A Design of the Spontaneous Chinese Learner Speech Corpus

Wu, Chen-huei

Shih, Chilin

(Citation)
Learner Corpus Studies in Asia and the World,2:115-124

(Issue Date)
2014-05-31

(Resource Type)
departmental bulletin paper

(Version)
Version of Record

(JaLCDOI)
https://doi.org/10.24546/81006694

(URL)
https://hdl.handle.net/20.500.14094/81006694

A Design of the Spontaneous Chinese Learner Speech Corpus

Chen-huei WU
National Hsinchu University of Education

Chilin SHIH
University of Illinois at Urbana-Champaign

Abstract

The Spontaneous Chinese Learner Speech Corpus consists of 185 hours of audio and video recordings, which was obtained from Chinese speech training classes on a weekly basis from 2004 to 2009 at University of Illinois at Urbana-Champaign. The speakers in this corpus includes 11 Chinese language teacher, 11 Korean-speaking learners, 23 English-speaking learners and 86 Chinese heritage learners. Two paradigms, namely, "Variety Show" and "Debate" were designed to fit in a 50-minute class. Speaker turns were marked with the video editing software ELAN to provide speaker codes and the precise time boundaries demarcating the hour-long recordings into speech turns. Based on the turn-markings, each snippet was displayed on a webpage to obtain a turn-sychronized transcriptions. The corpus data were used for perceptual ratings and acoustic analysis of fluency and foreign accent, language assessment, speech recognition etc. The database is a prolific resource with speech samples for various research topics.

Keywords

Spontaneous speech, Oral fluency, Foreign accent,
Second language acquisition, Chinese language learning

I Introduction

Due to the development in computational power, networks and computer storage, analyzing large amounts of spontaneous speech has recently become a possible task. The speech data in this study includes spontaneous and prepared speech which have been produced in a natural setting and presents phenomenon that may not be observed in experiments of reading word lists. One of the purposes of building this corpus is for perceptual ratings on oral fluency and foreign accent. In addition, acoustic attributes related to fluency and foreign accent were investigated.

This database is prolific recourse due to various speaker background. Before the corpus data can be used for further research, it required many layers of labor-intensive work, such as turn-marking, transcription. A subset of the data was selected for the perceptual rating and acoustic analysis. The general properties of the corpus and each step of the data management is explained in the following sections.

II The Corpus Design

The speech data was recorded in a Chinese speech training class on a weekly basis from Fall 2004 through Spring 2009 at University of Illinois at Urbana-Champaign (Shih, 2006). One hundred and eight five hours of audio and video recordings were collected from the third-year and fourth-year Chinese language classes. In the subsections, speaker background, task types, speaker turn-marking, transcription guidelines are address.

2.1 Speakers

The speakers in the corpus includes 11 Chinese teachers (9 females and 2 males), 86 Chinese heritage learners (28 females and 58 males), 11 Korean learners of Chinese (7 females and 4 males) and 23 English learners of Chinese (8 females and 15 males). The eleven Chinese instructors are Mandarin native speakers; among them, five are from Taiwan and six from Mainland China.

The heritage learners were students whose native language is Chinese, but who received education in English and grew up in the United States. Some were born in the U.S. and some arrived in the U.S. at a young age. Most of the Chinese heritage learners were from Mainland China and some were from Taiwan. Recently, the language learning development of heritage speakers has attracted the attention of many researchers in second language acquisition (Au, Knightly, Jun & Oh, 2002; Polinsky, 2006; Montrul, 2008). Heritage speakers are adult early bilinguals of minority languages. They might be the children of first generation immigrants or might have lived in an L2 country at some point during childhood. Under these conditions, the heritage language might not be completely acquired due to the fact that children of first generation immigrants have a strong desire to fit into the new society. These speakers also speak the native language on a limited basis and in a restricted environment. Therefore, the heritage language used at home might gradually be dominated by the majority language of the new society. The competence and performance of heritage speakers varies to different degrees because of incomplete L1 acquisition. Generally speaking, they have good speaking and listening abilities, with native-like pronunciation and fluency. Although L1 acquisition might not be completely acquired in heritage speakers' childhoods, some of them might go back to L2 classes to improve or maintain their L1.

