THE DESIGN OF PROSODICALLY ORIENTED MANDARIN SPEECH DATABASE

Fu-chiang Chou and Chiu-yu Tseng
Institute of Linguistics (Preparatory Office), Academia Sinica
Taipei, Taiwan R. O. C.
e-mail: moza@speech.ihp.sinica.edu.tw

ABSTRACT

While most of the current speech databases are designed based on the phonetically balance or phonetically rich principles, we proposed to design a prosodically oriented speech database for Mandarin. Based on the assumption that the intonation is determined by the sentence types, and the prosodic structure is closely related to the syntactic structure and patterns, we make our focus on the coverage of sentence types and phrasal syntactic patters. This paper describes the methodology for the design and collection of such a prosodically oriented speech database for Mandarin Chinese. There are 1600 short paragraphs selected to cover the three major sentence types, namely, interrogative, exclamatory, and declarative. The selected text scripts were recorded by six speakers and the recorded databases are served as the experimental environments for the research on Mandarin prosody.

1. INTRODUCTION

Most of the current speech databases are designed based on the phonetically balance or phonetically rich principles [1], which include our previous phonetically oriented Mandarin speech database [2]. The reason may be that most of the corpus is designed for speech recognition, and the research of prosody based on speech corpora is just started. Such kinds of databases are insufficient for the study of prosodic properties in both the basic research and applied technologies. Therefore, it is important to design a prosodically oriented speech database that covers the variability of prosodic phenomena in speech. We will describe the methodology for the design and collection of such a speech database in this paper.

As mentioned above, there are two main properties should be considered in the design of speech database, the phonetic property and the prosodic property. For the concern of phonetic property, to cover most of the contextual segmental variants can provide the sufficient units for the training of speech recognizers or the unit inventories of speech synthesizers. We call this requirement the "phonetically rich". For the design of such a phonetically rich corpus, the text scripts can be automatically selected from a large text corpus to meet some phonetic requirements. For Mandarin, the requirement is to include various intra-syllabic and inter-syllabic allophone structures. This property is the basic requirement for many applications such as the speech recognition and speech synthesis. Many similar approaches can be found in the literatures [3][4][5]. For the concern of prosodic property, to cover most of prosodic contexts can provide the sufficient information for the research of prosodic phenomena and the training of prosodic models for TTS systems. We call this requirement the "prosodically rich". For the design of such a prosodically rich corpus, the text scripts can also be automatically selected from a large text corpus to meet some requirements like the design of phonetically rich corpus. The problem is what are the requirements for a prosodically rich corpus. Compared with the problem of phonetic properties, there are fewer literatures talking about the prosodic properties, especially for Mandarin. Based on the assumption that prosodic structure is closely related to the syntactic structure of a sentence [6][7], we first make our focus on the sentence types and phrasal syntactic patterns.

The choice of sentence types for the selection of prosodic corpus is involved with some basic problems. The patterns of Mandarin intonation have not yet been well established. Chao gave four forms of intonation in [8] and gave thirteen types of intonation according to different speaker attitudes in [9]. Ho had studied the intonation of Mandarin Chinese according to the three modalities of utterances [10], interrogative, exclamatory, and declarative. However, most of the other Chinese linguists only mentioned the difficulties of investigating Mandarin intonation. In our case, the three modalities in Ho are chosen as the three major sentence types, and each major type consists of many sub-types.

Besides the sentence types that determine the intonation of a sentence, the other factors that determine the prosodic structure should also be considered. This is a more complex problem, because the factors maybe involved in the syntactic, semantic or the pragmatic levels. The only factor that we can consider in this stage with some automatic ways is the syntactic properties. We choose the part-of-speech (POS) as the major syntactic factor that should be considered in this stage. Therefore, we can use a similar selection program that used in the design of phonetically rich speech database. The only modification is the requirements are replaced to cover most of the POS patterns. Similar approach could be found in [11], which used a genetic algorithm to design the learning database for a task dependent text-to-speech synthesizer. Because the task is a yellow page service, it is designed to cover all the combinations of the three fields, the first name, the last name and the telephone number. In our case that is not designed for a specific task, we chose to cover the POS bigram and trigram instead.

The paper is organized as followed. The selection 2 describes the three sentence types and their sub-types. The section 3 describes the algorithm that design to select the sentences to cover most of the POS patters. The section 4 describes the recording and labeling of the designed corpus. At last are the concluding remarks.

