FOUR POINT SIX
(4) Sinica COSPRO—Corpus and Tools for Mandarin Fluent Speech Prosody

Chiu-yu Tseng

1 Introduction

This section reports the design and content of Sinica COSPRO (Mandarin Continuous Speech Prosody Corpora http://www.myet.com/COSPRO) and a Toolkit for prosody analysis, providing the first corpora of Mandarin fluent continuous speech. It includes eight sets of read speech and one set of spontaneous speech, all of which was recorded with a microphone at a sampling rate of 16,000 Hz in a 1-channel, 16-bit linear format, in soundproof rooms at the Phonetics Lab, Institute of Linguistics, Academia Sinica. A total of 7.7 GB of the corpus was manually tagged for multiple-phrase speech paragraphs and discourse prosody units by perceived boundaries (*break), and checked for both intra- and inter-transcriber consistency[1]. Table 1 summarizes the recording setups and data collected.

2 Design of Speech Databases

The major rationales of the COSPRO design and annotation are the explicit information structure of discourse prosody. Hence, both speech signals and boundary pause are treated as interacting discourse units. Implicit in the design are the psycho-, physio- and cognitive-linguistic contributions and interactions for output prosody.

2.1 COSPRO 01: Phonetically Balanced Speech Database

This database was designed for phonetic variations in continuous speech, which differ from single syllables, lexical words, or short simple sentences illicited in isloation. The database consists of 1455 phonotactically possible Mandarin tonal syllables and their combinations of two, three, and four syllables, and 599 discourse pieces ranging from 1 to 180 syllables[2], using the most frequent 27,000 lexical words from CKIP[3]. The

same text was also used later by the MAT (Mandarin across Taiwan)[4] speech database project.

TABLE 1 NINE SETS OF CORPORA IN SINICA CONPRO, LISTED WITH THEIR RESPECTIVE RECORDING TIME, RECORDING SETUP, EQUIPMENT USED, AND DIGITIZED DATA SIZE.

Corpus number	Speakers (male / female)	Recording time (hr:min)	Software	Recording equipment	Data size (GB)
01	6 (3/3)	18:38	In-house	1. SONY MZ-R2 MD tape recorder 2. Beyer dynamic M69N(C) microphone 3. TDK MD tapes	1.99
02	90 (40/50)	19:29	In-house	1. SONY MZ-R2 MD tape recorder 2. Beyer dynamic M69N(C) microphone 3. TDK MD tapes	2.08
03	7 (3/4)	31:19	In-house	1. SONY MZ-R2 MD tape recorder 2. Beyer dynamic M69N(C) microphone 3. TDK MD tapes	2.38
04	2 (1/1)	0:48	Pulsar	1. dbs386 Tube Pre digital amplifier 2. Creamw@re Pulsar recording sound card 3. AKG C410 headset microphone	0.243
05	2 (1/1)	35:50	Cool Edit 2000	On location at a radio station	0.574
06	2 (1/1)	7:30	Cool Edit 2000	1. dbs386 Tube Pre digital amplifier 2. Creamw@re Pulsar recording sound card 3. AKG C410 headset microphone	0.759
07	2 (1/1)	1:32	Cool Edit 2000	1. HHB Portadisc MDP500 MD tape recorder 2. AKG C410 headset microphone	0.577
08	2 (1/1)	16:50	Cool Edit 2000	1. HHB Portadisc MDP500 MD tape recorder 2. SONY ECM77B mini microphone	1.9
09	2 (1/1)	0:20	Cool Edit 2000	1. HHB Portadisc MDP500 MD tape recorder 2. AKG C410 headset microphone	0.107

Speech data from 6 native speakers, three males and three females from three age groups, namely, 25 and under, over 35, and over 50, was collected in 1996, This database provides quantities of speech from each speaker and is useful for both intra- and inter-speaker variation.

