Computational Linguistics and Chinese Language Processing
Vol. 10, No. 2, June 2005, pp. 219-236

MATBN: A Mandarin Chinese Broadcast News Corpus

Hsin-Min Wang*, Berlin Chen+, Jen-Wei Kuo+ and Shih-Sian Cheng*

Abstract

The MATBN Mandarin Chinese broadcast news corpus contains a total of 198 hours of broadcast news from the Public Television Service Foundation (Taiwan) with corresponding transcripts. The primary purpose of this collection is to provide training and testing data for continuous speech recognition evaluation in the broadcast news domain. In this paper, we briefly introduce the speech corpus and report on some preliminary statistical analysis and speech recognition evaluation results.

Keywords: broadcast news, corpus, speech recognition, Mandarin Chinese, transcription, annotation

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 [Stern 1997]. Since then, many research groups worldwide have paid attention to this challenging task and put much effort into collecting broadcast news corpora of various languages [Matsuoka *et al.* 1997, Federico *et al.* 2000, Graff 2002]. Though some Mandarin Chinese broadcast news corpora are available from LDC (Linguistic Data Consortium, USA)[1], they all exhibit the Mainland China accent, and the wording is quite different to that used in the Taiwan area. To support researchers and technology developers who are interested in studying Mandarin Chinese spoken in the Taiwan area, we have collected Mandarin Chinese news broadcast in Taiwan.

Due to the success of a previous project which collected **M**andarin speech data **a**cross **T**aiwan (MAT) [Wang 1997] and was completed by a group of researchers from several universities and research institutes in Taiwan, the same group of people decided to collaborate again on collecting spontaneous speech data in 2001. The first MAT project spanned the

* Institute of Information Science, Academia Sinica, Taipei, Taiwan
E-Mail: whm@iis.sinica.edu.tw
+ Graduate Institute of Computer Science and Information Engineering, National Taiwan Normal University, Taipei, Taiwan
1 Linguistic Data Consortium: http://www.ldc.upenn.edu

[Received November 30, 2004; Revised February 15, 2005; Accepted February 17, 2005]

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, and phonetically balanced sentences) and a small amount of spontaneous speech (short answering statements). The second MAT project spanned the period August 2001 through July 2004. We expected to collect both dialogue speech and broadcast news speech. In the broadcast news part, we expected to transcribe 220 hours of broadcast news in 3 years. The first 40-hour corpus was scheduled to be completed by the end of the first year (July 2002), the other 80 hours were due to be completed in July 2003, while the remaining 100 hours were planned to be ready for testing in July 2004. However, we were able to process only 198 hours of broadcast news because our transcribers spent much time correcting errors reported by our colleagues who participated in this project during the second and third years. The 198-hour corpus spanned the period November 17, 2001 through April 3, 2003.

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 that are used to align the text with the speech data. In the 198-hour broadcast news corpus, based on hand-segmentation results, there are 4,100 stories, 581 headlines, 197 weather forecasts, and 197 ending sections. Around 150 hours of speech from 10 weather forecasts and all the stories, headlines, and ending sections were carefully transcribed, while the remaining weather forecasts and segments containing advertising or pure music were just annotated with time stamps without orthographic transcripts. The transcripts contain around 2.3 million Chinese characters in total.

We have established a webpage for the corpus. On this webpage, there are tools that users can employ to query the corpus. Though the project is finished, we will continue to correct errors reported by users. Also, we have selected five one-hour shows as a development set and five more one-hour shows as an evaluation set, and conducted speech recognition experiments on them. 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, a preliminary assessment of the 198-hour Mandarin Chinese broadcast news corpus is given in section 4. The corpus webpage and corpus tools are introduced in section 5. Speech recognition evaluation results are discussed in section 6. Finally, conclusions are drawn in section 7.

2. Data Collection

The Public Television Service Foundation (Taiwan)[2] kindly agreed to share their broadcast

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

news with us. The recordings spanned the period November 7, 2001 through June 30, 2003. A Digital Audio Tape (DAT) recorder, which was connected to a broadcasting machine using an XLR balanced cable, was set up in the TV broadcasting studio. That is, the broadcast news speech was recorded at the same time it was broadcasted to avoid any modulation effect. Recordings are made in stereo with a 44.1kHz sampling rate and 16 bit resolution. Each recording consists of a broadcast news episode 60 minutes long.