2. SELECTION OF TEXT SCRIPTS

The selection of interrogative sentences is based on the classification in [12]. There are four types of interrogative sentences: interrogative-word question, particle question, disjunctive question, and tag question (A not A question). They are defined as followed:

(a) interrogative-word question:
The interrogative-words in Mandarin are the semantic equivalents of such English words as who, what, where, which, and so forth.
(b) disjunctive question
The disjunctive question is presented with a choice between two options. A simple example is composed by two declarative sentence joined by "or".
(c) tag question
The tag question is composed of a statement followed by an "A not A" form.
(d) particle question
The particle question is a question that signaled by the presence of the question particle in sentence-final position.

In general, these four types of interrogative sentences are determined by the particles, conjunctions, and interrogative words consisted in the sentences.

The exclamatory sentences are classified into three sub-types: with interrogative adverbs, with interrogative particles, and with exclamatory particles. In the same way, these types are determined with the particles and adverbs in the sentences.

The sentence types of declarative sentences are more complicated than interrogative and exclamatory sentences, and there are no standard classification methods in literature. Therefore, we try to select the declarative sentences according to their phrase patterns but not their entire sentence structures. The phrase patterns are defined as the combinations of POSs or punctuation marks. Consequently, we tried to increase the bigram and trigram coverage of the sequence of POSs and punctuation marks. The process will be described later.

The Academia Sinica Balance Corpus (Sinica corpus) [13] was used for the selection of the text scripts for recording the speech database. There are totally 280,000 sentences with manual tags. The tags include 46 POSs and 10 punctuation marks. The 8350 sentences ending with question marks and the 5471 sentences ending with exclamation marks was at first extracted for the selection of interrogative sentences and exclamatory sentences. After the manual exclusion of incomplete sentences, there are 4650 interrogative sentences and 3947 exclamatory sentences left. These sentences are automatically classified with the particles and interrogative words, and then manually tailored for relatively even distribution. Totally 550 interrogative sentences and 300 exclamatory sentences are selected in this stage. The distribution of the sentence types in the corpus is listed in Table 1.

Interrogative sentence	number	ratio	Exclamatory sentence	number	ratio
Interrogative-word	948	20.4%	fragment	323	8.2%
Disjunctive	445	9.6%	Interrogative adverbs	132	3.3%
Tag	23	0.5%	Interrogative particles	348	8.8%
Particle	2516	54.1%	Exclamatory particles	1694	42.9%
Other	718	15.4%	Other	1450	36.7%
Total	4650	100.0%	Total	3947	100.0%

Table 1. The distribution of the sub-sentence types in interrogative sentences and exclamatory sentences.

The declarative sentences were selected after the selection of interrogative and exclamatory sentences. In this stage, the selection process is designed to increase the bigram and trigram coverage of POSs and punctuation marks. The bigram and trigram coverage of POSs and punctuation marks for the selected interrogative and exclamatory sentences ware calculated at first. An interactive program was designed to further select the declarative sentences. The distributions of the bigram and trigram are first calculated and stored in a table. The score for each bigram or trigram is defined according to the number occurred in the whole corpus. The more frequently occurred items have lower scores. If the bigram or trigram has already been selected or the number is smaller than ten, the score is set to zero. The score for a sentence was defined as the summation of the scores for the bigrams and trigrams in that sentence. The sentence with highest score in the whole corpus will be shown on the screen. The operator has to decide to keep it or not according to the continuity and completeness of the sentence. If the sentence is selected, the score table will be updated and the process is iterated. If the sentence was not selected, it will be dropped and the sentence with inferior score will be shown instead. There were 800 sentences selected with this procedure. The statistics of the bigrams and trigrams of the whole corpus and the selected corpus are list in Table 2.

	Whole corpus	Whole corpus (No>=10)	Interrogative & exclamatory sentence	Whole selected corpus
No of Bigram	2715	2059	1101	1746
No of Trigram	46092	14595	4170	9394

Table 2. The bigram and trigram coverage of the selected corpus

3. RECORDING AND LABELING OF THE CORPUS

A program is designed for the recording of the speech corpus. The text scripts will be shown on the monitor and the recording will be activated when the speaker starting to read the sentence. Six speakers with three males and three females separately recorded the whole set of text scripts. The recordings took place at a sound proof studio and were supervised with an experienced assistant. The speakers were requested to read the scripts at a normal speaking rate. Moreover, if there were hesitations or mistakes, the speaker will be asked to read the sentence again until the whole sentence is clearly pronounced. This can reduce the errors in the further labeling procedures. The recording equipment includes AKG C410 condenser microphone and Ariel Proport 656 A/D converter. The recording time is about 7-8 hours, and the storage is about 450 Mbytes per speaker. The actual time and size of the speech database for the six speakers are listed in Table 3.

