholinguistics: Experiments in Spontaneous Speech.* London & New York: Academic Press.

Gomi, Hiroaki, and Mitsuo Kawato. 1996. Equilibrium-point Control Hypothesis examined by measured arm stiffness during multijoint movement. *Science* 272.5258: 117-120.

Goodwin, Charles. 1981. *Conversational Organization: Interaction between Speakers and Hearers.* New York: Academic Press.

Gordon, Matthew. 1996. The phonetic structures of Hupa. *UCLA Working Papers in Phonetics* 93:164-187.

Gordon, Matthew, and Peter Ladefoged. 2001. Phonation types: a cross-linguistic overview. *Journal of Phonetics* 29.4:383-406.

Gorman, Kyle, Jennifer Cole, Mark Hasegawa-Johnson, and Margaret Fleck. 2007. Automatic detection of turn-taking cues in spontaneous speech using prosodic features. Paper presented at the 2007 Annual Meeting of the Linguistic Society of America. Anaheim, California, USA.

Greenberg, Steven. 1999. Speaking in shorthand—A syllable-centric perspective for understanding pronunciation variation. *Speech Communication* 29.2-4:159-176.

Greenberg, Steven, Joy Hollenback, and Dan Ellis. 1996. Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus. *Proceedings of ICSLP-1996*, S24-27. Philadelphia, Pennsylvania, USA.

Grice, Herbert Paul. 1975. Logic and conversation. *Syntax and Semantics*, Vol. 3: *Speech Acts*, ed. by Peter Cole & Jerry L. Morgan, 41-58. New York: Academic Press.

Grice, Martine, and Stefan Baumann. 2002. Deutsche Intonation und GToBI [German intonation and GToBI]. *Linguistische Berichte* 191:267-298.

Grosz, Barbara J., and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics* 12.3:175-204.

Gupta, Narendra K., Srinivas Bangalore, and Mazin Rahim. 2002. Extracting clauses for spoken language understanding in conversational systems. *Proceedings of ICSLP-2002*, 361-364. Denver, Colorado, USA.

Guz, Umit, Sébastien Cuendet, Dilek Hakkani-Tür, and Gokhan Tur. 2007. Co-training using prosodic and lexical information for sentence segmentation. *Proceedings of INTERSPEECH-2007*, 2597-2600. Antwerp, Belgium.

Hain, Thomas, Phil Woodland, Gunnar Evermann, and Dan Povey. 2000. CU-HTK March 2000 Hub5e Transcription System. *NIST 2000 Workshop on Speech Transcription*. P10:11-14. College Park: University of Maryland.

Hakkani-Tür, Dilek, and Gokhan Tur. 2007. Statistical sentence extraction for information distillation. *Proceedings of ICASSP-2007*, Vol. 4, 1-4. Honolulu, Hawaii, USA.

Halliday, Michael A. K. 1967. *Intonation and Grammar in British English.* The Hague: Mouton.

Hanson, Helen M. 1997. Glottal characteristics of female speakers: acoustic correlates. *Journal of the Acoustical Society of America* 101.1:466-481.

Hanson, Helen M., and Erika S. Chuang. 1999. Glottal characteristics of male speakers: acoustic correlates and comparison with female data. *Journal of the Acoustical Society of America* 106.2:1697-1714.

Hanson, Helen M., Kenneth N. Stevens, Hong-Kwang J. Kuo, Marilyn Y. Chen, and Janet Slifka. 2001. Towards models of phonation. *Journal of Phonetics* 29.4:451-480.

Hasegawa-Johnson, Mark. 1996. *Formant and Burst Spectral Measurements with Quantitative Error Models for Speech Sound Classification*. Cambridge: MIT dissertation.

Hasegawa-Johnson, Mark. 2005. Speech Tools Minicourse. http://www.isle.uiuc.edu/courses/minicourse/index.html.

Hasegawa-Johnson, Mark. 2006. HTK Study Group: Recognizer Training and Testing Methods.

Hasegawa-Johnson, Mark, James Baker, Steven Greenberg, Katrin Kirchhoff, Jennifer Muller, Kemal Sönmez, Sarah Borys, Ken Chen, Amit Juneja, Karen Livescu, Srividya Mohan, Emily Coogan, and Tianyu Wang. 2005. Landmark-Based Speech Recognition: Report of the 2004 Johns Hopkins Summer Workshop. Johns Hopkins University Center for Language and Speech Processing.

Häsler, Katrin, Ingrid Hove, and Beat Siebenhaar. 2005. Die Prosodie des Schweizerdeutschen—Erkenntnisse aus der sprachsynthetischen Modellierung von Dialekten [The prosody of Swiss German—insights from a speech synthesis modeling of dialects]. *Linguistk Online* 24:187-224.

Haspelmath, Martin, Matthew S. Dreyer, David Gil, and Bernard Comrie. (eds.) 2005. *The World Atlas of Language Structures.* Oxford: Oxford University Press.

Heeman, Peter A., and James F. Allen. 1999. Speech repairs, intonational phrases and discourse markers: modeling speakers' utterances in spoken dialogue. *Computational Linguistics* 25.4:527-571.

Hermansky, Hynek. 1990. Perceptual linear predictive (plp) analysis of speech. *Journal of the Acoustical Society of America* 87.4:1738-1752.

Hickok, Gregory, and David Poeppel. 2000. Towards a functional neuroanatomy of speech perception. *Trends in Cognitive Sciences* 4.4:131-138.

Hickok, Gregory, and David Poeppel. 2007. Opinion—The cortical organization of speech processing. *Nature Reviews Neuroscience* 8.5:393-402.

Hillard, Dustin, Zhongqiang Huang, Heng Ji, Ralph Grishman, Dilek Hakkani-Tür, Mary Harper, Mari Ostendorf, and Wen Wang. 2006. Impact of automatic comma prediction on POS/name tagging of speech. *Proceedings of SLT-2006*. Aruba.

Hirose, Keikichi, and Nobuaki Minematsu. 2006. Use of prosodic features for speech recognition. *Proceedings of ICSLP-2004*, 1445-1448. Jeju Island, Korea.

Hirschberg, Julia. 1995. Prosodic and other acoustic cues to speaking style in spontaneous and read speech. *Proceedings of ICPhS-1995*, Vol. 2, 36-43. Stockholm, Sweden.

Hirschberg, Julia. 2002. Communication and prosody: functional aspects of prosody. *Speech Communication* 36.1-2:31-43.

Hirschberg, Julia, and Christine H. Nakatani. 1996. A prosodic analysis of discourse segments in direction-giving monologues. *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, 286-293. Santa Cruz, California, USA.

Hirschberg, Julia, Diane Litman, and Marc Swerts. 2004. Prosodic and other cues to speech recognition failures. *Speech Communication* 43.1-2:155-175.

Hirst, Daniel, and Albert di Cristo. 1998. *Intonation Systems: A Survey of Twenty Languages.* Cambridge & New York: Cambridge University Press.

Hofmann, Thomas. 1999. Probabilistic latent semantic analysis. *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence.* Stockholm, Sweden.

Honal, Matthias, and Tanja Schultz. 2005. Automatic disfluency removal on recognized spontaneous speech—rapid adaptation to speaker-dependent disfluencies. *Proceedings of ICASSP-2005*, 969-972. Pennsylvania, Philadelphia, USA.

Houston, Marsha, and Cheris Kramarae. 1991. Speaking from silence: method of silencing and of resistance. *Discourse and Society* 2.4:387-399.

Howell, Peter, and Karima Kadi-Hanifi. 1991. Comparison of prosodic properties between read and spontaneous speech material. *Speech Communication* 10.2:163-169.

Howitt, Andrew Wilson. 2000. Vowel landmark detection. *Proceedings of ICSLP-2000*, 628-631. Beijing, China.

