Collecting Mandarin Speech Databases for Prosody Investigations

Chiu-yu Tseng[1], Yun-Ching Cheng, Wei-Shan Lee and Feng-Lan Huang
Institute of Linguistics, Academia Sinica
128 Academia Rd. Sec. 2
Taipei, Taiwan 115
[1]cytling@sinica.edu.tw

Abstract

The prosody of Mandarin running speech is notably marked by grouping of short phrases into perceptually identifiable larger units in the speech flow. An organization of Mandarin speech prosody should not only account for the grouping phenomenon, but also offer some explanation for such grouping in relation to information of other linguistic levels as well as speech planning. The physical, phonetic, acoustic, semantic and syntactic characteristics prosodic units as well as their perceptual properties have been under investigation at our lab. How these units may relate to and combine with preceding as well as following silent portions in the speech flow to constitute the overall phrasing of running speech in general, and how they could be viewed from the perceptive of speech planning in particular have also been the focus of our investigation. We are very much aware of the fact that prosody varies for different speech styles, and therefore devised methods to collect speech corpora of different speech styles. Since our aim was to capture the characteristics that constitute the overall flow and rhythmic structure of connected speech, our speech samples were long utterances of prosodic phrases, utterances and prosody groups in different durations, and were longer than utterances normally found in syntactic investigations. In this paper, we will report methods we devised to collect speech data of the following speaking styles: read speech by untrained native speakers, read speech by radio announcers, spontaneous speech of specific topics and without topic specification, spontaneous speech of public speaking and monologue. Speech data included microphone speech recorded in sound proof chambers and telephone speech. Text used included well structured paragraphs as well as word salads.

1 Introduction

In this paper, we report the design and recording of speech databases established at the phonetics laboratory, Institute of Linguistics, Academia Sinica at Taiwan. We will summarize the aims and designs first, followed by a summary of the speech corpora collected, and finally by a discussion and a short conclusion.

2 Designs of Speech Databases

Our laboratory has collected 7 read speech corpora and 1 spontaneous speech corpus during the period of 1995 till 2003, focusing on different aspect of speech prosody of Mandarin Chinese, and in particular, the overall characteristics of phrasing and paragraphing in speech

flow. Databases of read speech include: (1.) phonetically balanced speech database, 599 paragraphs of text; (2.) our exclusive design of text for MAT, and our share of data collection, namely, 40 sets in total, each set contains 408 isolated monosyllables, 1062 lexical words and 20 utterances; (3.) prosody balanced phrase groups, 1654 utterances of text; (4.) stress balanced text, 161 utterances of text; (5.) lexically balanced corpus for Mandarin spoken in Taiwan and in China, 1217 sentences and 26 paragraphs of text; (6.) Prosody-Group oriented database, 18 discourses or 77 prosody groups of text and (7.) word salad, 80 balanced phrases, 25 utterances and 2 paragraphs of text. The only spontaneous speech database focuses also on speech flow rather than dialogue or conversation, and therefore is narration and monologue in nature.

2.1. Phonetically-Balanced Speech Database

The purpose of this database is to obtain both phonetic and prosody information usually not found in canonical forms of lexical words or phrases under 10 syllables. Five factors were controlled in the design. They are: (1.) all of the possible syllables in Mandarin, under 1300 in total, (2.) most frequently used lexical words from 2 to 4 syllables, (3.) all possible segmental combinations and concatenations, and (4.) all possible tonal combinations and concatenations, and (5.) paragraph length, 1 to 180 characters in total. We selected by software the most frequently used 27,000 or so lexical items (words) from the CKIP [1] lexical database as a set, and then chose all possible segmental and tonal combinations within, then hand-tailored them into paragraphs. 599 paragraphs were composed [2]. The database was designed and collected in 1994. The text part has served as the basis for the MAT (Mandarin across Taiwan) [3] speech database project subsequently, as well as much of our own later works at the lab.