Speaker	Time (min)	Storage (Mb)
M1	234	429
M2	258	473
M3	283	518
F1	263	481
F2	254	465
F3	313	573

Table 3. The time and storage of the speech database for the six speakers

The most tedious and important work for the design of a speech database is the labeling. Many speech databases have only orthographic transcription, which is not enough for many applications, especially for the research of prosody. Therefore, we tried to include both the segmental and prosodic labels in this speech database. We have designed a machine readable transcription system in both the segmental and prosodic domains [14]. The details of the prosodic labeling system and some initial experiments can be found in another paper of the same proceeding [15]. However, the work of manual labeling is very time-consuming. It is hard to label the whole speech database with manual efforts. Therefore, we have designed some programs to automatically label the speech database in both the segmental and prosodic domains [16]. With these automatic programs, the human processing time can be largely saved.

4. CONCLUDING REMARK

This is the first attempt to design a prosodically oriented speech database for Mandarin. Based on the assumption that the intonation is determined by the sentence types, and the prosodic structure is closely related to the syntactic structure, we designed the procedures and programs to select the text scripts. The speech database is recorded with these designed text scripts and labeled with a newly developed prosodic labeling system. This database is at first used for the training of a corpus-based TTS system for Chinese [17]. It was proved that such a prosodically oriented corpus can largely increase the naturalness of the synthetic speech compared with the original phonetically oriented corpus. The next step is to use the speech database for the investigation of the relation between syntax, semantics and prosody. After that, more factors can be added for the design of another prosodic oriented speech database.

REFERENCES

[1] Zue, V. , Stephanie, S. and James, G. 1990. Speech database development at MIT: Timit and beyond. *Speech Communication,* Vol.9. , No.4 , 351-356.

[2] Tseng, C. Y. 1995. A phonetically oriented speech database for Mandarin Chinese. *Proceedings of the 13th International Congress on Phonetic Sciences,* 326-329.

[3] Akira, K., Kazuya, T., Yoshinori, S., Shigeru, K., Hisao, K. and Kiyohiro, S. 1990. ATR Japanese speech database as a tool of speech recognition and synthesis. *Speech Communication*, Vol. 9, No. 4, 357-363

[4] Wang, H. M., Chang, Y. C. and Lee, L. S. 1994. An algorithm for automatically selecting phonetically balanced sentences from a large corpus for training and testing a speech recognition system. *Proceeding of the International Conference on Computer Processing of Oriental Language,* 507-510.

[5] Zu, Y. Q. 1997. Sentence design for speech synthesis and speech recognition database by phonetic rules. *Proceeding of the 5th European Conference on Speech Communication and Technology*, 743-746

[6] Selkirk, E. O. 1984. *Phonology and syntax: the relation between sound and structure.* MIT Press

[7] Price, P., Ostendorf, M., Shattuck-Hufnagel, S. and Fong, C. 1991. The use of prosody in syntactic disambiguation, *J. Acoust. Soc. Am. ,* 90, 2956-2970

[8] Chao, Y. R. 1933. Tone and intonation in Chinese. *Bulletin of the Institute of History and Philology*, 4, 2121-2134

[9] Chao, Y. R. 1968. *A grammar of spoken Chinese*, University of California Press

[10] Ho, A. T. 1977. Intonation variations in a Mandarin sentence for three expressions: interrogative, exclamatory, and declarative. *Phonetica*, 34, 446-456.

[11] Boeffard, O. and Emerard, F. 1997. Application-dependent prosodic models for Text-to-Speech synthesis and automatic design of learning database corpus using Genetic Algorithm. Proceeding of the 5th European Conference on Speech Communication and Technology, 2507-2510

[12] Li, C. N. and Thompson, S. A. 1981. *Mandarin Chinese: A Functional Reference Gramma*. University of California Press, 520-563.

[13] Academia Sinica Balance Corpus, Internet WWW page, at URL: <http://rocling.iis.sinica.edu.tw/CKIP/balance.htm>

[14] Tseng, C. Y. and Chou, F. C. 1998. Machine readable phonetic transcription system for Chinese dialects spoken in Taiwan. *Proceedings of the 1st Oriental COCOSDA Workshop*, 179-183

[15] Tseng, C. Y. and Chou, F. C. 1999. A prosodic labeling system for Mandarin speech database. *Proceedings of the 14th International Congress on Phonetic Sciences*

[16] Chou, F. C., Tseng, C. Y. and Lee, L. S. 1998. Automatic segmental and prosodic labeling of Mandarin speech database. *Proceedings of the 5th International Conference on Spoken Language Processing,*

[17] Chou, F. C. 1999 *Corpus-based technologies for Chinese text-to-speech synthesis.* PhD thesis, National Taiwan University