Hsu, Chun-chieh Natalie. 2006. Incremental processing and cues for head-final relative clauses in Chinese. Manuscript. Newark: University of Delaware.

Hsu, Hui-Chuan. 2003. A sonority model of syllable contraction in Taiwanese Southern Min. *Journal of East Asian Linguistics* 12.4:349-377.

Huang, C.-T. James. 1988. 'Shuo shi he you' [On 'be' and 'have' in Chinese]. *Bulletin of the Institute of History and Philology Academia Sinica* 59.1:43-64.

Huang, Jui-Ting, and Lin-Shan Lee. 2006. Improved large vocabulary Mandarin speech recognition using prosodic features. *Proceedings of Speech Prosody 2006.* Dresden, Germany.

Huffman, Marie K. 1987. Measures of phonation type in Hmong. *Journal of the Acoustical Society of America* 81.2:495-504.

Hutchby, Ian, and Robin Wooffitt. 1998. *Conversation Analysis: Principles, Practices, and Applications.* Cambridge & Malden: Polity Press.

Hymes, Dell H. 1974. *Foundations in Sociolinguistics: An Ethnographic Approach.* Philadelphia: University of Pennsylvania Press.

International Telecommunication Union (ITU). 1993. Pulse Code Modulation (PCM) of voice frequencies.

Isačenko, Alexander V., and Hans-Joachim Schädlich. 1964. *Untersuchungen über die deutsche Satzintonation* [*Investigations on German Sentence Intonation*]. Berlin: Akademie-Verlag.

Ishizaki, Masato, and Yasuharu Den. 2001. *Discourse and Dialogue*. Tokyo: University of Tokyo Press. (In Japanese)

Iwano, Koji, and Keikichi Hirose. 1999. Prosodic word boundary detection using statistical modeling of moraic fundamental frequency contours and its use for continuous speech recognition. *Proceedings of ICASSP-1999*, Vol. 1, 133-136. Phoenix. Arizona, USA.

Jakobson, Roman. 1960. Closing statement: linguistics and poetics. *Style in Language*, ed. by Thomas A. Sebeok, 350-377. Cambridge: MIT Press.

Janin, Adam, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. 2003. The ICSI Meeting Corpus. *Proceedings of ICASSP-2003*, Vol. 1, 364-367. Hong Kong, China.

Jaworski, Adam. 1993. *The Power of Silence: Social and Pragmatic Perspectives.* Newbury Park: Sage.

Jaworski, Adam. (ed.) 1997. *Silence: Interdisciplinary Perspectives.* Berlin & New York: Mouton de Gruyter.

Jefferson, Gail. 1984. On stepwise transition from talk about a trouble to inappropriately next-positioned matters. *Structures of Social Action: Studies in Conversation Analysis*, ed. by John. Maxwell Atkinson & John Heritage, 191-222. Cambridge & New York: Cambridge University Press.

Jefferson, Gail. 2004. Glossary of transcript symbols with an introduction. *Conversation Analysis: Studies from the First Generation*, ed. by Gene H. Lerner, 13-31. Amsterdam & Philadelphia: John Benjamins.

Jelinek, Frederick. 1976. Continuous speech recognition by statistical methods. *Proceedings of IEEE*, Vol. 64, 532-556.

Jones, Douglas, Wade Shen, Elizabeth Shriberg, Andreas Stolcke, Teresa Kamm, and Douglas Reynolds. 2005. Two experiments comparing reading with listening for human processing of conversational telephone speech. *Proceedings of INTERSPEECH-2005*, 1145-1148. Lisbon, Portugal.

Juang, Bin H., Stephen E. Levinson, and Man Mohan Sondhi. 1986. Maximum likelihood estimation for multivariate mixture observations of markov chains. *IEEE Transactions on Information Theory* 32.2:307-309.

Juneja, Amit, and Carol Espy-Wilson. 2003. Speech segmentation using probabilistic phonetic feature hierarchy and support vector machines. *Proceedings of IJCNN-2003*, 675-679. Portland, Oregon, USA.

Kaiki, Nobuyoshi, and Yoshinori Sagisaka. 1992. The control of segmental duration in speech synthesis using statistical methods. *Speech Perception, Production and Linguistic Structure*, ed. by Yoh'ichi Tohkura, Eric Vatikiotis-Bateson & Yoshinori Sagisaka, 391-402. Tokyo: Ohmsha.

Kamholz, David. 2005. An ontology for sounds and sound patterns. http://emeld.org/workshop/2005/papers/kamholz-paper.html. (last consulted 2006-10-22)

Kayne, Richard. 1989. Null subjects and clitic climbing. *The Null Subject Parameter*, ed. by Osvaldo Jaeggli and Kenneth J. Safir, 239-261. Dordrecht & Boston: Kluwer Academic Publishers.

Kayne, Richard. 1991. Romance clitics, verb movement and PRO. *Linguistic Inquiry* 22.4:647-686.

Keller, Eric. 1994. Fundamentals of phonetic science. *Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State of the Art, and Future Challenges*, ed. by Eric Keller, 5-21. Chichester & New York: John Wiley.

Kelly, John, and John Local. 1989. On the use of general phonetic techniques in handling conversational material. *Conversation: An Interdisciplinary Perspective*, ed. by Roger Derek & Peter Bull, 197-212. Clevedon: Multilingual Matters.

Kim, Heejin. 2006. *Rhythmic Shortening in American English: Effect of Prosodic Phrase Structure*. University of Illinois at Urbana-Champaign dissertation.

Kim, Joungbum, Sarah E. Schwarm, and Mari Ostendorf. 2004. Detecting structural metadata with decision trees and transformation-based learning. *Proceedings of HLT/NAACL 2004*, 137-144. Boston, Massachusetts, USA.

Kim, Ji-Hwan, and Philip C. Woodland. 2003. A combined punctuation generation and speech recognition system and its performance enhancement using prosody. *Speech Communication* 41.4:563-577.

King, Simon, and Paul Taylor. 2000. Detection of phonological features in continuous speech using neural networks. *Computer Speech and Language* 14.4:333-353.

Kirchhoff, Katrin, G. Fink, and G. Sagerer. 2000. Conversational speech recognition using acoustic and articulatory input. *Proceedings of ICASSP-2000*, Vol. 3, 1435-1438. Istanbul, Turkey.

Klatt, Dennis H. 1975. Vowel lengthening is syntactically determined in a connected discourse. *Journal of Phonetics* 3:129-140.

Koh, Youshaku, and Kiyoaki Shirai. 2005. Creating support of a Dialogue Acts Tagged Corpus. *Proceedings of the 11th Japanese Society for Language Processing*, 815-818. (In Japanese)

Kohler, Klaus. 1987. Categorical pitch perception. *Proceedings of ICPhS-1987*, 331-333. Tallin, Estonian SSR.

Kohler, Klaus J. 1991. *Studies in German Intonation*. Kiel: Institut für Phonetik und digitale Sprachverarbeitung, Universität Kiel.

Kohler, Klaus J. 1996. Articulatory reduction in German spontaneous speech. *1st ETRW on Speech Production Modeling: From Control Strategies to Acoustics*, 1-4. Autrans, France.

Kolar, Jachym, Elizabeth Shriberg, and Yang Liu. 2006. On speaker-specific prosodic models for automatic dialog act segmentation of multi-party meetings. *Proceedings of ICSLP-2006*, 2014-2017. Pittsburgh, Pennsylvania, USA.

Kushan, Surana, and Janet Slifka. 2006. Is irregular phonation a reliable cue towards the segmentation of continuous speech in American English? *Proceedings of Speech Prosody 2006.* Dresden, Germany.

Laan, Gitta P. M. 1997. The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and a read speaking style. *Speech Communication* 22.1:43-65.

