CASS: A PHONETICALLY TRANSCRIBED CORPUS OF MANDARIN SPONTANEOUS SPEECH [1]

LI Aijun (1), ZHENG Fang (2), William Byrne (3), Pascale Fung (4), Terri Kamm (3), LIU Yi (4), SONG Zhanjiang (2),
Umar Ruhi (5), Veera Venkataramani (3), CHEN XiaoXia (1)

Chinese Academy of Social Sciences, Beijing, China (1)
Tsinghua University, Beijing, China (2)
Johns Hopkins University, Baltimore, MD 21218 USA (3)
University of Science and Technology, Hong Kong (HKUST) (4)
University of Toronto, Canada (5)

1. INTRODUCTION

A collection of Chinese spoken language has been collected and phonetically annotated to capture spontaneous speech and language effects. The Chinese Annotated Spontaneous Speech (CASS) corpus contains phonetically transcribed spontaneous speech. This corpus was created to begin to collect samples of most of the phonetic variations in Mandarin spontaneous speech due to pronunciation effects, including allophonic changes, phoneme reduction, phoneme deletion and insertion, as well as duration changes. It is intended for use in pronunciation modeling for improved automatic speech recognition and will be used at the 2000 Johns Hopkins University Language Engineering Workshop by the project on Pronunciation Modeling of Mandarin Casual Speech.

2. CORPUS INFORMATION

The speech in the CASS corpus was provided by the Broadcast Station of Tsinghua University (BSTHU), Beijing, China, from their audio archives. The recordings are of university lectures by professors and invited speakers, student colloquia, and other public meetings. The collection consists primarily of impromptu addresses, delivered in an informal style without prompting or written aids. As such the collection is a rich source of spontaneous speech phenomena and is well suited for pronunciation modeling. The recordings were made in ordinary classrooms, amphitheatres, or school studios without the benefit of high quality tape recorders or microphones. As a result the recordings are of uneven quality and contain significant background noise.

The archived recordings were delivered on audio cassettes. These were digitized using a Sound Blaster audio card into single-channel audio files, sampled at 16KHz and at 16-bit precision. An initial review of the data yielded approximately 6 hours of speech judged to be spontaneous, unaccented Mandarin. These were segmented into short utterances of between 2.4 and 4.0 seconds using the XWaves waveform-editing program. From this initial sampling a 3.0-hour subset of relatively clear and noise-free speech was chosen for detailed annotation.

The detailed annotation consists of transcriptions at the word level, syllable level, and semi-syllable level with detailed pronunciation variants. Phoneticians from the CASS Dialect Department reviewed the speech to determine each speaker’s origin. All speakers have lived for several decades in Beijing, a Mandarin speaking city. While most speakers retained some accent due to their native dialect, the influence was judged not to be significant for all speakers but one.

Due to the spontaneous nature of the speech and the difficult acoustic conditions, the transcription process was very time consuming. After an initial word transcription of the data, the complete annotation of one hour of raw speech required about 380 hours of effort by a single transcriber.

Table 1. Corpus Characteristics

Speaker ID and Gender	Speech rate (syllables/sec.)	Dialect Background
F03	3.89	Wu
F04	4.68	Wu
M01	5.31	HuNan or JiangXi
M02	4.94	AnHui or HuBei
M04	4.99	Shanxi/AnHui
M05	4.14	Northwestern
M12	4.06	North Wu

[1] This work was supported by the National Science Foundation under Grant No. #IIS-9820687, and carried out for the 2000 Workshop on Language Engineering, Center for Language and Speech Processing, Johns Hopkins University. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or The Johns Hopkins University.

Details about the speakers in the corpus are given in Table 1. Except for the female speaker F03, all the speakers were very fast, where slow speech is 170-200 syllables per minute, moderate speed is about 230 syllables per minute, and more than 250 syllables per minute is considered fast [2].

3. LABELING PRINCIPLES

The corpus was designed according to the following principles:

(1) Use of the machine-readable phonetic alphabet, SAMPA-C [3,5]

(2) Accurate transcription of phonetic variability and spoken phenomena

(3) High transcriber consistency

The speech was transcribed in a *three-tiered* annotation. The three tiers are the syllable tier, the semi-syllable tier, and the miscellaneous tier. In the syllable tier, *pinyin* and *tone* of each syllable is transcribed orthographically, i.e. based on the standard pronunciation of the word transcription. In the semi-syllable tier, the *initial* and *final* of each syllable is labeled using SAMPA-C. Segmentation boundaries are also provided in the semi-syllable tier. Sound variability such as phoneme change, insertion and deletion are also transcribed on the semi-syllable tier. Tones after tone sandhi, or tonal variation, are attached to the finals. In the miscellaneous tier, phenomena of spoken discourse, such as coughing, laughing, and mouth noises, are transcribed.