Instructions for subjects were to read into the microphone in normal speaking rate at sound proof chambers at our lab. Speech data from 3 male and 3 female speakers were collected. Each male-female pair was defined by age to denote three generations of speakers. One of the male speakers had been a radio announcer; the others were untrained native speakers.

2.2. MAT Speech Database

The purpose of this database was to collect a wide variety of Mandarin Chinese spoken across Taiwan via telephone. A joint project was proposed [3], the outcome was later known as various versions of MAT. The aim of the design was to cover the following three features: (1.) all of the possible syllables in Mandarin excluding tone information, 408 in total, (2.) most frequently used lexical words from 2 to 4 syllables, 1062 in total, (3.) all possible segmental combinations, 1400 in total and (4.) all possible tonal combinations, over 300 in total. In 1995, the text for the MAT project was generated at our lab from the 599 phonetically balanced paragraphs. 8 institutions joined the project and began recording in 1996, each site receiving 40 subsets from the text provided by our lab, including 408 isolated monosyllables, 1062 lexical words and 20 utterances ranging from 9 to 20 characters in length. Instructions for subjects were to telephone toll free numbers and read into the telephone mouthpiece.

Our lab was also one of the 8 institutions that jointly carried out collection of speech data during the period of 1996-1999. We collected speech data from a total of 160 speakers, all of whom were untrained volunteers. The collected speech data were part of the MAT speech data for later distribution.

2.3. Prosody-Balanced Phrase Groups Database

The purpose of this database was to examine the role of intonation with respect to prosody grouping in Mandarin Chinese. During our analysis of the Phonetically-Balanced Speech Database, we noticed that on the one hand, the 599 phonetically and tonally balanced paragraphs proved to be insufficient for prosody investigation in the corpus sense for lack of a somewhat balanced distribution among phrase/sentence types, namely, declarative, interrogative and exclamatory. We also noticed, on the other hand, that the speech data we collected showed a clear grouping of utterances into perceptually identifiable but larger-than-phrase/sentence prosody units in speech flow. We termed the phenomenon Prosody Group [4, 5] and subsequently designed another text to obtain phrase-type as well as prosody-adverbial/particle balance. That is, on the basis of three sentence types, i.e., declarative, interrogative and exclamatory, we included all possible adverbials and particles within each type to exhaust possible occurrences and to further investigate variations. Four factors were controlled, namely, (1.) utterance type, (2.), particles and adverbials, (3.) distribution of utterance type and (4.) utterance length. This speech database was our first database to concentrate on the grouping effect in Mandarin speech and its prosodic characteristics. The entire text was hand tailored to make the paragraphs as close to spoken form as possible, removing and editing occurrence of literary expressions. A total of 1654 phrase groups, including 805 declaratives, 546 interrogatives and 303 exclamatories, were generated, ranging from 5 to 134 characters each in length. Since declaratives do not involve special particles and adverbials, the selection was based on lexical balance again from the CKIP database. Note that the utterances in this database were not as long as the phonetically-balanced speech database, but they were still longer utterances nonetheless. The database was designed and collected in1997.

Instructions for subjects were to read out in comfortable and normal speaking rate at sound proof chambers at our lab. Speech data from 3 male and 4 female untrained speakers were collected. All of the speakers were untrained native speakers.

2.4. Stress-Balanced Speech Database

In order to obtain stress information for our studies of focus and prominence in speech flow and further understand the relationship between lexical stresses from utterance focus, we designed a stress balanced database and subsequently tagging systems for prominence as well [6, 7]. Three factors were controlled for stress balance, namely, (1.) stress type, (2.) cross-listener perceptual consistency and (3.) duration of lexical items. We chose from the above two texts lexical words ranging from 2 to 7 characters and balanced the stress distribution among the lexical items. Since the distribution of lexical words from the first two texts was insufficient to cover stress distribution, we also added more lexical word from the CKIP database and drew examples from the media for update. A total of 161 phrase groups were generated, ranging from 9 to 66 characters in length. The database was designed and collected in 2000.