The non-native L2 learners are students whose native language is English or Korean. The English learners of Chinese are learners who had no prior background in Chinese before they attended the college-level Chinese classes. Different from English learners of Chinese, most of the Korean learners had prior background in Chinese during their high-school education.

2.2 Task

Students in the Chinese classes received speech training in two paradigms, namely, "Variety Show" and "Debate" (Shih, 2006). Each of the paradigms was designed to fit in a 50-minute class. In the *Variety Show* format, there are 4 main sessions: opening, talk show, formal speech and comments. Learners are asked to play roles, such as to be the chair for the whole show, to be the talk show host, or to be the speech makers. The chair opens each session in the show with an introduction; the talk show host prepares several topics and selects students from the audience to step up in front of the stage and answer questions. The speech makers give prepared, formal, 4 to 6 minute speeches. The *Variety Show* format incorporates a few frequently encountered social interactions, such as giving an opening/closing remarks, introducing guest speakers, and giving formal speeches. Through weekly practices, learners had multiple chances to play each role and to observe many performances by their classmates and instructors.

In the *Debate* format, students are divided into two sides, a proposition side and an opposition side. A specific topic is given in advance. Some of the learners prepare a formal speech to express their positions on the given topic; some prepare questions to ask the opposing side; and some have to answer questions on the spot. The *Debate* format trains learners to argue and speak clearly, logically and convincingly under time pressure.

Based on different formats, there are two speech styles, namely: (1) spontaneous speech in which students speak without advanced preparation, i.e., some questions and all answers in the Variety Shows and Debates; and (2) prepared speech, i.e., speeches made by the chair or host, the formal speeches prepared by students in Variety Shows and the statements students made in Debates. If students prepared their speeches beforehand, most of them read their speech and held drafts in their hands. Thus, prepared speech can be recognized in the video clips. Overall, the speech style of Debate is more formal than that of the Variety Show.

2.3 Speaker Turn

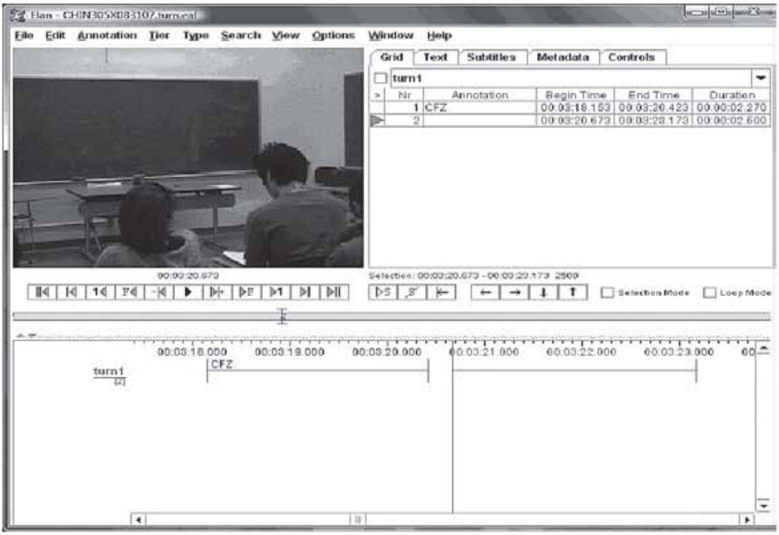
Speaker turn-marking is a labor-intensive and time-consuming step in the procedure of database management. The goal of turn-marking is to provide a proper unit for speech sample selection and time-synchronized transcription. Speaker turn-markings were annotated with a precise start time and end time for each turn and identified with speaker codes. There are several reasons why turn-marking is necessary for future research projects.

i. An individual speaker's turn, as a unit, facilitates speech sampling for individual speakers. The annotation of speaker codes enables researchers to create pools of speech samples from the same speakers and do sample selection. If speech is overlapped by multiple speakers in a speech snippet, for instance, it will be difficult for raters to evaluate speech. It also increases the difficulty in acoustic analysis.

ii. Discourse coherent speech is preferred in perceptual human-rated experiments. Speech samples starting from a random point in the speech might bias fluency or accentedness scores. With speaker turn-marking, we are able to select speech samples from the beginning of turns.

iii. Long speech files increase the chance of misalignment between speech and text when using automatic speech recognition technology (ASR).