Labov, William. 1972. *Sociolinguistic Patterns.* Philadelphia: University of Pennsylvania Press.

Ladd, Dwight Robert. 1980. *The Structure of Intonational Meaning: Evidence from English.* Bloomington: Indiana University Press.

Ladd, Dwight Robert. 1996. *Intonational Phonology.* Cambridge & New York: Cambridge University Press.

Ladefoged, Peter. 1971. *Preliminaries to Linguistic Phonetics*. Chicago: Chicago University Press.

Ladefoged, Peter, and Ian Maddieson. 1997. *The Sounds of the World's Languages*. Oxford: Blackwell.

Lakoff, Robin T. 1995. Cries and whispers: the shattering of the silence. *Gender Articulated: Language and the Socially Constructed Self*, ed. by Kira Hall & Mary Bucholtz, 25-50. New York: Routledge.

Lalljee, Mansur, and Mark Cook. 1973. Uncertainty in first encounters. *Journal of Personality and Social Psychology* 26.1:137-141.

Laver, John. 1980. *The Phonetic Description of Voice Quality.* Cambridge & New York: Cambridge University Press.

Lee, Kai-Fu, and Hsiao-Wuen Hon. 1989. Speaker-independent phone recognition using Hidden Markov Models. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 37.11:1641-1648.

Lee, Tzu-Lun, Ya-Fang He, Yun-Ju Huang, Shu-Chuan Tseng, and Robert Eklund. 2004. Prolongation in spontaneous Mandarin. *Proceedings of ICSLP-2004*, 2181-2184. Jeju Island, Korea.

Leech, Geoffrey N. 1983. *Semantics: The Study of Meaning.* New York: Longman.

Lehiste, Ilse. 1972. The timing of utterances and linguistic boundaries. *Journal of the Acoustical Society of America* 51.6:2018-2024.

Lendvai, Piroska, Antal van den Bosch, and Emile Krahmer. 2003. Memory-based disfluency chunking. *Proceedings of ISCA Tutorial and Research Workshop on Disfluency in Spontaneous Speech* (*DiSS'03*), 63-66. Göteborg, Sweden.

Leonard, Laurence B., Marc E. Fey, and Marilyn Newhoff. 1981. Phonological considerations in children's early imitative and spontaneous speech. *Journal of Psycholinguistic Research* 10.2:123-133.

Levelt, William J. M. 1983. Monitoring and self-repair in speech. *Cognition* 14.1:41-104.

Levelt, William J. M. 1989. *Speaking: From Intention to Articulation.* Cambridge: MIT Press.

Levelt, William J. M., and Anne Cutler. 1983. Prosodic marking in speech repair. *Journal of Semantics* 2.1:205-217.

Levinson, Stephen C. 1983. *Pragmatics.* Cambridge & New York: Cambridge University Press.

Li, Charles N., and Sandra A. Thompson. 1982. The Gulf between spoken and written language: a case study in Chinese. *Spoken and Written Language: Exploring Orality and Literacy*, ed. by Deborah Tannen, 77-88. Norwood: Ablex.

Lickley, Robin J. 1996. Juncture cues to disfluency. *Proceedings of ICSLP-1996*, 2478-2481. Philadelphia, Pennsylvania, USA.

Lieberman, Philip, William Katz, Allard Jongman, Roger Zimmerman, and Mark Miller. 1985. Measures of the sentence intonation of read and spontaneous speech in American English. *Journal of the Acoustical Society of America* 77.2:649-657.

Lin, Che-Kuang, and Lin-Shan Lee. 2005. Improved spontaneous Mandarin speech recognition by disfluency interruption point (IP) detection using prosodic features. *Proceedings of INTERSPEECH-2005*, 1621-1624. Lisbon, Portugal.

Lin, Che-Kuang, Shu-Chuan Tseng, and Lin-Shan Lee. 2005. Important and new features with analysis for disfluency interruption point (IP) detection in spontaneous Mandarin speech. *Proceedings of ISCA Tutorial and Research Workshop on Disfluency in Spontaneous Speech* (*DiSS'05*), 117-121. Aix-en-Provence, France.

Lin, Che-Kuang, and Lin-Shan Lee. 2006. Latent Prosodic Modeling (LPM) for speech with applications in recognizing spontaneous Mandarin speech with disfluencies. *Proceedings of ICSLP-2006*, 2390-2393. Pittsburgh, Pennsylvania, USA.

Litman, Diane J., and James F. Allen. 1990. Discourse processing and commonsense plan. *Intentions in Communication*, ed. by Philip R. Cohen, Jerry L. Morgan & Martha E. Pollack, 365-388. Cambridge: MIT Press.

Liu, Jun. 2002. Negotiating silence in American classrooms: three Chinese cases. *Language and Intercultural Communication* 2.1:37-54.

Liu, Sharlene Anne. 1995. *Landmark Detection for Distinctive Feature-based Speech Recognition*. Cambridge: MIT dissertation.

Liu, Vincent. 2005. *Utterance-final Particles in Mandarin Conversations: Function and Representation of A*. Taipei: Fu Jen Catholic University MA thesis.

Liu, Yuehua. (ed.) 1998. *On Directional Complements.* Beijing: Beijing Language and Culture University. (In Chinese)

Liu, Yang, Elizabeth Shriberg, and Andreas Stolcke. 2003. Automatic disfluency identification in conversational speech using multiple knowledge sources. *Proceedings of EUROSPEECH-2003*, 957-960. Geneva, Switzerland.

Liu, Yang, Elizabeth Shriberg, Andreas Stolcke, and Mary Harper. 2004. Using machine learning to cope with imbalanced classes in natural speech: evidence from sentence boundary and disfluency detection. *Proceedings of ICSLP-2004.* Jeju Island, Korea.

Liu, Yang, Elizabeth Shriberg, Andreas Stolcke, and Mary Harper. 2005. Comparing HMM, maximum entropy, and conditional random fields for disfluency detection. *Proceedings of INTERSPEECH-2005*, 3313-3316. Lisbon, Portugal.

Liu, Yang, Elizabeth Shriberg, Andreas Stolcke, Barbara Peskin, Jeremy Ang, Dustin Hillard, Mari Ostendorf, Marcus Tomalin, Phil Woodland, and Mary Harper. 2005. Structural metadata research in the EARS program. *Proceedings of ICASSP-2005*, 957-960. Philadelphia, Pennsylvania, USA.

Liu, Yang, Elizabeth Shriberg, Andreas Stolcke, Dustin Hillard, Mari Ostendorf, and Mary Harper. 2006. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Transactions on Audio, Speech, and Language Processing* 14.5:1526-1540.

Livescu, Karen. 2005. *Feature-Based Pronunciation Modeling for Automatic Speech Recognition*. Cambridge: MIT dissertation.

Livescu, Karen, and James Glass. 2004. Feature-based pronunciation modeling with trainable asynchrony probabilities. *Proceedings of INTERSPEECH-2004*, 677-680. Jeju Island, Korea.

Livescu, Karen, Özgür Çetin, Mark Hasegawa-Johnson, Simon King, Chris Bartels, Nash Borges, Arthur Kantor, Partha Lal, Lisa Yung, Ari Bezman, Stephen Dawson-Hagerty, Bronwyn Woods, Joe Frankel, Mathew Magimai-Doss, and Kate Saenko. 2007. Articulatory Feature-Based Methods for Acoustic and Audio-visual Speech Recognition: 2006 JHU Summer Workshop Final Report. Johns Hopkins University Center for Language and Speech Processing.

Llisterri, Joaquim, and Dolors Poch. (eds.) 1991. *Proceedings of the ESCA Workshop 'Phonetics and Phonology of Speaking Styles: Reduction and Elaboration in Speech Communication'.* Barcelona, Spain.