3.1 SAMPA-C: A Chinese Segmental Labeling Convention

The transcriptions in the semi-syllable tier use the SAMPA-C labeling convention. SAMPA-C is a labeling set of machine-readable IPA symbols adapted for Chinese languages from the SAMPA Speech Assessment Methods Phonetic Alphabet [4]. It includes canonical symbols for consonants and vowels, the initials and finals, the retroflexed finals, and in addition has labels for tones, sound variability, and spontaneous phenomena. A detailed description of SAMPA-C is given elsewhere in these proceedings [3].

3.2 Transcriber Consistency

To assess the agreement between annotators, 15 minutes of speech was transcribed in common by all four transcribers and their agreement was measured. The consistency of their transcriptions was measured in terms of number of transcriber pairs agreeing on the labeling of each particular segment; in these measurements, tonal information is discarded.

The average transcriber agreement was found to be 84.23% in the semi-syllable tier when agreement on silence labels (including the closure segments before some initials) was counted. The average transcriber agreement at this Pinyin level is 86.12%, counting silences. Comparing Pinyin and semi-syllable agreement simultaneously, the transcriber agreement is 85.04%. The agreement was observed to be 88.92%, 85.88% and 87.14% without counting the silence labels for the above three situations.

The agreement in the semi-syllable tier and Pinyin tier, are reported individually and jointly in Table 2. Two measurements of agreement are given for each pair of transcribers. The number to the left of '/' is the percentage of labels in agreement counting silence labels, while the number to the right indicates percent agreement disregarding silence.

Table 2. Transcriber Consistency Measurements.

Transcriber Pairs	SAMPA-C	Pinyin	Pinyin+ SAMPA-C
A-B	84.10 / 85.72	85.36 / 88.18	84.64 / 86.73
A-C	84.88 / 86.12	86.98 / 89.36	85.77 / 87.45
A-D	82.39 / 83.49	84.96 / 87.08	83.49 / 84.97
B-C	86.07 / 88.25	87.34 / 90.77	86.61 / 89.29
B-D	82.99 / 85.04	84.85 / 88.01	83.78 / 86.27
C-D	84.97 / 86.72	87.31 / 90.19	85.97 / 88.16

4. SOUND VARIABILITY

There are ten types of sound variability annotated in the semi-syllabic tier: Insertion, Deletion, Pharyngrealization, Voiced, Voiceless, Nasalization, Rounding, Aspiration, Centralization and Phoneme Change.

Rounding and aspiration are marked if either phenomenon is present to a degree judged more than usual.

Table 3. Transcription of modal sound changes.

Preceding final (IPA)	Character transcription of '啊'	Pinyin transcription of '啊'	SAMPA-C
[-i,-y,-a]	呀	ya0	ia_”
[-n]	哪	na0	na_”
[-u]	哇	wa0	ua_”
[-o,-°,-Q]	呀	ya0	ia_”
[-Ð]	啊	nga0	Na_”
[¡]	啊	za0	tsa_”
[Ÿ]	啊	ra0	ra_”

Retroflex finals and modal sound changes “啊 (a0)” are two special sound changes in spoken Chinese. Retroflexed syllables are labeled in the PinYin tier, even though they should more properly be described as sound changes, i.e. marked in the semi-syllable level. In this corpus they are marked by an expanded inventory of finals that contains a retroflexed version of each standard final.

As presented in Table 3, the orthographic text “啊”, can be pronounced as “a0”, “ya0”, “wa0”, “na0”, “ra0”, “nga0’’ or “za” depending on the final of the preceding syllable[1]. When “啊” is pronounced as ya0 or wa0, it is labeled as “ya0” or “wa0” on Pinyin tier since there is Chinese character corresponding to it; the SAMPA-C pronunciation follows from the standard character pronunciation. When “啊” is pronounced as “ra0” or “nga0”, it is labeled as “a0” on the Pinyin tier and either “r” or “N” appears as an inserted symbol in the SAMPA-C tier.

