MATBN 2002: A MANDARIN CHINESE BROADCAST NEWS CORPUS

Hsin-min Wang

Institute of Information Science, Academia Sinica
Taipei, Taiwan
Email: whm@iis.sinica.edu.tw

ABSTRACT

The MATBN 2002 Mandarin Chinese broadcast news corpus contains a total of 40 hours of broadcast news from Public Television Service Foundation (Taiwan) with corresponding transcripts. The primary motivation for this collection is to provide training and testing data for continuous speech recognition evaluation in the broadcast domain. We expect to collect and process 220 hours of Mandarin Chinese broadcast news speech over 3 years. At the end of the first year, the 40 hour broadcast news corpus has been completed on schedule and is scheduled to be releasable in early 2003. According to our plan, we expect to release the interim 120 hour broadcast news corpus in late 2003 and the final 220 hour broadcast news corpus in late 2004.

1. INTRODUCTION

Starting in 1995, the Defense Advanced Research Projects Agency of the United States (DARPA) directed its research program for continuous speech recognition to focus on automatic transcription of broadcast news [1]. Since then, many research groups worldwide have paid attention to this challenging task and spent great efforts in the collection of broadcast news corpora of various languages [2-4]. Though some Mandarin Chinese broadcast news corpora are available from LDC (Linguistic Data Consortium, USA) [5], they are all in the Mainland China accent and the wording is quite different to that used in the Taiwan area. To support the researchers and technology developers who are interested in studying Mandarin Chinese used in the Taiwan area, we want to collect Mandarin Chinese broadcast news in the Taiwan area.

Due to the success of the previous project to collect **M**andarin speech data **a**cross **T**aiwan (MAT) [6], which was conducted by a group of researchers from several universities and research institutes in Taiwan, the same group of people decided to collaborate again to collect spontaneous speech data in 2001. The previous MAT project spanned the period August 1995 through July 1998. Speech files were collected through telephone networks. The content included read speech (numbers, Mandarin syllables, words of 2 to 4 syllables, phonetically balanced sentences) and a little spontaneous speech (short answering statements). The new MAT project spans the period August 2001 through July 2004. We expect to collect both dialogue speech and broadcast news speech. In the broadcast news part, we expect to transcribe 220 hours of broadcast news in 3 years. The first 40 hour corpus is scheduled to be completed at the end of the first year (July 2002), the other 80 hours are due in July 2003, while the remaining 100 hours should be ready for testing in July 2004.

The transcripts are in Big5-encoded form, with SGML tagging to annotate acoustic conditions, background conditions, story boundaries, speaker turn boundaries and audible acoustic events, such as hesitations, repetitions, vocal non-speech events, external noises, etc. These tags include time stamps to align the text with the speech data. In the first interim 40 hour broadcast news corpus, according to the hand-segmentation, there are 779 stories, 104 headlines, 40 weather reports, and 40 ending sections. Around 33 hours of speech from 10 weather reports and all the stories, headlines, and ending sections are carefully transcribed while the remaining weather reports and segments containing advertising or pure music are just annotated with time stamps without orthographic transcripts.

The rest of this paper is organized as follows: The data collection procedures and the details of transcription and annotation are presented in Sections 2 and 3, respectively. Then, the preliminary assessment of the interim 40 hour Mandarin Chinese broadcast news corpus is discussed in Section 4. Finally, conclusions are made in Section 5.

2. DATA COLLECTION

Public Television Service Foundation (Taiwan) [7] has kindly agreed to share their broadcast news with us. The recordings span the period November 7, 2001 through March 31, 2003 (expected). A Digital Audio Tape (DAT) recorder, which is connected to the broadcasting machine using the XLR balanced cable, has been set up in the TV broadcasting studio. That is, the broadcast news speech was recorded synchronously while broadcasting to avoid the modulation effect. Recordings are in stereo with 44kHz sampling rate and 16 bit resolution. Each recording consists of a broadcast news episode of 60 minutes.

Each DAT was manually processed to transfer the digital speech samples into a single Microsoft Windows wave file and stored in the hard disk. Then, the signal was down-sampled to 16kHz with a resolution of 16 bits. During this operation, only the left channel was selected. Thus, the broadcast news speech in 16KHz, 16 bit resolution, mono, was used for further transcription and annotation. Until now, about 200 hours of broadcast news have been recorded in this way.