Local, John. 1992. Continuing and restarting. *The Contextualization of Language*, ed. by Peter Auer & Aldo di Luzio, 273-296. Amsterdam & Philadelphia: John Benjamins.

Local, John. 1996. Conversational phonetics: some aspects of news receipts in everyday talk. *Prosody in Conversation: Interactional Studies*, ed. by Elizabeth Couper-Kuhlen & Margret Selting, 175-230. Cambridge & New York: Cambridge University Press.

Local, John. 2003. Phonetics and talk-in-interaction. *Proceedings of ICPhS-2003*, 115-118. Barcelona, Spain.

Local, John, John Kelly, and William Wells. 1986. Towards a phonology of conversation: turn-taking in Tyneside English. *Journal of Linguistics* 22.2:411-437.

Local, John, and John Kelly. 1989. *Doing Phonology: Observing, Recording, Interpreting.* Manchester: Manchester University Press.

Local, John, and Gareth Walker. 2004. Abrupt-joins as a resource for the production of multi-unit multi-action turns. *Journal of Pragmatics* 36.8:1375-1403.

Lu, Jianming. 1999. '*De*' zi jiegou han '*suo*' zi jiegou [On the *de* construction and the *suo* construction]. *Xiandai Hanyu Xuci Sanlun* [*Essays on Functional Particles in Modern Chinese*], by Jianming Lu & Zhen Ma, 243-258. Beijing: Language and Culture Press.

Luo, Xiaoqiang, and Frederick Jelinek. 1999. Probabilistic classification of hmm states for large vocabulary continuous speech recognition. *Proceedings of ICASSP-1999*, 353-356. Phoenix, Arizona, USA.

Maclay, Howard, and Charles E. Osgood. 1959. Hesitation phenomena in spontaneous English speech. *Word* 15.1:19-44.

Maddieson, Ian, and Susan Hess. 1987. The effects of F0 of the linguistic use of phonation type. *UCLA Working Papers in Phonetics* 67:112-118.

Maekawa, Kikuo. 2003. Corpus of Spontaneous Japanese: is design and evaluation. *Proceedings of ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition* (*SSPR-2003*), 7-12. Tokyo Institute of Technology, Tokyo, Japan.

Maekawa, Kikuo. 2004. Design, compilation, and some preliminary analyses of the Corpus of Spontaneous Japanese. *Spontaneous Speech: Data and Analysis*, ed. by Kiyoko Yoneyama & Kikuo Maekawa, 87-108. Tokyo: The National Institute for Japanese Language.

Maekawa, Kikuo. 2005. Quantitative analysis of word-form variation using a spontaneous speech corpus. *Proceedings of Corpus Linguistics 2005.* Birmingham, UK.

Maekawa, Kikuo. 2005. Toward a pronunciation dictionary of Japanese: analysis of CSJ. *Proceedings of Symposium on Large-Scale Knowledge Resources* (*LKR2005*), 43-48. Tokyo Institute of Technology 21st Century COE Program.

Maekawa, Kikuo, Hideaki Kikuchi, Yosuke Igarashi, and Jennifer Venditti. 2002. X-JToBI: an extended J_ToBI for spontaneous speech. *Proceedings of ICSLP-2002*, 1545-1548. Denver, Colorado, USA.

Maekawa, Kikuo, Hanae Koiso, Hideaki Kikuchi, and Kiyoko Yoneyama. 2003. Use of a large-scale spontaneous speech corpus in the study of linguistic variation. *Proceedings of ICPhS-2003*, 643-646. Barcelona, Spain.

Maekawa, Kikuo, and Hideaki Kikuchi. 2005. Corpus-based analysis of vowel devoicing in spontaneous Japanese: an interim report. *Voicing in Japanese*, ed. by Jeroen Maarten van de Weijer, Kensuke Nanjo & Tetsuo Nishihara, 205-228. Berlin & New York: Mouton de Gruyter.

Maekawa, Kikuo, and Yosuke Igarashi. 2006. Prosodic independence of bimoraic accented particles: analysis of the Corpus of Spontaneous Japanese. *Journal of the Phonetic Society of Japan* 10.2:33-42. (In Japanese)

Makhoul, John, Alex Baron, Ivan Bulyko, Long Nguyen, Lance Ramshaw, David Stallard, Richard Schwartz, and Bing Xiang. 2005. The effects of speech recognition and punctuation on information extraction performance. *Proceedings of INTERSPEECH-2005*, 57-60. Lisbon, Portugal.

Marcu, Daniel. 2000. The rhetorical parsing of unrestricted texts: a surface-based approach. *Computational Linguistics* 26.3:395-448.

Markel, John D. 1972. The sift algorithm for fundamental frequency estimation. *IEEE Transactions on Audio and Electroacoustics* 20:367-377.

Matusov, E., D. Hillard, M. Magimai-Doss, D. Hakkani-Tür, Mari Ostendorf, and H. Ney. 2007. Improving speech translation with automatic boundary prediction. *Proceedings of the INTERSPEECH-2007*, 2449-2452. Antwerp, Belgium.

McAllister, Don, Lawrence Gillick, Francesco Scattone, and Michael Newman. 1998. Fabricating conversational speech data with acoustic models: a program to examine model-data mismatch. *Proceedings of the ICSLP-1998*, 1847-1850. Sydney, Australia.

Mendoza-Denton, Norma. 1995. Pregnant pauses: silence and authority in the Anita Hill-Clarence Thomas hearings. *Gender Articulated: Language and the Socially Constructed Self*, ed. by Kira Hall & Mary Bucholtz, 51-66. New York: Routledge.

Mesgarani, Nima, Shihab A. Shamma, and Malcolm Slaney. 2004. Speech discrimination based on multiscale spectrotemporal features. *Proceedings of ICASSP-2004*, Vol. 1, 601-604. Montreal, Quebec, Canada.

Meteer, Marie, and Ann Taylor. 1995. Dysfluency Annotation Stylebook for the Switchboard Corpus. Linguistic Data Consortium.

Mrozinski, Joanna, Edward W. D. Whittaker, Pierre Chatain, and Sadaoki Furui. 2006. Automatic sentence segmentation of speech for automatic summarization. *Proceedings of ICASSP-2006.* Toulouse, France.

Musselman, Carol, and Gonul Kircaali-Iftar. 1996. The development of spoken language in deaf children: explaining the unexplained variance. *Journal of Deaf Studies and Deaf Education* 1.2:108-121.

Nagata, Masaki, and Tsuyoshi Morimoto. 1994. First steps towards statistical modeling of dialogue to predict the speech act type of the next utterance. *Speech Communication* 15.3-4:193-203.

Nakatani, Christine H., and Julia Hirschberg. 1994. A corpus-based study of repair cues in spontaneous speech. *Journal of the Acoustical Society of America* 95.3:1603-1616.

Nam, Hosung, and Elliot Saltzman. 2003. A competitive, coupled oscillator model of syllable structure. *Proceedings of ICPhS-2003*, Vol. 3, 2253-2256. Barcelona, Spain.

National Institute for Japanese Language. 2006. Construction of the Corpus of Spontaneous Japanese. *NIJL Research Report*, No.124. Tokyo: National Institute for Japanese Language. (In Japanese)

Neti, Chalapathy, Gerasimos Potamianosand Juergen Luettin, Iain Matthews, Hervé Glotin, Dimitra Vergyri, June Sison, Azad Mashari & Jie Zhou. 2000. Audio-Visual Speech Recognition: Final Report. Johns Hopkins University Center for Language and Speech Processing.

Ní Chasaide, Ailbhe, and Christer Gobl. 1997. Voice source variation. *The Handbook of Phonetic Sciences*, ed. by William J. Hardcastle & John Laver, 427-461. Cambridge: Blackwell.