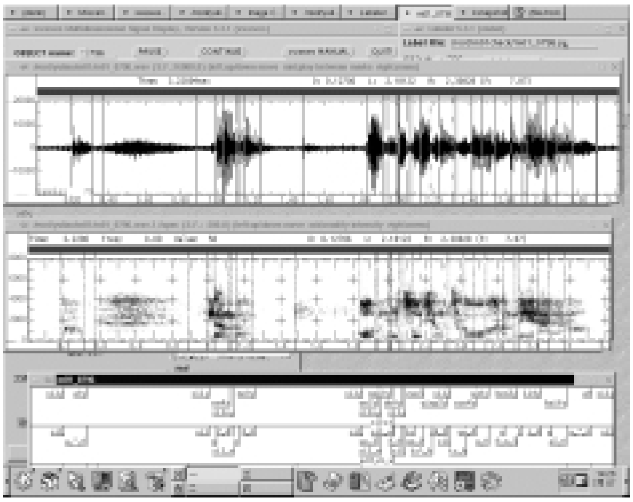

Fig 1. The modal “啊 a0”, unchanged after vowel “I” in sentence “na4 me0 wo3 men0 de0 zuo3 qing3 cuo4 wu4 hen3 li4 hai4 a0”.

Descriptions of pronunciation variability in Chinese differentiate sound changes which occur regularly in specified contexts, from free sound changes which are not governed by any pronunciation rules. In the first case, proper pronunciation requires sounds to change in certain contexts (first two rows in Table 3). For example, “啊 a0” should change to “na” following a nasal coda “n”; after a “i”, “a”, or “y”, however, “啊 a0” should be pronounced as “ya0”. It is interesting to note that in this corpus the sound changes of “啊 a0” do not follow convention. In Figure 1 “啊 a0” remains unchanged despite the preceding vowel “i”. This was the case throughout the corpus: many instances were found of “啊 a0” not changing according to the prescribed sound change rules.

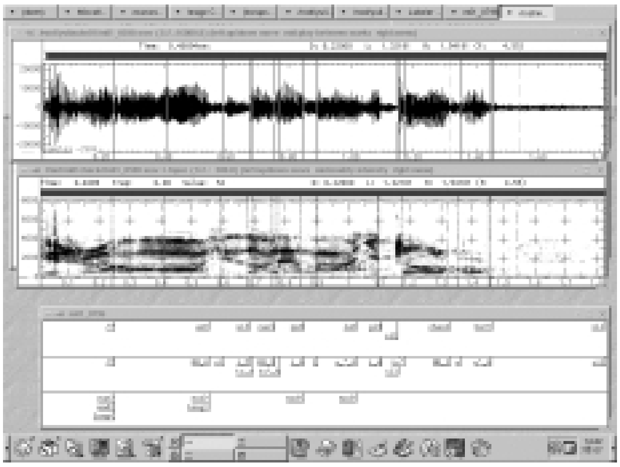

Figure 2. The sound change of “qu4” to “qi0” and “hu1” to a retroflexed “hur0” in sentence “xi3 zao3 qu4 la0 yi4 zhao1 hu0”.

Additional examples of sound changes are given in Figures 2 and 3. In Figure 2, the final of the syllable “qu4” changes to “i4” and the final of the last syllable is retroflexed. An example is given in Figure 3 of changes in Pinyin resulting from deletion and merging of initials and finals. The final of “gen1”is deleted and the initial and the syllable “wo3” are merged to form a new syllable “guo” in sentence “na4 ci4 gen1 wo3 men0 yi2 kuai4 jiang3 shuo1 wen2 ge2 qi1 jian1”.

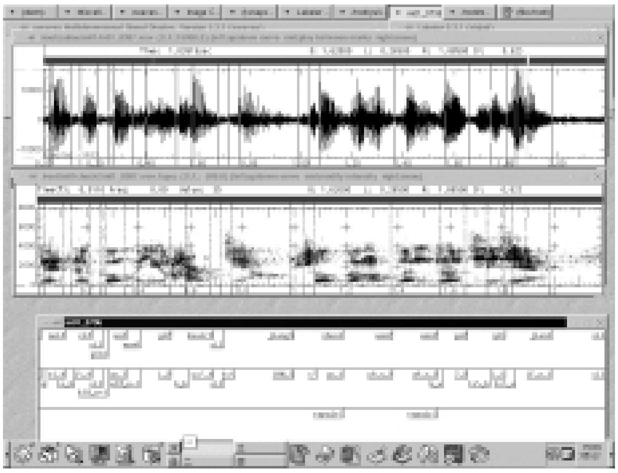

Figure 3. Example of resyllabification resulting from deleted final.