In this procedure, speaker turns were marked and annotated by trained research assistants using the video editing software ELAN (Hellwig, n.d.). An example of ELAN annotation is given in Figure 1. The annotator uploaded classroom video and audio into ELAN, dropped cursors to indicate the precise time of turn boundaries and entered the speaker code. The 50-minute recordings were thus demarcated with speaker codes and time stamps indicating the precise turn boundaries of their speech. The information enables synchronization of speech and transcription and facilitates sampling speech for perceptual ratings, acoustic measures, and ASR training.



□ turn1

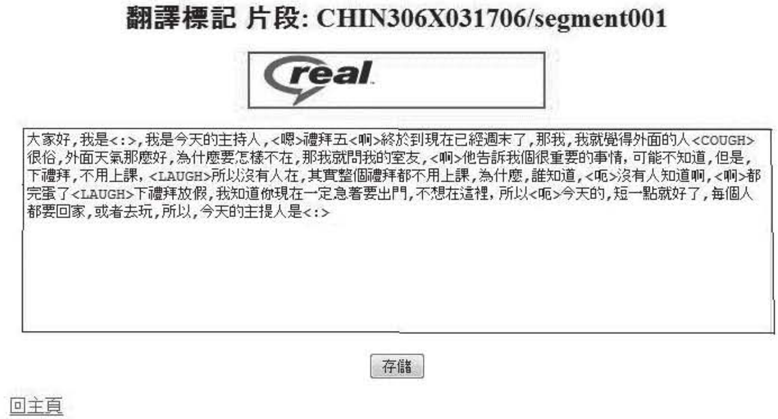
Nr	Annotation	Begin Time	End Time	Duration
1	CFZ	00:03:18.153	00:03:20.423	00:00:02.270
2		00:03:20.673	00:03:23.173	00:00:02.500

Selection: 00:03:20.673 - 00:03:23.173 2500

Fig. 1: An example of ELAN annotation

2.4 Transcription

The transcription was obtained through a transcription website, where each speaker turn was presented individually with a link to the audio/video files. A text area is provided for verbatim transcribing. The transcribers can make edits and revisions and all versions of the revisions were saved. Speech data from both the classroom recordings and the picture telling corpus were transcribed. The transcription was done using traditional Chinese characters by trained transcribers. Speech recordings produced in English in the picture telling tasks were transcribed in English. A snapshot of the transcription website is given in Figure 2.



翻譯標記 片段: CHIN306X031706/segment001

大家好,我是<:>,我是今天的主持人,<嗯>禮拜五<啊>終於到現在已經週末了,那我,我就覺得外面的人<COUGH>很俗,外面天氣那麼好,為什麼要怎樣不在,那我就問我的室友,<啊>他告訴我個很重要的事情,可能不知道,但是,下禮拜,不用上課,<LAUGH>所以沒有人在,其實整個禮拜都不用上課,為什麼,誰知道,<呃>沒有人知道啊,<啊>都完蛋了<LAUGH>下禮拜放假,我知道你現在一定急著要出門,不想在這裡,所以<呃>今天的,短一點就好了,每個人都要回家,或者去玩,所以,今天的主提人是<:>

Fig. 2. A snapshot of the transcription website

Specific linguistic and non-linguistic phenomena, such as disfluencies, speech errors and laughter were labelled with a pair of angle brackets < >. Below are the transcription guidelines.