Each DAT was manually processed to convert the digital speech samples into a single Microsoft Windows wave file and stored in a 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, broadcast news speech in mono, down-sampled to 16kHz with 16 bit resolution was used for further transcription and annotation. More than 250 one-hour broadcast news shows were recorded in this way. However, we were able to transcribe only 198 of them.

Since video can provide visual clues to facilitate transcription and annotation, video recordings were also made simultaneously with the audio recordings. The recordings were made on VHS video tapes. Because we did not have enough space to store several hundred video tapes, each recording was first converted into an MPEG1 file and then stored on a CD-ROM. After conversion was completed, the video tape was reused again. With the video recording, the broadcast news speech corpus can be expanded into a video broadcast news corpus, though at this stage, we are only focusing 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” [Barras *et al.* 2001]. Two full-time transcribers were engaged in this project. They were educated native Mandarin speakers. They worked next to each other on separate machines, so they could easily share their experiences and discuss the problems they encountered on the job. Every sound file was transcribed by one transcriber. When the transcription of a sound file was completed, an additional verification step was performed by the other transcriber to eliminate errors and to ensure consistency of the data. In addition, the first author of this paper and the two transcribers held a regular meeting every week, at which further checking of the transcription and annotation work was performed, and some specific problems encountered by the transcribers were discussed and solved.

Sometimes it was hard to correctly identify the speakers or background conditions by only listening to the audio recording. In each such case, the transcribers would look for clues in the corresponding video file. In addition to the original conventions used in the DGA&LDC Transcriber, we also included the other two sets of annotation tags. The first of these was designed by Dr. Shu-chuan Tseng for annotating Mandarin conversational dialogue speech

[Tseng 2004], and the second one was provided by Dr. Chiu-yu Tseng, which was originally designed for annotating spontaneous monologue speech. All the annotation tags frequently used in this corpus are listed in the Appendix.

The studio anchor speech always exhibits a high standard of fluency, good pronunciation, and good acoustic quality. Most of the field reporter speech also exhibits a high standard of fluency and good pronunciation, but sometimes the acoustic quality is low. Some of the interviewee speech is of very low quality and intelligibility with background speech and noises of various types, and the speech itself sometimes contains mispronunciations, particles, repetitions, repairs, etc. As a result, much more time was required to transcribe and annotate the interviewee speech. The segments containing dialects or foreign languages were annotated with the language identity and time stamps without 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 a suitable framework for formatting the text. The document structure used in all the transcripts is as follows [Barras et *al.* 2001].

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

Each Episode contains a series of "Section" elements, which correspond 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, a Section or a Segment where there is a change in the presence of music, background voices, or other noises, a "Background" element is inserted to mark the change; the Background attributes identify the type of background (music, speech, other, and shh) and the point in time at which the change occurs.

3.2 Automatic generation of initial transcriptions

After we examined the anchor scripts on the website of the TV broadcaster, we found that they matched the content of the studio anchor speech rather well. This observation led us to design

an automatic tool to generate initial transcription files for our transcribers to start with. The flowchart of the tool for generating initial transcriptions is depicted in Figure 1. It primarily consists of a story segmentation module, a speech recognition module, an anchor script alignment module, and an output module.

