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. Further-

more, 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).

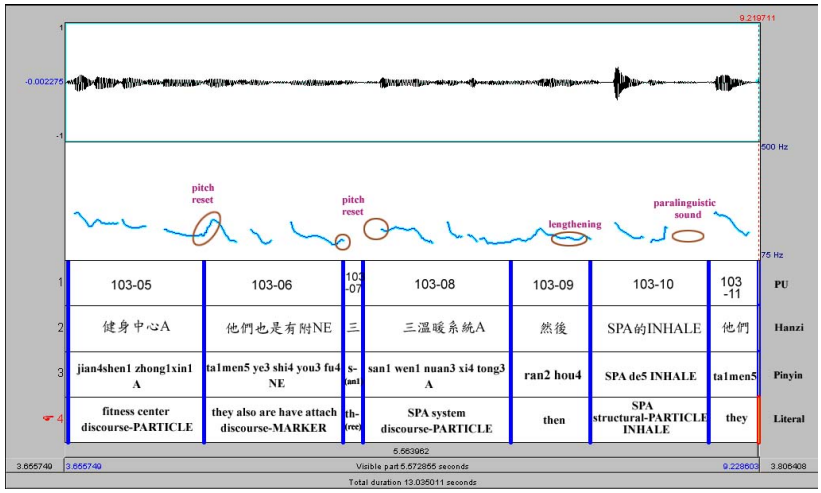

Figure 1: Prosodic units in spoken Mandarin

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).

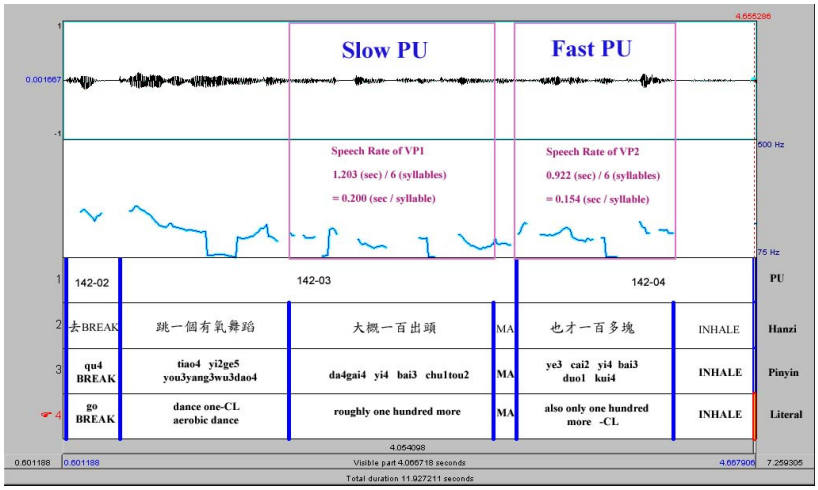

Figure 2: Prosodic units in spoken Mandarin

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 [$_{VP1}$ 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 [$_{VP2}$ 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.

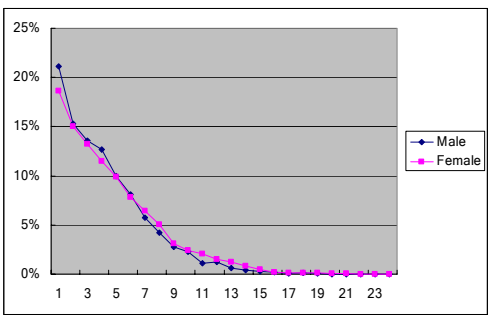

Figure 3a: Unit size in gender

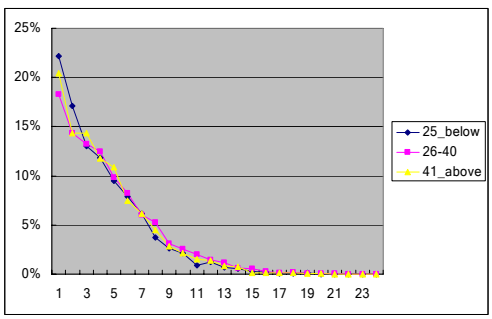

Figure 3b: Unit size in age

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.