Since video can provide some clues to help facilitating the transcription and annotation work, video recordings were also made simultaneously. Recordings are on VHS video tapes. Because we do not have the space to store several hundreds of video tapes, each recording is first converted into an MPEG1 file and then stored on a CD-ROM. After the conversion is completed, the video tape can be reused again. With the video, the broadcast news speech corpus can be expanded to a video broadcast news corpus, though at this stage, we only focus on the audio track.

3. TRANSCRIPTION AND ANNOTATION

The corpus has been segmented, labeled and transcribed manually using a tool developed by DGA (Délégation Générale pour l'Armement, France) and LDC, called "Transcriber"[8]. Two full-time transcribers were hired for this project. They were educated native Mandarin speakers. They worked next to each other and could easily share their experiences. Every sound file was transcribed by one transcriber. When the transcription of a sound file was completed, an additional verification was done by the other transcriber. In addition, the author and the two transcribers have a regular meeting every week, at which further checking was performed and specific problems were discussed and solved.

Sometimes it is hard to correctly identify the speakers or background conditions only by listening to the sound. In such a case, the transcribers can look for clues from the corresponding video file. Furthermore, the transcribers can easily get the story title[1] from the video so they do not need to create one by themselves. In addition to the original conventions used in the DGA&LDC Transcriber, we have also included the other two annotation tag sets. The first of these was designed by Dr. Shu-chuan Tseng for annotating Mandarin conversational dialogue corpus [9] and the second one was provided by Dr. Chiu-yu Tseng originally for annotating spontaneous monologue speech.

The anchor speech is always of a high standard of fluency, good pronunciation and a good acoustic quality. Most of the field reporter speech is also of a high standard of fluency and good pronunciation, but sometimes of a low acoustic quality, while some interviewee speech is of a very low quality and intelligibility with possible background sounds of various types or the speech itself may contain many inappropriate pronunciations, particles, repetitions, repairs, etc. As a result, it takes much longer to transcribe and annotate the interviewee speech. The segments containing dialects or foreign languages are annotated with the language identity and time stamps but no orthographic transcripts.

3.1. SGML Structure of Transcriptions

Owing to the complexity and hierarchical nature of the additional information needed in the transcripts, SGML was chosen by the DGA&LDC Transcriber as the most suitable framework to use in formatting the text. The document structure used for all transcripts is as follows [8]:

For each waveform file (a full 60 minute program here), there is one transcript file, containing a single "Episode" element; the Episode has attributes to identify the file name, the transcriber, and the release version.

Each Episode contains a series of "Section" elements, which equate to the topical units (stories, advertising, etc.) in the Episode; the Section attributes identify the type of unit, and the points in time at which the Section begins and ends in the corresponding waveform file.

Within each Section containing material to be transcribed, there are one or more "Segment" elements, corresponding to speaker turns within the Section; the Segment attributes identify the speaker, the speaking mode, the channel fidelity, and the points in time at which the speaker turn begins and ends.

At any point within an Episode, Section or Segment where there is a change in the presence of music, background voices or other noise, a "Background" element is inserted to mark the change; the Background attributes identify the type of background condition (music, speech, other, shh) and the point in time at which the change occurs.

4. PRELIMINARY ASSESSMENT

The first interim 40 hour Mandarin Chinese broadcast news corpus has been completed and some preliminary assessments have been conducted.

Figure 1 depicts a partial transcription of a broadcast news show. The transcription has three hierarchically embedded layers of segmentation (orthographic transcription, speaker turns, and sections (stories)), plus a fourth layer of segmentation (acoustic background conditions) which is independent of the other three. Some frequent situations are as follows: the non-speech part between the speech segments of two distinct speakers could be chopped into several distinct short segments according to their acoustic foreground and background conditions. Moreover, a speaker turn could be separated into several segments by short silence segments.

[1] In this corpus, every story has been attached a very short topical description.