Instructions for subjects were to read into the microphone in normal speaking rate at sound proof chambers at our lab. Speech data from 1 male and 1 female untrained native speaker each were collected.

2.5. Lexically-Balanced Speech Database

Although Mandarin Chinese is the official spoken language for both China and Taiwan, it is common knowledge among Chinese that many lexical items differ. By lexical balance here, we mean coverage and distribution of lexical words used in Taiwan Mandarin and Beijing Mandarin. This aspect is essential for development in Mandarin speech technology to obtain some systematic knowledge of the lexical difference as well as pronunciation variation. In an unofficial collaboration with Tsinghua University at Beijing, we exchanged text of recording materials to achieve lexical balance and recorded speech data respectively. The lexically-balanced speech database on our side included 217 phonetically balanced (9-20 characters) sentences, 26 paragraphs (85-982 characters) and 1000 relatively short sentences (16-25 characters). The 217 sentences were selected from the MAT text, originally constructed at our lab. The 26 paragraphs came from two sources. We hand compiled 23 paragraphs from our previous 599 phonetically balanced paragraphs and added 3 paragraphs constructed by Tsinghua University at Beijing. The 1000-sentence set was materials from Tsinghua University at Beijing. The speech data was collected in 2002.

Instructions for subjects were to read out in normal speaking rate. Speech data from 1 male and 1 female radio announcer, both under 35 years of age, were collected.

2.6. Prosody-Group Oriented Speech Database

The database was designed to further investigate the following phenomena; (1.) grouping of phrases and paragraphing in speech flow, (2.) boundaries and units involved, (3.) global planning and local specification of prosody units, and (4.) focus and prominence in speech flow, and (5.) interaction between syntax, semantics and prosody. We collapsed text from the above five texts and also transcriptions from our spontaneous speech data (See 2.8.) and came up with discourses ranging from 500 to 600 characters. Each 500-600-character discourse piece was punctuated into paragraphs ranging from 75 to 150 characters. The text for a total of 18 discourses containing a total 77 prosody groups and ranging from 347 to 712 characters was designed in 2003. Speech data was collected afterwards.

Speech data from 1 male and 1 female untrained native speaker each were collected. However, instructions for reading out the text differed. Three different readings of the same text were obtained in the following order. The first reading was identical from our previous read speech, namely, subjects were asked to read out the text in normal speaking rate. Then the subjects were asked to identify and hand mark on hard copies of the text the portions they intended to emphasize in each paragraphs. The subjects then read out the text according to their own emphasis specification for the second time. For the third reading, the subjects were given the same text with hand marked emphases, this time not by themselves by the research staff at our lab , and to read out the marked emphases accordingly.

2.7. Word Salad Speech Database

This database was to further test grouping effect in Mandarin running speech by removing syntactic and semantic information altogether. Simple in-house software was used to randomly pick out characters from our previous texts and coin them into utterances ranging from 10 to 60 characters. Four factors were controlled, i.e., (1.) tones, (2.) function words, (3.) word frequency and (4.) pronunciation. For tones, we controlled even distribution of the 4 Mandarin tones and 1 neutral tone for the paragraph-final syllable, and then controlled the second-to-last character for even distribution of the 4 tones one more time. The consideration

of tone control spanned for two syllables backwards was to include even distribution as well as to capture possible disyllabic effect at the word level. For function words, one control barred their occurrence at utterance initial position, and a second control went for their distribution within the text. The consideration was to see if function words would affect grouping towards the left or right in the speech flow in the linear sense, possibly marked by a pause. For this purpose, the consistency of word numbers which may have possible effect of the grouping as well as rhythm was also maintained. In other words, the number of characters between function words remained odd if the character string was odd in number before function word insertions; whereas the same principle applied to even number character strings as well. Less frequent characters were removed from the text reduce hesitation or confusion in the reading task. Characters with more than one pronunciation were also removed for the same reason. All of the text generated was subsequently hand tailored to remove any possible meaningful reading or proper names.