Niyogi, Partha, and Padma Ramesh. 1998. Incorporating voice onset time to improve letter recognition accuracies. *Proceedings of ICASSP-1998*, 13-16. Seattle, Washington, USA.

Niyogi, Partha, and Chris Burges. 2002. Detecting and Interpreting Acoustic Features by Support Vector Machines. Chicago: Computer Science Department, University of Chicago.

Nossair, Zaki B., and Stephen A. Zahorian. 1991. Dynamic spectral shape features as acoustic correlates for initial stop consonants. *Journal of the Acoustical Society of America* 89.6:2978-2991.

O'Connell, Dennis C., and Sabine Kowal. 1983. Pausology. *Computers in Language Research*, Vol. 1, ed. by Walter A. Sedelow & Sally Yeates Sedelow, 221-301. Berlin & New York: Mouton.

Odell, J. J., P. C. Woodland, and S. J. Young. 1994. Tree-based state clustering for large vocabulary speech recognition. *Proceedings of the International Symposium on Speech, Image Processing & Neural Networks Proceedings of the International Symposium on Speech, Image Processing & Neural Networks*, 690-693. Hong Kong, China.

Ogura, Hideki, Masaya Yamaguchi, Kenya Nishikawa, Kyoko Ishizuka, and Mutsuko Kimura. 2004. Summary of morphological information in Corpus of Spontaneous Japanese version 1.0. http://www.kokken.go.jp/katsudo/seika/corpus/public/manuals/pos.pdf. (In Japanese)

O'Shaughnessy, Douglas. 1995. Timing patterns in fluent and disfluent spontaneous speech. *Proceedings of ICASSP-1995*, 600-603. Detroit, Michigan, USA.

Ostendorf, Mari. 2000. Incorporating linguistic theories of pronunciation variation into speech-recognition models. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences* 358.1769:1325-1338.

Ostendorf, Mari, Patricia J. Price, and Stefanie Shattuck-Hufnagel. 1995. The Boston University Radio Speech Corpus. Linguistic Data Consortium.

Ostendorf, Mari, I. Shafran, Stefanie Shattuck-Hufnagel, L. Carmichael, and W. Byrne. 2001. A prosodically labeled database of spontaneous speech. *Proceedings of ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding* 22:119-121. Red Bank, New Jersey, USA.

Park, Joseph Sung-Yul. 2002. Cognitive and interactional motivations for the intonation unit. *Studies in Language* 26.3:637-680.

Parsons, Thomas W. 1987. *Voice and Speech Processing.* New York: McGraw-Hill.

Patterson, E. K., S. Gurbuz, Z. Tufecki, and J. N. Gowdy. 2002. CUAVE: a new audio-visual database for multimodal human-computer interface research. *Proceedings of ICASSP-2002*, Vol. 2, 2021-2024. Orlando, Florida, USA.

Peperkamp, Sharon. 1999. Prosodic words. *Glot International* 4.4:15-16.

Peters, Benno, Klaus J. Kohler, and Thomas Wesener. 2005. Melodische Satzakzentmuster in prosodischen Phrasen deutscher Spontansprache. Statistische Verteilung und sprachliche Funktion [Melodic sentence accent patterns in prosodic phrases of German spontaneous speech. Statistical distribution and linguistic function]. *Prosodic Structures in German Spontaneous Speech*, ed. by Klaus J. Kohler, Felicitas Kleber & Benno Peters, 7-54. Institut für Phonetik und Digitale Sprachverarbeitung, Universität Kiel.

Peters, Jörg. 2004. *Intonatorische Variation im Deutschen. Studien zu ausgewählten Regionalsprachen* [*Intonational Variation in German: Studies on Selected Regional Varieties*]. Habilitationsschrift: Universität Potsdam.

Pierrehumbert, Janet. 1980. *The Phonology and Phonetics of English Intonation*. Cambridge: MIT dissertation.

Pierrehumbert, Janet. 1989. A preliminary study of the consequences of intonation for the voice source. *STL-QPSR* 4:23-36.

Press, William H., Brian P. Flannery, Saul A. Teukolsky, and William T. Vetterling. 1988. *Numerical Recipes in C: The Art of Scientific Computing.* Cambridge & New York: Cambridge University Press.

Pruthi, Tarun, and Carol Y. Espy-Wilson. 2004. Acoustic parameters for automatic detection of nasal manner. *Speech Communication* 43.3:225-239.

Rao, Sharath, Ian Lane, and Tanja Schultz. 2007. Optimizing sentence segmentation for spoken language translation. *Proceedings of INTERSPEECH-2007*, 2845-2848. Antwerp, Belgium.

Redi, Laura, and Stefanie Shattuck-Hufnagel. 2001. Variation in the realization of glottalization in normal speakers. *Journal of Phonetics* 29.4:407-429.

Reithinger, Norbert, and Elisabeth Maier. 1997. Dialogue act classification using language models. *Proceedings of EUROSPEECH-1997*, 2235-2238. Rhodes, Greece.

Reynolds, Douglas. 2003. Channel robust speaker verification via feature mapping. *Proceedings of ICASSP-2003*, Vol. 3, 53-56. Hong Kong, China.

Richardson, Matt, Jeff Bilmes, and Chris Diorio. 2000. Hidden-articulator Markov Models: performance improvements and robustness to noise. *Proceedings of ICSLP-2000*, 131-134. Beijing, China.

Richmond, Korin, Simon King, and Paul Taylor. 2003. Modeling the uncertainty in recovering articulation from acoustics. *Computer Speech and Language* 17.2-3: 153-172.

Riley, Michael, William Byrne, Michael Finke, Sanjeev Khudanpur, Andrej Ljolje, John McDonough, Harriet Nock, Murat Saraclar, Charles Wooters, and George Zavaliagkos. 1999. Stochastic pronunciation modelling from hand-labelled phonetic corpora. *Speech Communication* 29.2-4:209-224.

Rose, Richard C., and Giuseppe Riccardi. 1999. Modeling disfluency and background events in ASR for a natural language understanding task. *Proceedings of ICASSP-1999*, 341-344. Phoenix, Arizona, USA.

Rotman, Brian. 1987. *Signifying Nothing: The Semiotics of Zero.* New York: St. Martin's Press.

Sacks, Harvey. 1984. Notes on methodology. *Structures of Social Action: Studies in Conversation Analysis*, ed. by John Maxwell Atkinson & John Heritage, 21-27. Cambridge: Cambridge University Press.

Sacks, Harvey. 1987. On the preferences for agreement and contiguity in sequences in conversation. *Talk and Social Organisation*, ed. by Graham Button & John R. E. Lee, 54-69. Clevedon: Multilingual Matters.

Sacks, Harvey, Emanuel A. Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language* 50.4:696-735.

Saltzman, Elliot L., and Kevin J. Munhall. 1989. A dynamical approach to gestural patterning in speech production. *Status Report on Speech Research* (July-December 1989), 38-68. New Haven: Haskins Laboratories.

Samuel, Ken, Sandra Carberry, and K. Vijay-Shanker. 1998. Dialogue act tagging with transformation-based learning. *Proceedings of COLING-ACL-98*, 1150-1156. Montreal, Quebec, Canada.

Sassen, Claudia. 2005. *Linguistic Dimensions of Crisis Talk: Formalising Structures in A Controlled Language.* Amsterdam & Philadelphia: John Benjamins.

de Saussure, Ferdinand. 1916. *Cours de linguistique générale. Redigé par Charles Bally et Albert Séchehaye*. Paris & Lausanne: Payot.

Saville-Troike, Muriel. 1985. The place of silence in an integrated theory of communication. *Perspectives on Silence*, ed. by Deborah Tannen & Muriel Saville-Troike, 3-18. Norwood: Ablex.

Schapire, Robert E., and Yoram Singer. 2000. Boostexter: a boosting-based system for text categorization. *Machine Learning* 39.2-3:135-168.