2.2 COSPOR 02: Multiple Speaker Speech Database

This database was designed for segmental features and speaker variation for speech recognition research. The text was designed to elicit maximum coverage of cross-speaker segmental information. Ten sets of lexical words, short sentences and paragraphs were constructed. Each set consists of 83 lexical words, 100 short sentences and 5 paragraphs. The lexical words and paragraphs were randomly selected from COSPRO 01, the short sentences were from COSPRO 03. Each set was read by four native male speakers and five female speakers. Speech data from a total of 90 speakers were collected in 1996 to provides a wide range of cross-speaker segmental and prosody variations from words, sentences and discourse pieces.

2.3 COSPRO 03: Intonation-Balanced Speech Database

This database was designed to collect phrase grouping and intonation variation in continuous Mandarin speech. Four factors were balanced: sentence type, distribution by sentence type, particles and adverbials, and sentence length. Three sentence types

were included, namely, the declarative, interrogative, and exclamatory sentences, all of which were based on lexical balance from the CKIP text database and balanced for adverbials and particles. A total of 1654 sentences were generated, with 805 declarative, 546 interrogative, and 303 exclamatory in structure, ranging from 5 to 134 characters in length. Speech data from 7 untrained speakers, three males and four female, was collected in 1997. This speech database provides a wide range of intonation variations over various sentence types, both within and across speakers.

2.4 COSPRO 04: Stress-Pattern-Balanced Speech Database

This database was designed to illicit lexical stress, sentence focus, and speaker-intended prominence (pitch accent) in continuous speech. The distribution of lexical stress was balanced, while the focus and prominence were elicited by specificaitons[5]. Lexical words of 2 to 7 characters were chosen from the texts of COSPRO 01, COSPRO03 and the CKIP database. A total of 161 short paragraphs were generated, ranging from 9 to 66 characters. Speech data from one male and one female untrained speaker was collected in 2000.

2.5 COSPRO 05: Lexically Balanced Speech Database

This database was designed for the distribution balance of frequently used lexical items from both Taiwan Mandarin and Putonghua. A total of 217 phonetically balanced sentences (9-20 characters), 26 paragraphs (85-982 characters), and 1000 relatively short sentences (16-25 characters) were constructed to cover a wide range of frequently used words in both dialects. Speech data from two radio announcers, one male and one female, was collected in 2002.

2.6 COSPRO 06: Focus-Balanced Prosody Group Speech Database

The database was designed for global prosody of larger discourse units, discourse focus and prominence, and interaction of discourse prosody with syntax and semantics. By combining text used for COSPRO 01 to 05 and text transcriptions of spontaneous speech data (subsection 2.8.), a total of 18 discourses of 347 to 712 characters were constructed. Manual punctuation of the 18 discourses resulted in 77 paragraphs ranging from 75 to 150 characters. Three readings were specified, namely, normal relaxed reading, with self-selected emphasis, and with marked emphasis specified by the research team. The speech data was collected in 2003 from two untrained native speakers, one male and one female.

2.7 COSPRO 07: Speech Database Varying in Text Type and Speaking Style

This database was designed to remove both syntactic and semantic information. 80 sets of word salad from 10 to 60 characters were generated from random characters of previously constructed text pieces. 40 sets werer presented without punctuation, while

the other 40 were presented with randomly assigned punctuation marks. In addition, 25 meaningful utterances from 17 to 83 characters and 2 meaningful paragraphs from 393 to 461 characters were also generated. Four factors were balanced: distribution of tones, word frequency, function words by utterance position, and homonyms. Speech data from two untrained native speakers, one male and one female, was collected in 2003.

2.8 COSPRO 08: Prosody-Balanced Monosyllable Database

This database was designed to elicit Mandarin monosyllables for syllable-based synthesis. A 30-character, 3-phrase complex sentence was designed as a carrier sentence, in which the 1455 Mandarin syllables were embedded at the sentence-initial, -medial, and -final positions to elicit the characteristics of global prosody with tones. Speech data from two untrained native speakers, one male and one female, was collected in 2004.

2.9 COSPRO 09: Comparable Spontaneous/Read Speech Database

This database was designed to compare the prosodic characteristics of read and spontaneous narratives of identical materials performed by the same speakers. The spontaneous speech was elicited as follows: The speakers were provided with written materials to study for 30 minutes, during which time they were allowed to take notes. After a short coffee break, the speakers were then asked to give an impromptu oral report on the materials by using their notes. The spontaneous speech was then transcribed into text. Days later, the same speakers returned and read the text transcriptions of their own spontaneous speech. A total of 6 discourses were obtained. The speech data was collected in 2003 from the same two speakers used for COSPRO 07.