80 sections of word salad, 25 utterances and 2 paragraphs were generated. The 80 sections ranged from 10 to 60 characters, 40 of them were presented in the form of character strings without any punctuation marks; the other 40 with punctuation marks randomly assigned by software and ended in periods. The 25 utterances with punctuations ranged from 17 to 83 characters; the two paragraphs 393 and 461 characters respectively. Both the utterances and paragraphs were samples from our earlier databases. These utterances and paragraphs served as reference of the speakers' regular reading speech for the comparison purposes. The text was generated in the spring of 2003. Speech data from 1 male and 1 female untrained native speaker each were collected.

Instructions for subjects were to read out in comfortable and normal speaking rate at sound proof chambers at our lab. Speech data from 1 male and 1 female untrained native speaker each, both of whom under 30 years of age, were collected.

2.8. Comparable Spontaneous-and-Read Speech Database

The goal of our spontaneous-and-read speech data was to (1.) begin investigation of prosody of spontaneous speech, (2.) compare and derive prosodic characteristics for both read and spontaneous speaking styles, and (3.) study possible paralinguistic as well as non-linguistic effects on prosody manifestation. Our focus was also running speech and as a result relative longer monologues were collected.

Instructions for the subjects varied most for this speech database. We devised three phases of recording to achieve our goal. The first phase overlapped with data collection described in 2.7. That is, speakers were asked to read the word salad. This turned out to be a difficult job for the subjects, both for the content and for the recording processes involved. But as recording time increased, the speakers became familiar with the set-up and the sound proof chamber, and grew more experienced and relaxed for the recording sessions. At this point, we were ready to begin the second phase of data collection, namely, spontaneous speech. We set out to collect spontaneous speech on specific topics as well as free monologues. For specific topics, we chose the outbreak of SARS and related reports in the spring of 2003. Subjects were to asked read without sounding out a piece of SARS related text we provided for a period of 30 minutes, and subjects were allowed to take notes of the materials during reading. The text consisted of headlines and front page coverage on SARS. When the 30 minutes were up, subjects were asked to enter the sound proof chamber with their own notes and give an oral report of what they had just read into the microphone. We made a point to use head sets microphones (AKG C410) so that the subjects were not inhibited from hand gesturing and body movements while speaking. This turned out to be

much easier for untrained speakers to generate monologue in length. We then offered a coffee break, and the subjects were asked to record again free monologues. By this time our subjects were completely relaxed and offered a narration of some treasured personal experiences, with an experimenter at the side as the loyal listener, reciprocating with eye contacts and proper body language such as nodding. This concluded the second phase of data collection, during which we obtained two kinds of spontaneous speech. The third phase was days later, after we made orthographic transcription of the spontaneous speech obtained during the second phase of recording. The same speakers were asked to read out their own previous spontaneous speech. After the third phase, we would be able to obtain both spontaneous and read speech data of identical content for future phonetic as well as prosodic comparisons. The same speakers for the word salad speech database also served as subjects for the present comparable spontaneous-and-read speech database. This database was designed in the spring of 2003. So far we have collected 6 discourses.

3 Speech Data Collected

Tables 1 and 2 summarize the data collected.