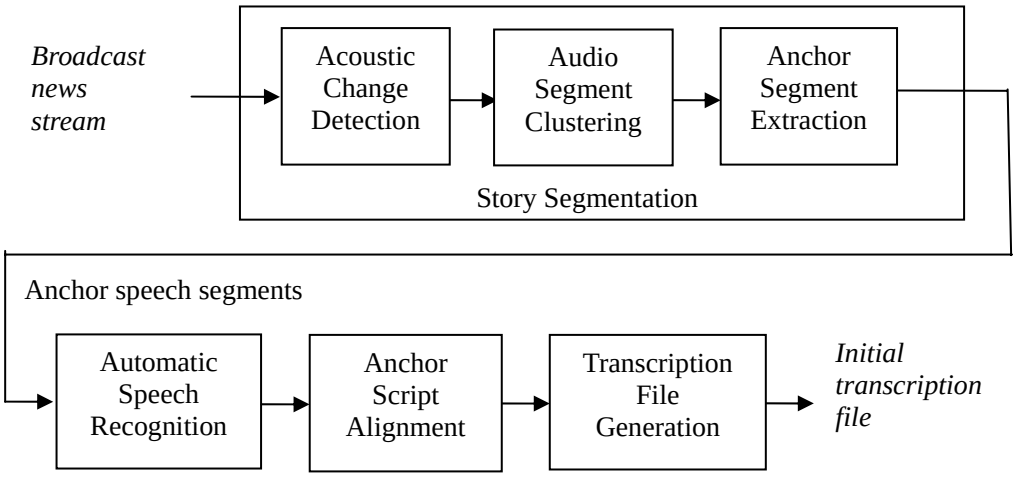

Figure 1. The flowchart of our approach to automatically generating initial transcription files

The story segmentation module first performs speaker and environment change detection, followed by hierarchical clustering of audio segments. We assume that the largest cluster is the studio anchor cluster, and that every studio anchor speech segment is the first segment of a story. Thus, the number of studio anchor speech segments corresponds to the number of stories in the audio stream, and the starting time of a story is the starting time of its studio anchor speech segment. The speech recognition module transcribes all the studio anchor speech segments, and the anchor script alignment module aligns the anchor scripts with the recognition output. Here, we use a vector-space-model-based information retrieval approach for alignment. For each studio anchor speech segment, the recognition output is converted into feature vectors, where the weight of an indexing term (which can be a syllable, a character, a word, or an overlapping N-gram combination) is represented as $tf \times idf$, the term frequency multiplied by the inverse document frequency. Each anchor script is also represented by feature vectors in a similar way. The Cosine measure is used to estimate the relevance between the recognition output and the anchor script. The anchor script with the largest relevance value is aligned with the studio anchor speech segment. Details about the story segmentation approach and the information retrieval approach to story alignment can be found in [Wang *et al.* 2004]. Finally, the output module generates an initial transcription file

consisting of time stamps, topical descriptions[3], and anchor scripts of the stories that correspond to the audio stream.

There is much room for improvement in the present tool; e.g., the IDs of the studio anchor and field reporters could be identified by applying speaker identification techniques. Nevertheless, with this initial transcription, the efficiency was improved to some extent.

4. Corpus Assessment

4.1 A brief description

Each one-hour news show usually has two or three parts separated by advertisements. Sometimes, however, there is no advertising at all within a show. Each part starts with headlines with background music, followed by a number of stories. Because the news shows were collected from a non-profit public TV station, the advertising was composed of public service announcements and previews rather than commercials. Generally, a one-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 forecast section, and an ending section.

Figure 2 depicts a partial transcription of a broadcast news show. The transcription has three hierarchically embedded segmentation layers (orthographic transcription, speaker turns, and sections (stories)), plus a fourth segmentation layer (acoustic background conditions), which is independent of the other three. Frequently, the non-speech part between speech segments produced by two different speakers is chopped into several distinct short segments according to their acoustic foreground and background conditions. Moreover, a speaker turn may be separated into several segments by short silence segments.

4.2 Preliminary statistical analysis

Some preliminary assessments of the 198-hour Mandarin Chinese broadcast news corpus have been conducted. There are seven distinct studio anchors; three are male, and four are female. The distribution of studio anchors is summarized in Table 1. It is obvious that the distribution is quite unbalanced; one female anchored 83.84% of the news shows, while two males each anchored only one show. According to the hand-segmentation results, there are 4,100 news stories in total. The total length is around 143 hours, and the average length per story is 2.1 minutes. Some other brief statistical information about these 4,100 stories is summarized as follows:

[3] The topical descriptions were copied from the website of the TV broadcaster because every anchor script was associated with a manually-generated topical description there.