We have tabulated the occurrences of sound changes in initials and finals and syllables and their summary is presented in Table 4. In addition to this, the miscellaneous speech and non-speech events found in 4,000 spontaneous sentences are listed in Table .

Table 4 Sound changes relative to standard pronunciation (in percent) in the CASS database.

	Initials	Finals	Syllables
F03	29.5	25.4	20.9
F04	34.4	22.7	24.1
M01	17.7	20.3	14.2
M02	31.3	22.8	22.0
M03	40.2	23.6	27.8
M04	45.5	24.5	32.9
M12	25.5	24.9	17.6
Total	27.46 (11040/40209)	12.02 (5647/46999)	29.24 (13742/46999)

5. DISCUSSION

In addition to the factors quantified the previous section, the following factors were observed during transcription to have significant influence on spontaneous pronunciation.

Dialect Effects: In this corpus, Speaker M12 has a heavy accent due to his native dialect. For example, rather than using the standard pronunciation “zheng4 zhong3 cheng2 zhang3”, his speech is transcribed as “ze4 zhong3 cheng2 zang3”, in which the retroflexed affricative “zh” becomes the dentalvelar “z”. This is a common sound change made by native Wu speakers.

Speaking Rate: Speaking rate is generally considered to be an influential factor in speech variability, with sound changes more likely to occur in faster speech. While the transcribers observed that speaking rate was a strong factor in speech variability, we note that the overall measurements of speaking rate and sound change do not reflect this.

Speaking Style: Although none of the speakers in this corpus were native Mandarin speakers, they were fluent enough that only expert phoneticians were able to identify their origins. Some of them were judged to have been in Beijing long enough to pick up the habit, commonly attributed to Pekinese speakers, of speaking with little tongue and lip movement. These habits lead to frequent deletion and neutralization of segments.

Prosodic Structure and Context Effects: As discussed and demonstrated in the previous section, sound changes of zero-initials behave differently depending on their acoustic context and the proximity to prosodic boundaries. These contextual effects lead to difficulties in distinguishing an actual zero initial from a fricative, glottal stop, or approximate.

Table 5. Summary of miscellaneous speech and non-speech phenomena in 4,000 spontaneous sentences.

No.	Spoken phenomena	Occurring times
1	Lengthening	409
2	Breathing	401
3	Laughing	40
4	Crying	0
5	Coughing	65
6	Disfluency	230
7	Noisy	627
8	Murmur/uncertain	567
9	Modal or exclamation	1511
10	Lip Smacking	40
11	Not Chinese	18

6. CONCLUSION

We have presented a collection of phonetically transcribed fluent Mandarin speech collected in realistic speaking conditions. The collection is rich in the speech phenomena found only in spontaneous speech. The corpus is intended for use by linguists studying Chinese speech, and in developing automatic speech recognition and synthesis systems. The CASS corpus is under active development; updated analyses will be made available at http://www.clsp.jhu.edu/ws2000/groups/mcs.

7. REFERENCE

[1] Lin Tao and Wang Lijia, “The Course of Phonetics”, PKU Publishing House, Beijing, Second Edition, 1999.

[2] Lin Tao, “The Overview of Experimental Phonetics”, PKU Publishing House, Beijing.

[3] Li Aijun, Chen Xiaoxia, Sun Guohua, Hua Wu, Yin Zhigang, Zu Yiqing, Zheng Fang, Song Zhanjiang, “The phonetic labeling on read and spontaneous discourse corpora,” to appear in the proceedings of International Conference on Spoken Language Processing (ICSLP), Oct. 2000, Beijing.

[4] John Wells, “Computer-coding the IPA: a proposed extensions of SAMPA”. Unpublished notes, Department of Phonetics and Linguistics, University College London, http://www.phon.ucl.ac.uk/home/sampa/home.htm

[5] Chen Xiaoxia, Li Aijun, etc , “An Application of SAMPA-C for Standard Chinese”. to appear in the proceedings of International Conference on Spoken Language Processing (ICSLP), Oct. 2000, Beijing.