1. Non-linguistic events, such as laughter, claps, coughs, and other loud noises are transcribed as <LAUGH>, <CLAP>, <COUGH>, and <NOISE>, respectively.
2. Filled pauses are annotated with corresponding Chinese characters, such as <嗯>, <>, and <呃>.
3. Unclear speech is labelled with <SKIP>.
4. English in speech is transcribed in English enclosed in a pair of angle brackets. If it is not understandable, then it is marked as <ENG>.
5. If speech overlaps significantly, the speech is not transcribed but is instead tagged as <OVERLAP>.
6. Speech errors are annotated with the expression of <erroneous syllable/intended

syllable>. The erroneous syllables or actual spoken sounds are transcribed in Zhuyin, followed by the intended and correct Chinese character, separated by a slash and all enclosed in a pair of angle brackets, for example, <ㄉㄧㄢ 3/等>. This is a case where the speaker intends to say 等一下 'deng3 yi1 xia4' 'wait a minute', but she said ㄉㄧㄢ 3 一下 'dian3 yi1 xia4'. The numeral indicates the Chinese tone. Learners' systematic pronunciation errors are not annotated.

An example of the transcription is provided as below.

<嗯>, 大家好。大家好。非常高興大家今天<ㄋ>來參加這一個活動。我先簡單的自我介紹一下。我叫做<WZH> ,我是四年級中文課的 <TA> 。那今天我們要舉行一場辯論的活動。辯論的題目是,中國傳統應該保持。坐在這邊的五位是我們正方的,參加同學,然後在這邊左手邊的是我們,五位,還有一位還沒來,我們,的,反方的同學。那,在這裏我先簡單的講一下我們<ㄉㄧㄢ 3/等> 一下辯論進行的方式。

<uh> Hello, everyone. Hello, everyone. I am very glad that all of you came to join this activity. Let me briefly introduce myself. My name is <WZH>, I am a fourth-year Chinese <TA>. Today, we are going to have a debate. The topic of the debate is ``Chinese traditions should be preserved". Sitting here are the five students of the proposition side, while the other five sitting at the left-hand side are the students of the opposition side. There is another who has not come yet. Well, here I am going to briefly explain the procedure of the debate. <CLAP>.

- <WZH> is the annotation of speaker code.
- <TA> is the annotation of English words when the speaker said 'TA'.
- <ㄉㄧㄢ 3/等> is the annotation of speech error.
- <CLAP> is the clapping after this segment of speech.

2.5 Sampling Design

Good sampling design is an important aspect of research which can lead to reliable statistical inference and predictions. In this study, speech samples should be long enough for evaluating and analyzing fluency and foreign accents[1]. In addition, each speech sample should include the inclusion of multiple sentences if possible, rather than being restricted to sentence fragments or single sentences. However, there is no universal agreed-upon length of speech samples for perceptual ratings. A study by Ambady and Rosenthal (1993) demonstrated that student's ratings of instructor's nonverbal behaviors based on 30 seconds of silent video clips composed of three 10 seconds clips

[1] The results of perceptual ratings on oral fluency and foreign accent has been published (Wu, 2013).

from the same teacher, or even thinner slices of 6 seconds and 15 seconds, successfully predicted end-of-semester teaching evaluation. This finding suggested that impressions can be formed extremely quickly. Derwing et al. (2006) used 20-second speech samples for evaluating fluency and foreign accent and observed that 20 seconds was sufficient for raters to make reliable judgment. Nevertheless, there is an inevitable trade-offs between the length of the speech samples and duration of the experiment. Using longer speech samples increases not only the duration of the experiment, but also the demands on raters. Another limitation of longer speech samples is that it is difficult to obtain long spontaneous speech from language learners if their oral proficiency is low. Statistically, the more samples for each individual speaker, more representative the results will be. Due to all these concerns, one-minute of speech for each speaker composed of four 15-second snippets at different times in a spontaneous speech style (mainly from the questions/answers in the *Variety Show* was randomly selected from the corpus. Speech samples from different blocks of semesters were chosen.