Schegloff, Emanuel A. 1982. Discourse as an interactional achievement: some uses of 'uh huh' and other things that come between sentences. *Analyzing Discourse: Text and Talk*, ed. by Deborah Tannen, 71-93. Washington, DC: Georgetown University Press.

Schmid, Karl. 1915. *Die Mundart des Amtes Entlebuch im Kanton Luzern* [*The Dialect of the Entlebuch District in the Canton of Lucerne*]. Frauenfeld: Huber.

Schwenk, Holger, and Jean-Luc Gauvain. 2000. Combining multiple speech recognizers using voting and language model information. *Proceedings of ICSLP-2000*, 915-918. Beijing, China.

Selkirk, Elisabeth O. 1981. *The Phrase Phonology of English and French.* Bloomington: Indiana University Linguistics Club.

Selkirk, Elisabeth O. 1984. *Phonology and Syntax: The Relation between Sound and Structure.* Cambridge: MIT Press.

Selting, Margret. 1995. *Prosodie im Gespräch* [*Prosody in Conversation*]. Tübingen: Max Niemeyer.

Selting, Margret. 2003. Treppenkonturen im Dresdenerischen [Step contours in the Dresden variety]. *Zeitschrift für germanistische Linguistik* 31:1-43.

Selting, Margret, Peter Auer, Birgit Barden, and Jörg Bergmann. 1998. Gesprächsanalytisches Transkriptionssystem GAT [Conversation analytic transcription system GAT]. *Linguistische Berichte* 173:91-122.

Shimazu, Akira, Hiroshi Taguchi, and Masahito Kawamori. 2000. Characteristics of acceptance utterances and their relations to dialog strategies. *Proceedings of the Third International Workshop on Human-Computer Conversation*, 152-157. Bellagio, Italy.

Shriberg, Elizabeth. 1994. *Preliminaries to a Theory of Speech Disfluencies*. Berkeley: University of California dissertation.

Shriberg, Elizabeth. 1999. Phonetic consequences of speech disfluency. *Proceedings of ICPhS-1999*, 619-622. San Francisco, California, USA.

Shriberg, Elizabeth. 2001. To ‘errr’ is human: ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association* 31.1:153-169.

Shriberg, Elizabeth. 2005. Spontaneous speech: how people really talk and why engineers should care. *Proceedings of INTERSPEECH-2005*, 1781-1784. Lisbon, Portugal.

Shriberg, Elizabeth, Andreas Stolcke, Dilek Hakkani-Tür, and Gokhan Tur. 2000. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication* 32.1-2:127-154.

Shriberg, Elizabeth, and Andreas Stolcke. 2004. Prosody modeling for automatic speech recognition and understanding. *Mathematical Foundations of Speech and Language Processing*, ed. by Mark E. Johnson, 105-114. New York: Springer.

Shriberg, Elizabeth, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The ICSI Meeting Recorder Dialog Act (MRDA) Corpus. *Proceedings of 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, 97-100. Cambridge, Massachusetts, USA.

Siebenhaar, Beat. 2004. Berner und Zürcher Prosodie. Ansätze zu einem Vergleich. Alemannisch im Sprachvergleich [Bernese and Zurich prosody. Approaches to a comparison]. *Beiträge zur 14. Arbeitstagung für alemannische Dialektologie in Männedorf (Zürich) vom 16.-18.9.2002*, ed. by Elvira Glaser, Peter Ott & Rudolf Schwarzenbach, 419-437 Stuttgart: Franz Steiner.

Siebenhaar, Beat et al. 2004. Prosody of Bernese and Zurich German. What the development of a dialectal speech synthesis system tells us about it. *Regional Variation in Intonation*, ed. by Peter Gilles & Jörg Peters, 219-238. Tübingen: Niemeyer.

Siegel, Sidney, and N. John Castellan, Jr. 1988. *Nonparametric Statistics for the Behavioral Sciences* (2nd edition). New York: McGraw-Hill.

Siegman, Aron Wolf, and Benjamin Pope. 1965. Effects of question specificity and anxiety-producing messages on verbal fluency in the initial interview. *Journal of Personality and Social Psychology* 2.4:522-530.

Sievers, Eduard. 1912. *Rhythmisch-melodische Studien* [*Rhythmic-melodic Studies*]. Heidelberg: Carl Winter.

Sifianou, Maria. 1997. Silence and politeness. *Silence: Interdisciplinary Perspective*, ed. by Adam Jaworski, 63-84. Berlin & New York: Mouton de Gruyter.

Silverman, Kim, Mary Beckman, John Pitrelli, Mari Ostendorf, Colin Wightman, Patti Price, Janet B. Pierrehumbert, and Julia Hirschberg. 1992. ToBI: A standard for labelling English prosody. *Proceedings of ICSLP-1992*, 867-870. Banff, Alberta, Canada.

Smith, Vicki L., and Herbert H. Clark. 1993. On the course of answering questions. *Journal of Memory and Language* 32.1:25-38.

Sönmez, M. Kemal, Larry Heck, Mitch Weintraub, and Elizabeth Shriberg. 1997. A lognormal tied mixture model of pitch for prosody-based speaker recognition. *Proceedings of EUROSPEECH-1997*, 1391-1394. Rhodes, Greece.

Steedman, Mark. 2000. Information structure and the syntax-phonology interface. *Linguistic Inquiry* 31.4:649-689.

Stephens, Jane, and Geoffrey Beattie. 1986. On judging the ends of speaker turns in conversation. *Journal of Language and Social Psychology* 5.2:119-134.

Stevens, Ken N., Sharon Y. Manuel, Stefanie Shattuck-Hufnagel, and Sharlene Liu. 1992. Implementation of a model for lexical access based on features. *Proceedings of ICSLP-1992*, Vol. 1, 499-502. Banff, Alberta, Canada.

Stolcke, Andreas, H. Bratt, J. Butzberger, Horacio Franco, Anand Venkataraman, Madeleine Plauche, C. Richey, Elizabeth Shriberg, Kemal Sonmez, Fu-Liang Weng, and Jing Zheng. 2000. The SRI March 2000 Hub-5 Conversational Speech Transcription System. *Proceedings of NIST Workshop on Speech Transcription 2000*, P11:11-14. College Park, Maryland, USA.

Stolcke, Andreas, H. Franco, R. Gadde, M. Graciarena, K. Precoda, M. Venkataraman, D. Vergyri, W. Wang, J. Zheng, Y. Huang, B. Peskin, I. Bulyko, Mari Ostendorf, and K. Kirchhoff. 2003. Speech-to-Text Research at SRI-ICSI-UW. NIST RT-03. Workshop. Boston, Massachusetts, USA.

Stolcke, Andreas, Barry Chen, Horacio Franco, Venkata Ramana Rao Gadde, Martin Graciarena, Mei-Yuh Hwang, Katrin Kirchhoff, Arindam Mandal, Nelson Morgan, Xin Lin, Tim Ng, Mari Ostendorf, Kemal Sönmez, Anand Venkataraman, Dimitra Vergyri, Wen Wang, Jing Zheng, and Qifeng Zhu. 2006. Recent innovations in speech-to-text transcription at SRI-ICSI-UW. *IEEE Transactions on Audio, Speech, and Language Processing* 14.5:1729-1744.

Strassel, Stephanie, and Meghan Glenn. 2003. Creating the annotated TDT-4 Y2003 Evaluation Corpus. TDT 2003 Evaluation Workshop, NIST.

Sundaram, Ram, Aravind Ganapathiraju, Jonathan Hamaker, and Joseph Picone. 2000. ISIP 2000 conversational speech evaluation system. NIST Evaluation of Conversational Speech Recognition over the Telephone.