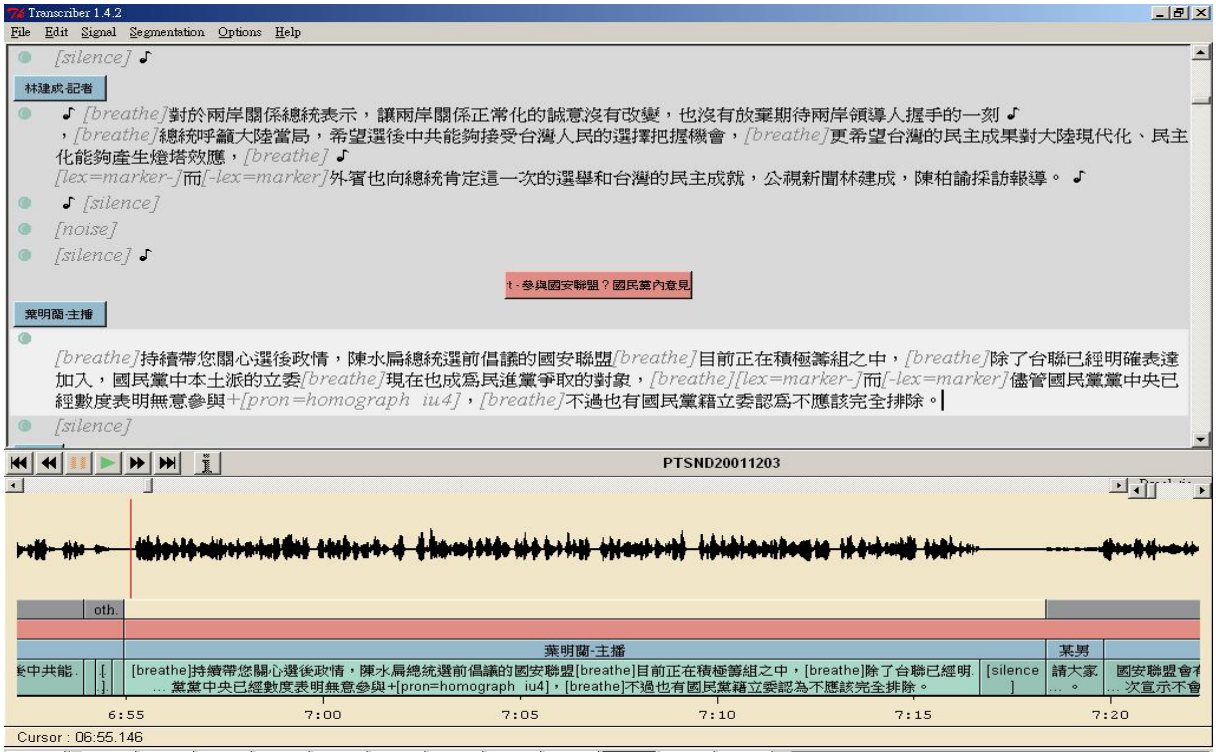

Figure 1. A partial transcription of a broadcast news show

Each 1-hour news show usually contains two or three parts separated by advertising. But sometimes there is no advertising at all within a show. Each part starts with headlines in the presence of background music followed by a number of stories. Because the news shows were collected from a non-profit public TV station, the advertising is comprised of featurettes or previews rather than commercials. As a whole, each 1-hour news show contains one to three headline sections, zero to two advertising sections depending on the number of headline sections, a number of news stories, a weather report section, and an ending section.

There are 3 distinct studio anchors; two are male and one is female. According to the hand-segmentation, there are 779 news stories in total. The total length is around 30 hours and the average length per story is 2.3 minutes. Some other brief statistical information about these 779 stories is summarized as follows:

- There are around 130 distinct field reporters. Of these, around 50 field reporters are male. The true identities of around 50 field reporters are undetermined even though our transcribers have referred to the video. It's likely that some of the unidentified field reporters in different stories in fact respectively correspond to the same reporter. Therefore, the real number of distinct field reporters could be lower.
- There are around 1300 distinct interviewees. Of these, around 900 interviewees are male. It's interesting to find that, unlike the case of the field reporters, the percentage for the male speakers is relatively higher. Moreover, even though the identities of some interviewees are also unknown, it's very likely that the unknown interviewees in different stories are in fact diverse.
- The total lengths of the studio anchor speech, the field reporter speech, and the interviewee speech are around 300 minutes, 850 minutes, and 650 minutes, respectively. Some segments contain overlapping speech.
- The frequency counts of occurrence of the most frequently used tags in the corpus are tabulated in Table 1. It was found that the top 5 most frequently used tags for the studio anchor speech, the field reporter speech, and the interviewee speech are almost the same. The overall top 5 most frequently used tags are "pause", "breathe", "silence", "particle", and "marker", respectively. "Pause", "silence", and "breathe" are common to different types of speech, but very high percentages of the "particle", "marker", "restart", and "repetition" tags are found in the interviewee speech. There are some inappropriate pronunciations (pronunciation errors) in the corpus, in particular in the interviewee speech. We also found that interviewees speak in dialects