	Phonetically Balanced	**MAT-**	**Prosody Balanced Phrase Groups**	**Stress-Balanced**	**Lexically Balanced**	**Prosody-Group Oriented**	**Word Salad**
Speakers	6 3 males 3 females	160 80 males 80 females	7 3 males 4 females	2 1 male 1 female	2 1 male 1 female	2 1 male 1 female	2 1 male 1 female
Recording time	18:38	Over 80 hrs	31:19	0:48	35:50	7:30	1:32
Recording Software	Lab developed	VCORDER	Lab developed	1.Pulsar 2.Triple DAT	Cool Edit 2000	Cool Edit 2000	Cool Edit 2000
Recording equipment	1. SONY MZ-R2 MD tape recording 2. beyer-dynamic M69N(C) mic 3. TDK MD tapes	1. Intel 486 or Pentium PC 2. 16 bits sound card 3. Dialogic D/41D phone card	1.SONY MZ-R2 MD tape recorder 2.AKG C410 headset mic 3.TDK MD tapes	1. dbs386 Tube Pre digital amplifier Creamw@re 2. Pulsar recording sound card 3. AKG C410 headset mic	On location at the radio station	1. dbs386 Tube Pre digital amplifier 2. Creamw@re Pulsar recording sound card 3. AKG C410 headset mic	1. HHB Portadisc MDP500 MD tape recorder 2. AKG C410 headset mic
Size of digitized data	2047.8MB	5507 MB	3441MB	244MB	568.3MB	2.06GB	626.7MB

	Phonetically Balanced	**MAT-**	**Prosody Balanced Phrase Groups**	**Stress-Balanced**	**Lexically Balanced**	**Prosody-Group Oriented**	**Word Salad**
Data format	*.wav	*.vat	*.wav	*.wav	*.wav	*.wav	*.wav
Type of data	Microphone speech	telephone speech	Microphone speech	Microphone speech	Microphone speech	Microphone speech	Microphone speech
Content and tally	599 paragraphs	1. Q & A 2. digit string 3. 408 mono-syllables 4. 1062 lexical words, 2-4 syllables 5. 400 utterances	805 declaratives 303 exclam-atories 546 inter-rogatives	161 stress balanced utterances	1. 217 utterances from MAT 2. 26 out of 599 phonetically balanced paragraphs 3. 1000 utterances from Beijing Tsinghua U	18 discourses (77 prosody groups in total)	80 sets of word salad 25 utterances 2 paragraphs
# of syllables/ utterance	1-180	9-20	5-134	9-66	1. 9-20/ utterance 2. 85-982/ paragraph 3. 16-25/ utterances	374-712	1. 10-60 2. 17-83 3. 393-461

Table 1. Databases of Read Speech

# of speakers	2, 1 male 1 female
Recording time	20m
Software	Cool Edit 2000
Equipment	1. AKG C410 headset microphone 2. HHB Portadisc MDP500 MD tape recording
Digitized Data	107.4MB
Data format	*.wav
Type of Data	Microphone speech
Content and Talley	6 discourses
Syllables	

Table 2. Database of Spontaneous Speech

4 Discussion

The major difference of the speech database collected by our lab lies in our interest to investigate the grouping-of-phrases effect in Mandarin speech, and our upcoming investigation of rhythmic structuring of speech flow. We used long paragraphs and discourses from the beginning, and have identified the perceptually identifiable prosody unit in Mandarin speech flow that is larger than single phrases or sentences [4]. We believe that understanding the prosodic group is essential to development of speech technology, in particular, application to speech synthesis in unlimited TTS in for bettering output naturalness. We devised a prosody tagging system in the spirit of ToBI that included this phenomenon [5, 6], and have also studied the physical, phonetic, and acoustic properties of other prosody phenomenon such as focus and prominence [7, 8] towards a possible organization of Mandarin speech prosody. We have emphasized on how speech flow is broken into various prosodic units separated by various lengths of silence in between, investigated these breaks in detail [9, 10, 11]. The above speech databases were a series of attempts to capture speech prosody from read to spontaneous speech. Our main goal was to take a step away from the traditional bottom-up canonical-form oriented approach to study phonetic as well as prosodic units and systems and their roles in prosody organization. As a result, our emphases were not segments or tones, nor words or phrases in the lexical sense, or even short (under 10 syllable) utterances. Along that particular thread of thinking, it is no surprise that we do not see speech flow merely as how to connect shorter units into longer ones. Rather, we have been looking at speech prosody from the perspective of language planning. For the time being, although the prosodic phenomenon presented, namely the grouping effect, appears more language specific to Mandarin Chinese, we believe that it may not be that language specific in nature since similar evidence could be easily observed in languages other than Chinese.