Swerts, Marc. 1998. Filled pauses as markers of discourse structure. *Journal of Pragmatics* 30.4:485-496.

Swerts, Marc, and Ronald Geluykens. 1994. Prosody as a marker of information flow in spoken discourse. *Language and Speech* 37:21-43.

Swerts, Marc, Eva Strangert, and Mattias Heldner. 1996. F0 declination in read-aloud and spontaneous speech. *Proceedings of the ICSLP-1996*, Vol. 3, 1501-1504. Philadelphia, Pennsylvania, USA.

Swerts, Marc, and Raymond Veldhuis. 2001. The effect of speech melody on voice quality. *Speech Communication* 33.4:297-303.

Syrdal, Ann K. 1996. Acoustic variability in spontaneous conversational speech of American English talkers. *Proceedings of ICSLP-1996*, Vol. 3, 438-441. Philadelphia, Pennsylvania, USA.

Taboada, Maite. 2006. Spontaneous and non-spontaneous turn-taking. *Pragmatics* 16.2-3:329-360.

Takanashi, Katsuya, Kiyotaka Uchimoto, and Takehiko Maruyama. 2004. Manual of clause unit annotation in Corpus of Spontaneous Japanese Version 1.0. http://www.kokken.go.jp/katsudo/seika/corpus/public/manuals/clause.pdf. (In Japanese)

Takano, Kenji, and Akira Shimazu. 2004. Analysis of a transport route guidance dialogue: local structures. *Proceedings of the 10th Japanese Society for Language Processing*, 181-184. (In Japanese)

Takeda, Kazuya, Yoshinori Sagisaka, and Hisao Kuwabara. 1989. On sentence-level factors governing segmental duration in Japanese. *Journal of the Acoustical Society of America* 86.6:2081-2087.

Takeuchi, Kazuhiro, Ikuyo Morimoto, Katsuya Takanashi, and Hitoshi Isahara. 2004. On discourse boundary information in Corpus of Spontaneous Japanese Version 1.0. http://www.kokken.go.jp/katsudo/seika/corpus/public/manuals/discourse.pdf. (In Japanese)

Talkin, David. 1995. A robust algorithm for Pitch Tracking (RAPT). *Speech Coding and Synthesis*, ed. by W. Bastiaan Kleijn & Kuldip K. Paliwal, 495-518. Amsterdam & New York: Elsevier.

Tang, Ting-chi. 1980. Cleft and pseudo-cleft constructions in Chinese: structure, function and constraint. *Journal of National Taiwan Normal University* 25:249-296.

Tannen, Deborah. 1982. Oral and literate strategies in spoken and written narratives. *Language* 58.1:1-21.

Tannen, Deborah. 1985. Silence: anything but. *Perspectives on Silence*, ed. by Deborah Tannen & Muriel Saville-Troike, 93-111. Norwood: Ablex.

Tao, Hongyin. 1996. *Units in Mandarin Conversation: Prosody, Discourse and Grammar.* Amsterdam & Philadelphia: John Benjamins.

Taylor, Paul. 2000. Analysis and synthesis of intonation using the Tilt model. *Journal of the Acoustical Society of America* 107.3:1697-1714.

Thiesmeyer, Lynn Janet. (ed.) 2003. *Discourse and Silencing: Representation and the Language of Displacement.* Amsterdam & Philadelphia: John Benjamins.

Thongkum, Therapan L. 1987. Another look at the register distinction in Mon. *UCLA Working Papers in Phonetics* 67:132-165.

Ting, Jen. 2003. The nature of the particle *suo* in Mandarin Chinese. *Journal of East Asian Linguistics* 12.2:121-139.

Ting, Jen. 2005. On the syntax of the *suo* construction in classical Chinese. *Journal of Chinese Linguistics* 33.2:233-267.

Ting, Jen. 2006a. Clitic climbing and *suo* in Mandarin Chinese and its implications for universal grammar. Paper presented at the 80th Annual Meeting of the Linguistic Society of America. Albuquerque, New Mexico, USA.

Ting, Jen. 2006b. On the form and function of the particle *suo* in Mandarin Chinese. *The Proceedings of the 18th North American Conference on Chinese Linguistics*, ed. by Janet Zhiqun Xing, 518-534. Los Angeles: GSIL.

Ting, Jen. 2008. The nature of the particle *suo* in the passive constructions in classical Chinese. *Journal of Chinese Linguistics* 36.1:30-72.

Traum, David R. 1994. *A Computational Theory of Grounding in Natural Language Conversation*. Rochester: University of Rochester dissertation.

Traum, David R., and Christine H. Nakatani. 1999. A two-level approach to coding dialogue for discourse structure: activities of the 1998 DRI Working Group on higher-level structures. *ACL-99 Workshop Towards Standards and Tools for Discourse Tagging*, 101-108. Collage Park, Maryland, USA.

Trippel, Thorsten, Felix Sasaki, Benjamin Hell, and Dafydd Gibbon. 2003. Acquiring lexical information from multilevel temporal annotations. *Proceedings of EUROSPEECH-2003*, 2265-2268. Geneva, Switzerland.

Trippel, Thorsten, Dafydd Gibbon, Alexandra Thies, Jan-Torsten Milde, Karin Looks, Benjamin Hell, and Ulrike Gut. 2004. CoGest: a formal transcription system for conversational gesture. *Proceedings of LREC-2004.* Lisbon, Portugal.

Trudgill, Peter. 1974. *The Social Differentiation of English in Norwich.* Cambridge: Cambridge University Press.

Tsai, Yu-Fang, and Keh-Jiann Chen. 2003. Context-rule Model for POS tagging. *Proceedings of PACLIC-17*, 146-151. Singapore.

Tseng, Chiu-yu, Shao-Huang Pin, Yehlin Lee, Hsin-Min Wang, and Yong-Cheng Chen. 2005. Fluent speech prosody: framework and modeling. *Speech Communication* 46.3-4:284-309.

Tseng, Shu-Chuan. 1999. *Grammar, Prosody and Speech Disfluencies in Spoken Dialogues*. Bielefeld: Bielefeld University dissertation.

Tseng, Shu-Chuan. 2004. Processing spoken Mandarin corpora. *Traitement automatique des langues* 45.2:89-108.

Tseng, Shu-Chuan. 2005. Syllable contractions in a Mandarin conversational dialogue corpus. *International Journal of Corpus Linguistics* 10.1:63-83.

Tseng, Shu-Chuan. 2006a. Repairs in Mandarin conversation. *Journal of Chinese Linguistics* 34.1:80-120.

Tseng, Shu-Chuan. 2006b. Linguistic markings of units in spontaneous Mandarin. *Chinese Spoken Language Processing*: *5th International Symposium, ISCSLP 2006, Singapore, December 13-16, 2006 Proceedings*, ed. by Qiang Huo, Bin Ma, Eng-Siong Chng & Haizhou Li, 43-54. Lecture Notes in Computer Science 4274. Berlin & Heidelberg: Springer-Verlag.

Turk, Alice E., and James R. Sawusch. 1997. The domain of accentual lengthening in American English. *Journal of Phonetics* 25.1:25-41.

Vaissière, Jacqueline. 1983. Language-independent prosodic features. *Prosody: Models and Measurements*, ed. by Anne Cutler & Dwight Robert Ladd, 54-66. Berlin & New York: Springer-Verlag.

van Kuijk, David, and Loe Boves. 1999. Acoustic characteristics of lexical stress in continuous telephone speech. *Speech Communication* 27.2:95-111.

Vance, Timothy J. 1987. *An Introduction to Japanese Phonology.* Albany: State University of New York Press.

Venditti, Jennifer. 1997. Japanese ToBI labeling guidelines. *Ohio State University Working Papers in Linguistics* 50:127-162.