Studio anchor speech		Field reporter speech		Interviewee speech		Overall	
Tags	# occurrence	Tags	# occurrence	Tags	# occurrence	Tags	# occurrence
breathe	2430	pause	5437	pause	7815	pause	14405
pause	1153	breathe	5433	particle	4711	breathe	10412
silence	1091	silence	3783	breathe	2549	silence	7177
marker	122	inapp_pronun	488	silence	2303	particle	5083
English	76	marker	367	marker	2272	marker	2761
inapp_pronun	71	particle	329	inapp_pronun	1175	inapp_pronun	1734
particle	43	English	242	restart	756	restart	825
restart	38	Min-Nan	51	Min-Nan	321	English	584
repair	17	restart	31	repetition	282	Min-Nan	377
repetition	11	Formosan*	11	English	266	repetition	302

Table 1. The frequency counts of occurrence of the most frequently used tags in the corpus. Min-Nan is a common dialect and Formosan denotes all the aboriginal languages used in the Taiwan area. The "inapp_pronun" tag is used for annotating inappropriate pronunciations (pronunciation errors). *The number of occurrence for the "uncertain" tag is also 11.

more often than studio anchors and field reporters. Moreover, it's a common situation among different types of speech that some speech segments contain English terms.

In addition to the 779 news stories, there are 104 headline sections (~80 minutes), 21 advertising sections (~13 minutes), 40 weather report sections (~190 minutes), and 40 ending sections (~12 minutes). All the headline sections and ending sections were also carefully transcribed. The weather reports of 10 shows were also carefully transcribed but the weather reports of the remaining 30 shows and all the advertising sections were just annotated with time stamps without orthographic transcripts.

5. CONCLUDING REMARKS

After one year of hard work, we are pleased to announce that the first 40 hour Mandarin Chinese broadcast news corpus has been completed on schedule. The paperwork to make this corpus releasable is in progress. Hopefully, the corpus will be ready in early 2003. As to the project, we are now in the second year. Hopefully, at the end of the second year, the other 80 hour corpus will be ready for testing as well. As mentioned above, we expect to collect and process 220 hours of broadcast news before the summer of 2004. There's still a long way to go.

6. ACKNOWLEDGEMENTS

This project was funded by the National Science Council of the Republic of China under grant No. NSC 90-2213-E-009-109. The author would like to thank Public Television Service Foundation (Taiwan) for sharing their broadcast news with us and their employees for helping us to set up the recording machines in their broadcasting studio and operating them regularly. Acknowledgments go to Dr. Chiu-yu Tseng and Dr. Shu-chuan Tseng for their valuable assistance and comments on the transcription and annotation, Prof. Sadaoki Furui and his colleagues for sharing their experiences with us, Ms. Kuan-jung Chen, Ms. Mei-li Chang, and Ms. Tzau-fang Yan for their hard work on transcription and annotation, Mr. Kuo-hsian Wang and Mr. Yi-hsiang Chao for cloning speech data on the DAT to PC, and Mr. Shi-sian Cheng for doing preliminary statistical analysis on the corpus. Acknowledgements also go to all the colleagues from universities and research institutes that participated in this project.

7. REFERENCES

[1] R. M. Stern, "Specification of the 1996 Hub 4 Broadcast News Evaluation," *Proc. DARPA Speech Recognition Workshop*, 1997.

[2] D. Graff, "An Overview of Broadcast News Corpora," *Speech Communication*, 37, pp. 15-26, 2002.

[3] M. Federico, D. Giordani, P. Coletti, "Development and Evaluation of an Italian Broadcast News Corpus," *Proc. LREC'2000*.

[4] T. Matsuoka, Y. Taguchi, K. Ohtsuki, S. Furui and K. Shirai, "Toward automatic transcription of Japanese broadcast news," *Proc. EUROSPEECH'97*.

[5] Linguistic Data Consortium: http://www.ldc.upenn.edu.

[6] H. C. Wang, "MAT - A Project to Collect Mandarin Speech Data through Telephone Networks in Taiwan," *Computational Linguistics and Chinese Language Processing*, 2(1), pp. 73-89, 1997.

[7] Public Television Service Foundation (Taiwan): http://www.pts.org.tw.

[8] C. Barras, E. Geoffrois, Z. B. Wu, M. Liberman, "Transcriber: Development and Use of S tool for Assisting Speech Corpora Production," *Speech Communication*, 33, pp. 5-22, 2001.

[9] S. C. Tseng and Y. F. Liu, "Mandarin Conventional Dialogue Corpus," *MCDC Technical Note 2001-01*, Institute of Linguistics, Academia Sinica.