For native speakers, all 11 Chinese instructors who fully acquired their L1, Mandarin, served as the baseline for comparing the results with heritage and English learners of Chinese. Four 15-second snippets were randomly selected from the database. For language learners, each semester had 15 recordings of class sessions, which were divided into 3 blocks. Two snippets were chosen from the first block, the beginning of the semester (the first five weeks) and two snippets were chosen from the last block, the end of the semester (the last five weeks). Between the blocks at the beginning and at the end of the semester, there as a four or five week gap. Speech samples of 17 heritage speakers (5 females and 12 males) and 20 English learners of Chinese (5 females and 15 males) were randomly chosen based on the block design. Two hundred and thirty-six speech files were chosen for analysis.

III Acoustic Analysis of Fluency

Phone labels were obtained by using the Penn Phonetics Lab Forced Aligner (Yuan & Liberman, 2008). In order to improve the alignment, the authors added a dictionary containing Zhuyin symbols, speakers codes corresponding to the name pronunciation used in speech, noise and disfluency transcription. The outcome of the automated phone segmentation was inspected and corrected manually by the author. With phone segmentation, acoustic attributes as listed below were automatically extracted for further analysis (Cucchiarini et al., 2000; Ramus et al., 1999).

1. FPct: Number of filled pauses such as uh's and um's
2. FPdur: Duration of filled pauses
3. Articulation Rate (AR): Number of vowels (syllable nuclei) / utterance duration without silence

4. Rate of Speech (RS): Number of vowels (syllable nuclei) / utterance duration including silence
5. Phonation time ration (PTR): Utterance duration without silent pauses / utterance duration with silent pauses
6. Percentage Vowels (Pv): Duration of vowels (vocalic segments) / utterance duration without silence
7. Standard deviation of vowel duration (stdV)

The acoustic attributes, FPct, FPdur, AR, RS, PTR, Pv and stdV measured the temporal properties of the speech produced by individual speakers in each snippet. Table 1 shows the mean values of the acoustic measures for speaker groups. The distribution of the acoustic attributes is presented in Figure 3, where English learners have most FPs and longest duration of FPs, followed by heritage learners and native speakers. The AR, RS, and PTR all reveal that native speakers speak faster than heritage learners and English learners. Pv shows that native speakers have slightly longer vowel production than other speaker groups, and stdV shows more variation in vowel length.

Table 1: Mean value of acoustic measures for speaker groups

Speaker Groups	FPct	FPdur (seconds)	AR (syl/sec)	RS (syl/sec)	PTR(%)	Pv(%)	stdV
Native Mandarin	1.40	0.38	5.36	4.16	77	46	0.46
Heritage Learners	1.98	0.67	4.42	2.87	65	43	0.39
English Learners	3.70	1.48	3.28	1.86	57	42	0.36

A linear relationship between acoustic measures and fluency rating was conducted. The R-squared values indicate that RS, PTR and AR are able to explain, respectively, 68%, 51%, and 38% of the variance in the fluency rating. The frequency and duration of FPs have a negative relationship with fluency rating, meaning that the more/longer the FPs, the lower the fluency rating. Other acoustic measures related to speaking rates, such as AR, RS, PTR show a positive linear regression, indicating that the fast the speaking rate, the higher the fluency ratings. It is not surprising that native speakers have faster speaking rate than heritage and English learners and native speaker received higher fluency scores, while the fluency rating of heritage and English learner speech varies greatly.