Venditti, Jennifer. 2005. The J_ToBI model of Japanese intonation. *Prosodic Typology: The Phonology of Intonation and Phrasing*, ed. by Sun-Ah Jun, 172-200. Oxford & New York: Oxford University Press.

Venkataraman, A., Andreas Stolcke, W. Wang, D. Vergyri, V. R. R. Gadde, and J. Zheng. 2004. SRIs 2004 broadcast news speech to text system. *EARS Rich Transcription 2004 Workshop*. Palisades, New York, USA.

Vergyri, Dimitra, Andreas Stolcke, Venkata R. R. Cadde, Luciana Ferrer, and Elizabeth Shriberg. 2003. Prosodic knowledge source for automatic speech recognition. *Proceedings of ICASSP-2003*, 208-211. Hong Kong, China.

Vetsch, Jakob. 1910. *Die Laute der Appenzeller Mundarten* [*The Sounds of the Appenzell Dialects*]. Frauenfeld: Huber.

Wajskop, Max, Joaquim Llisterri, and Dolors Poch-Olive. (eds.) 1992. *Speech Communication* (Special Issue on Phonetics and Phonology of Speaking Styles: Reduction and Elaboration in Speech Communication) 11.4-5.

Wang, Li. 1958. *Hanyu Shigao* [*History of the Chinese Language*]. Beijing: Science Press.

Warnke, V., R. Kompe, H. Niemann, and E. Nöth. 1997. Integrated dialog act segmentation and classification using prosodic features and language models. *Proceedings of EUROSPEECH-1997*, 207-210. Rhodes, Greece.

Wasow, Thomas. 1997. Remarks on grammatical weight. *Language Variation and Change* 9:81-105.

Watanabe, Michiko, Keikichi Hirose, Yasuharu Den, Shusaku Miwa, and Nobuaki Minematsu. 2006. Factors influencing ratios of filled pauses at clause boundaries in Japanese. *Proceedings of ISCA Tutorial and Research Workshop on Experimental Linguistics*, 253-256. Athens, Greece.

Weintraub, Mitch, Kelsey Taussig, Kate Hunicke-Smith, and Amy Snodgras. 1996. Effect of speaking style on LVCSR performance. *Proceedings of ICSLP-1996*, 16-19. Philadelphia, Pennsylvania, USA.

Wennerstrom, Ann, and Andrew F. Siegel. 2003. Keeping the floor in multiparty conversations: intonation, syntax, and pause. *Discourse Processes* 36.2:77-107.

Wightman, Colin W., Stefanie Shattuck-Hufnagel, Mari Ostendorf, and Patti Price. 1992. Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America* 91.3:1707-1717.

Wipf, Elisa. 1910. *Die Mundart von Visperterminen im Wallis* [*The Dialect of Visperterminen in the Canton of Wallis*]. Frauenfeld: Huber.

Wolf, Florian, and Edward Gibson. 2005. Representing discourse coherence: a corpus-based analysis. *Computational Linguistics* 31.2:249-287.

Wooters, Chuck, James Fung, Barbara Peskin, and Xavier Anguera. 2004. Towards robust speaker segmentation: the ICSI-SRI Fall 2004 Diarization system. RT-04F Workshop.

Xia, Fei, and Lap Cheung. 2006. Features, bagging, and system combination for the Chinese POS tagging task. *Proceedings of 5th SIGHAN Workshop on Chinese Language Processing*, 25-32. Sydney, Australia.

Yamazumi, Kenji, Takayuki Kagomiya, Yohichi Maki, and Kikuo Maekawa. 2005. Psychological scale for the impression rating of monologue. *The Journal of the Acoustical Society of Japan* 61.6:303-311. (In Japanese).

Yoneyama, Kiyoko, Janice Fon, and Hanae Koiso. 2003. Durational and prosodic patterning at discourse boundaries in Japanese spontaneous monologs. *Proceedings of ICPhS-2003*, 2637-2640. Barcelona, Spain.

Yoon, Tae-Jin. 2007. *A Predictive Model of Prosody through Grammatical Interface: A Computational Approach*. Urbana: University of Illinois at Urbana-Champaign dissertation.

Yoon, Tae-Jin, Sandra Chavarria, Jennifer Cole, and Mark Hasegawa-Johnson. 2004. Intertranscriber reliability of prosodic labeling on telephone conversation using ToBI. *Proceedings of INTERSPEECH-2004*, 2729-2732. Jeju Island, Korea.

Yoon, Tae-Jin, Jennifer Cole, Mark Hasegawa-Johnson, and Chilin Shih. 2005. Acoustic correlates of non-modal phonation in telephone speech. *Journal of the Acoustical Society of America* 117.4:2621.

Young, Steve, Gunnar Evermann, Thomas Hain, Dan Kershaw, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, and Phil Woodland. 2002. *The HTK Book.* Cambridge: Engineering Department, Cambridge University.

Young, Steve, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, and Phil Woodland. 2005. *The HTK Book* (version 3.3). Cambridge: Engineering Department, Cambridge University.

Zhang, Jing. 1981. Hanyu jufa jiegou de jiben leixing (shang) [Some basic types of grammatical constructions in Chinese (part I)]. *Zhongguo Yuwen* 1981.3:190- 214.

Zhang, You. 2000. *Information Fusion for Robust Audio-Visual Speech Recognition*. Urbana: University of Illinois at Urbana-Champaign dissertation.

Zheng, Yanli, Mark Hasegawa-Johnson, and Sarah Borys. 2004. Stop consonant classification by dynamic formant trajectory. *Proceedings of INTERSPEECH-2004*, 2481-2484. Jeju Island, Korea.

Zhu, Qifeng, Andreas Stolcke, Barry Y. Chen, and Nelson Morgan. 2005. Using MLP features in SRIs conversational speech recognition system. *Proceedings of INTER-SPEECH-2005*, 2141-2144. Lisbon, Portugal.

Zue, Victor W., and Martha Laferriere. 1979. Acoustic study of medial /t, d/ in American English. *Journal of the Acoustical Society of America* 66.4:1039-1050.

Zue, Victor W., Stephanie Seneff, and James Glass. 1990. Speech database development at MIT: TIMIT and beyond. *Speech Communication* 9.4:351-356.

Zweig, G., J. Bilmes, K. Filali, K. Livescu, P. Xu, K. Jackson, Y. Brandman, E. Sandness, E. Holtz, J. Torres, and B. Byrne. 2002. Structurally discriminative graphical models for automatic speech recognition—results from the 2001 Johns Hopkins Summer Workshop. *Proceedings of ICASSP-2002*, 183-190. Denver, Colorado, USA.

LANGUAGE AND LINGUISTICS

LANGUAGE AND LINGUISTICS is a department of the INSTITUTE OF LINGUISTICS, ACADEMIA SINICA. It furthers the Institute's objective of excellence in research, scholarship, and education by publishing pioneering articles and monographs in Linguistics and other related fields. All publications have been reviewed rigorously according to academic standards.

LANGUAGE AND LINGUISTICS MONOGRAPH SERIES NUMBER A25

Edited by Shu-Chuan Tseng

Linguistic Patterns in Spontaneous Speech

《語言暨語言學》

《語言暨語言學》隸屬於中央研究院語言學研究所，以出版語言學及相關領域之最新研究成果爲宗旨。所有出版品均經過嚴格學術審查。

《語言暨語言學》專刊甲種之二十五

Linguistic Patterns in Spontaneous Speech

曾淑娟　編輯

出版：中央研究院語言學研究所
發行：中央研究院語言學研究所
115 台北市 南港區 研究院路 2 段 128 號
http://www.ling.sinica.edu.tw

印刷：文盛彩藝事業有限公司
初版：中華民國九十八年四月 [2009]
定價：新台幣 600 元／US$35

ISBN-13: 978-986-01-7635-3
GPN: 1009800245