5 Conclusion

An organization of Mandarin speech prosody is inadequate if it only addresses short phrases and simple syntactic sentences as the largest speaking unit, treating longer utterances as simple concatenation of shorter units without a larger framework. Our colleagues in speech technology who develop speech synthesis for TTS systems have shown us just that over the years. A prosody organization or framework should not only account for the grouping or paragraphing phenomenon in speech flow, but also offer some explanation to the nature and origin as well. We believe that this grouping effect is largely related to speech planning, while interacting with syntactic and semantic levels of knowledge and information before and during speech production. And we speculate that semantics may be the governing factor of the planning. In short, we argue that unless we investigate speech prosody from this top-down perspective and look for a meeting ground with the findings from the bottom-up perspective, we would not be able to reach a more comprehensive understanding of speech prosody. Our presentation in this paper recalls the designs and rationales for speech databases catered to this orientation so far.

6 Acknowledgements

This paper summarizes part of three Academia Sinica Theme Research Projects, "Collaborating Researches on Chinese Information Processing" (1994-99), "Knowledge

Representation and Language Engineering for Mandarin Chinese" (1999-02) "New Directions for Mandarin Speech Synthesis: From Prosodic Organization to More Natural Output" (2003-05).

The first author wishes to thank Fu-chiang Chou, Jia-chuan Si, Wei-jia Lee, Da-deh Chen, Ru-chun Tseng and Chi-ching Chen for participating in the earlier part of the research from 1995 to 2000. Co-authors Yun-Ching Cheng and Wei-Shan Lee joined the research group in 2000, Feng-Lan Huang in 2001. Communications should be addressed to Chiu-yu Tseng.

7 References

1. http://godel.iis.sinica.edu.tw/CKIP/
2. Tseng, C., 1995. "A phonetically oriented speech database for Mandarin Chinese." ICPhS'95, Stockholm, Sweden, vol. 3, pp.326-329
3. MAT (Design and Implementation of Mandarin Speech Database), NSC project 85-2213-E-007-045, 1996-1999
4. Chiu-yu Tseng, "Prosodic Group: Suprasegmental Characteristics of Mandarin Connected Speech from a Speech Database". The Sixth International Conference on Chinese Linguistics(ICCL-6) (June 18-21, 1997), Leiden, the Netherlands.
5. Tseng, C. and F. Chou, "Machine readable phonetic transcription system for Chinese dialects spoken in Taiwan", *The Journal of the Acoustical Society of Japan (E)*, Vol.20, No.3, (May 1999): 215-223
6. Tseng, C. and F. Chou, "A Prosodic Labeling System for Mandarin Speech Database", XIVth International Congress of Phonetic Science, (Aug. 1-7, 1999), San Francisco, California.2379-2382
7. Tseng, C. "Focus and Prominence: More Investigation of Mandarin Prosodic Properties through Speech Database", The 9th International Conference on Chinese Linguistics, Jan. 26-28, 2000) , Singapore.
8. Tseng, C. "Some notes on Mandarin Prominence", Corpus and Computational Linguistics Symposium : Theory and Practice, Beijing, China.(May.26-28, 2001)
9. Tseng, C., "The Major Features of Prosody Organization of Speech Flow." Proceedings of The 6th National Conference on Man-Machine Speech Communication (NCMMSC-6), Xu, Ming-xing ed. Shenzhen, China (Nov.19-24, 2001), pp. 169-172
10. Tseng , C., "The Prosodic Status of Breaks in Running Speech: Examination and Evaluation", Speech Prosody 2002, Aix-en-Provence, France, April, 11-13, 2002) .667-670.
11. Tseng, C. and S. Bin, "Mandarin Prosody Organization and the Role of Silence in Speech Prosody", Joint International Conference of SNLP-Oriental COCOSDA 2002, Hua Hin, Prachuapkirikhan, Thailand, (May, 9-11, 2002) .324.