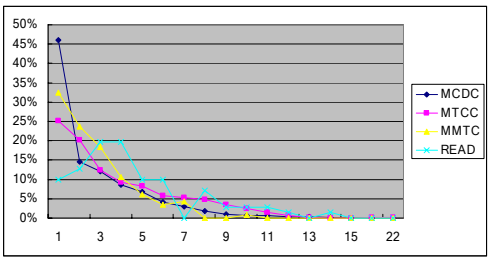

Figure 4a: Unit size, overall data

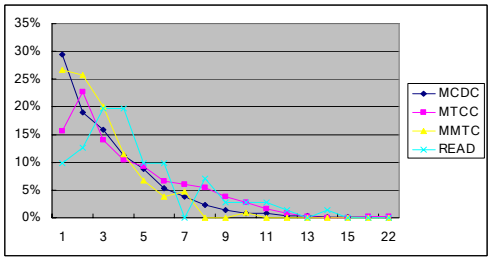

Figure 4b: Unit size, without one-word fillers

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).

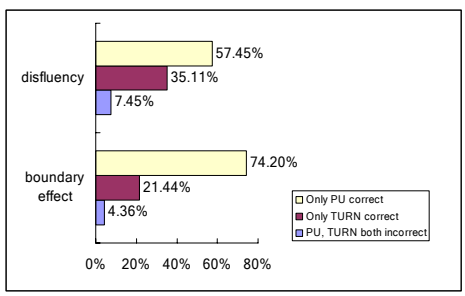

Figure 5a: POS inconsistency

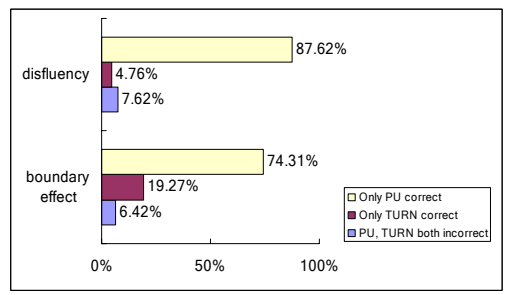

Figure 5b: Word segmentation inconsistency

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.

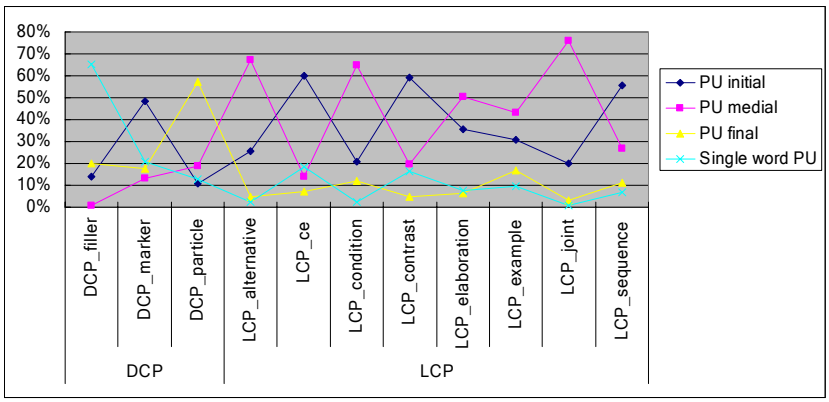

Figure 6: Distribution on cue phrases—multiple speakers

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.

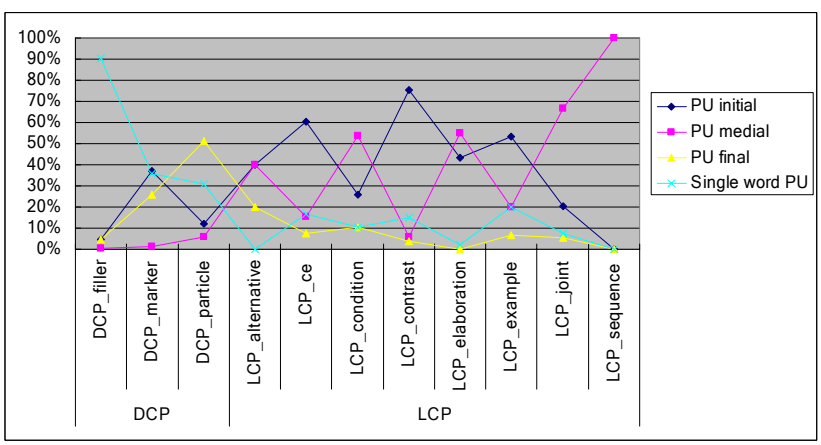

Figure 7: Single speaker's dataset

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