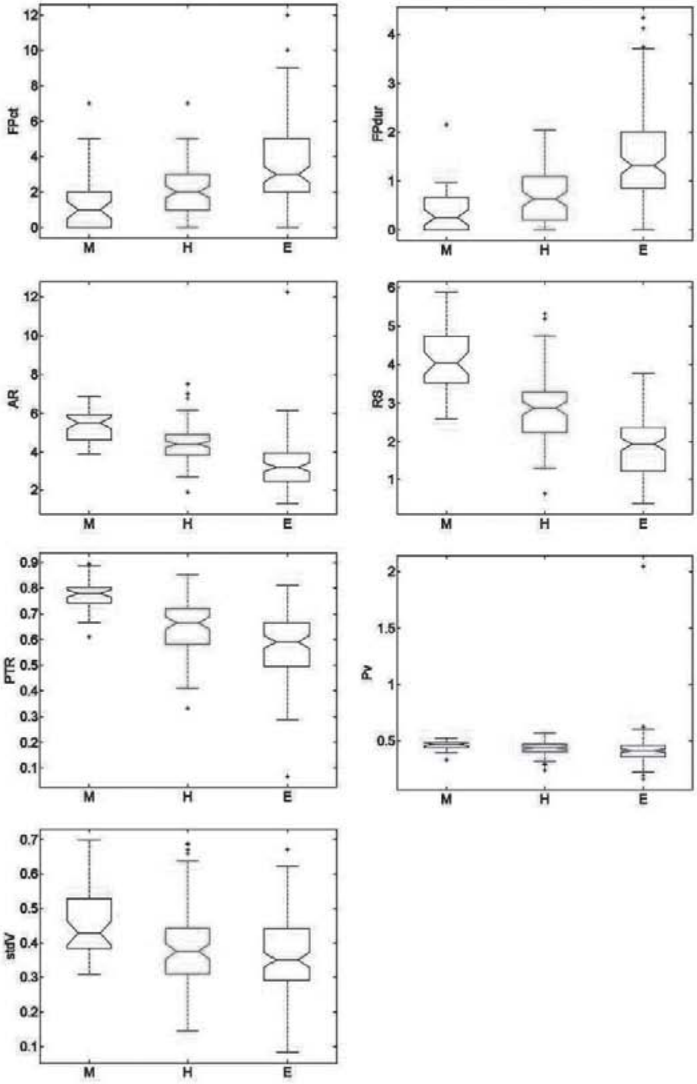


Fig. 3: Boxplots of acoustic attributes

IV Conclusion

This paper reported the design and data management of the spontaneous Chinese learner speech corpus. A subset data elicited from this corpus has been used for perceptual ratings and acoustic analyses on oral fluency and foreign accent. Spontaneous speech is closer to the natural speech that people speak in a daily basis. The result based on the corpus data can better reflect the speech performance by second language learners.

Acknowledgements

This work was supported by the Critical Research Initiatives Grant, UIUC, the NSF-funded project "DHB: An Interdisciplinary Study of the Dynamics of Second Language Fluency," and the 2010 Jiede Empirical Research Grant, CLTA. We thank colleagues on the grants for valuable discussions at UIUC and friends in Taiwan for collecting rating data. Last but not least, we thank students and TAs in Chinese language classes for participating in this project.

References

Ambady, N., & Rosenthal, R. (1993). Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness, *Journal of Personality and Social Psychology, 64*(3), 431-441.

Au, T. K-f., Knightly, L. M., Jun, S.-A., & Oh, J. S. (2002). Overhearing a language during childhood. *Psychological Science, 13*, 238-243.

Cucchiarini, H. S., & Boves, L. (2000). Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *Journal of the Acoustical Society of America, 107*(2), 989-999.

Derwing, T. M., Thomson, R. I., & Munro, M. J. (2006). English pronunciation and fluency development in Mandarin and Slavic speakers. *System, 34*(2), 183-193.

Hellwig, B. (n.d.). *Elan-linguisitc annotator*: http://www.lat-mi.eu/tools/elan/.

Mntrul, S. A. (2008). *Incomplete acquisition in bilingualism: Re-examining the age factor*. Amsterdam: John Benjamins.

Polinsky, M. (2006). Incomplete acquisition: American Russian. *Journal of Slavic Linguistic, 14*, 191-162.

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition, 73*, 265-292.

Shih, C. (2006). The language class as a community: A task design for speaking proficiency training. *Journal of the Chinese Language Teachers Association, 41*(2), 1–22.

Wu, Chen-huei. (2013). The perception of second language fluency and foreign accents. *Journal of Chinese Language Teaching, 10*(2), 117-141.

Yuan, J., & Liberman M.(2008). Speaker identification on the SCOTUS corpus. *Proceeding of Acoustics 2008*, 5687-5690.