3 Design of Annotation

The annotation is designed to identify phrase grouping and discourse information by perceived boundaries in continuous speech, instead of identifying intonation units independently. A hierarchical framework, called the Hierarchy of Prosodic Phrase Groups (HPG), specifies the discourse prosody organization and layered prosodic units, consisting of the syllable (SYL), prosodic word (PW), prosodic phrase (PPh), breath group (BG), and multiple phrase group (PG). The corresponding discourse boundaries are B1, B2, B3, B4, and B5, whereby SYL<PW<PPh<BG<PG and B1<B2<B3<B4<B5[6].

4 Conclusions

Sinica COSPRO and Toolkit has facilitated corpus-based phonetic research on Mandarin discourse prosody through quantitative accounts,and computational modeling[7]

References

1. Tseng, C. and Chou, F. Machine readable phonetic transcription system for Chinese dialects spoken in Taiwan, *The Journal of the Acoustical Society of Japan (E)*, 20, 3, pp.215-223. 1999
2. http://godel.iis.sinica.edu.tw/CKIP/
3. Tseng, C. A phonetically oriented speech database for Mandarin Chinese, *Proc. the 13th International Congress of Phonetic Science 3*, pp.326-329. 1995.
4. Wang, H., Seide, F., Tseng, C. and Lee, L. MAT-2000: Design, collection and validation of a Mandarin 2000-speaker telephone speech database, *Proc. ICSLP.* 2000
5. Tseng, C. Focus and prominence: more investigation of Mandarin prosodic properties through speech database, *Proc. the 9th International Conference on Chinese Linguistics*. 2000
6. Tseng, C., Pin, S., Lee, Y., Wang, H. and Chen, Y. Fluent speech prosody: framework and modeling *Speech Communication*, 46, 3-4, pp.284-309. 2005
7. http://phslab.ling.sinica.edu.tw/.

FOUR POINT SEVEN

Multilingual Corpora

(1) Multilingual Telephony Speech Corpora of Indian Languages

Samudravijaya K

1 Introduction

India is a multilingual country. There are 22 officially recognized (scheduled) languages and 234 mother tongues, each of which spoken by at least 10,000 speakers[1]. Machine recognition of spoken languages is very relevant since a significant fraction of Indians are uncomfortable with English-oriented input devices. Automatic speech recognition (ASR) technology will enable such people to gain access to information by just speaking into a microphone. Getting information on an "anytime, anywhere" basis is now possible by integrating ASR technology with the rapidly expanding mobile telephone systems in India. Most prevalent ASR systems follow statistical pattern recognition techniques and hence need a large amount of annotated speech data. In this chapter, we present a summary of telephony speech corpora in various Indian languages.

The majority of telephony speech corpora of Indian languages are either coordinated by an organization or carried out in-house by an institution. Accordingly, the information about speech corpora is presented under headings of organizations.

The speech data of all of the corpora were digitized at 8000Hz. The various corpora differ in characteristics such as text (isolated word, sentence, conversation), storage format (coding: mu-law, 16-bit PCM; header: SPHERE, wav), intended application (ASR, speaker/language/accent recognition), transcription (no/some transcription, annotation at lexical level, speech disfluencies noted), location of data collection (India, abroad), etc.

2 LDC

The Linguistic Data Consortium (LDC)[2] distributes speech corpora in many languages, including Indian languages. LDC distributes the following telephony speech corpora for two Indian languages: Hindi (Indo-Aryan family) and Tamil (Dravidian family). All telephone calls in these corpora were placed inside North America and all of the participants were native speakers.

The OGI Multilanguage Corpus (LDC94S17) is the earliest Indian language speech corpus made available for public usage. The corpus, meant for ASR applications, consists of responses to prompts spoken over telephone lines and contains about 175 calls per language. The speech data were recorded at 16bit PCM and are compressed, with NIST SPHERE headers.