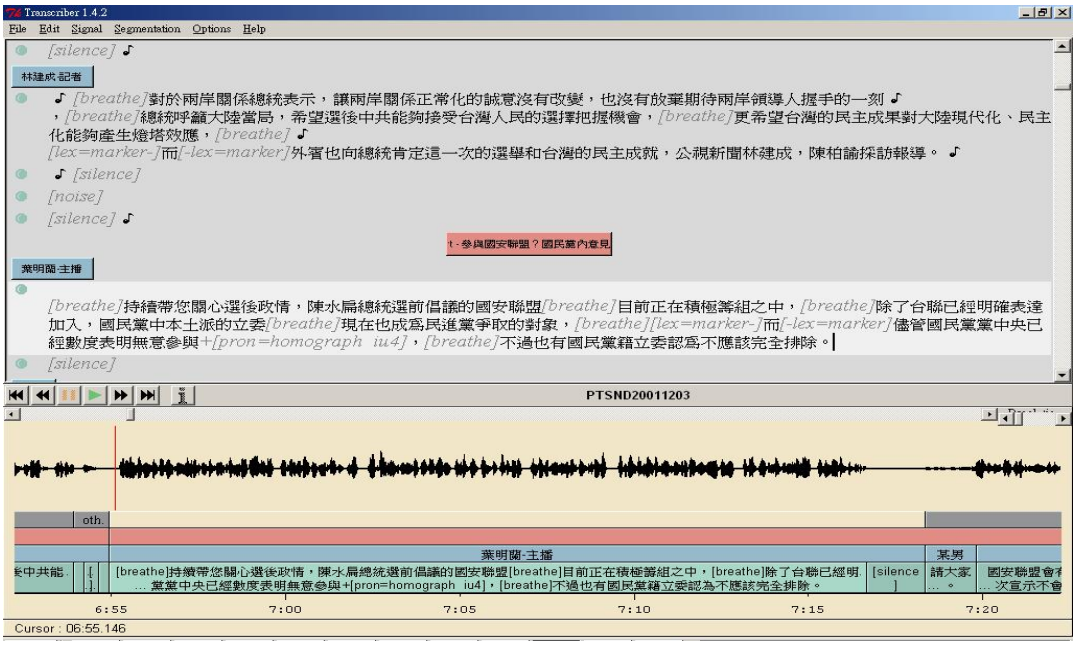

Figure 2. A partial transcription of a broadcast news show

Table 1. The distribution of studio anchors in the 198-hour broadcast news corpus

Gender	Identity	Number of shows	Percentage
Male	Male 1	15	7.58%
	Male 2	1	0.51%
	Male 3	1	0.51%
Female	Female 1	166	83.84%
	Female 2	7	3.54%
	Female 3	5	2.53%
	Female 4	3	1.52%
Sum		198	100%

(1) There are 386 distinct field reporters. Of these, 77 field reporters are male. The identities of 22 male and 70 female field reporters can be identified. The true identities of 294 field reporters are undetermined even though our transcribers have referred to the video recording for clues. It is very likely that some of the unidentified field reporters in different stories in fact correspond to the same reporter, while some correspond to the 92

identified reporters. Therefore, the exact number of distinct field reporters may be much lower than 386 but slightly higher than 100.

(2) There are around 5,900 distinct interviewees. Of these, around 4,000 interviewees are male. It is interesting to find that, unlike the situation with the field reporters, the percentage for male speakers is relatively high. Moreover, even though the identities of some interviewees are also unknown, it is very likely that the unknown interviewees in different stories are in fact different people.

(3) The total lengths of the studio anchor speech, field reporter speech, and interviewee speech are around 27 hours, 71 hours, and 45 hours, respectively. The numbers of characters transcribed from them are around 493,000, 1,167,000, and 616,000, respectively. The speaking rates are 4.9, 5.1, and 4.5 syllables per second, respectively. Some segments contain pure music or noise. In addition, some speech segments contain foreign languages, dialects, or aboriginal languages. If these segments are omitted, then the total lengths of the studio anchor speech, field reporter speech, and interviewee speech are around 25 hours, 58 hours, and 35 hours, respectively. Some speech segments contain overlapping speech.

(4) The frequency counts of the most frequently used tags in the corpus are shown in Table 2. The overall top 5 most frequently used tags are “breathe,” “pause,” “mispronunciation,” “particle,” and “discourse marker,” respectively. The “breathe” and “pause” tags are common to different types of speech, but very high percentages of the “mispronunciation,” “particle,” and “discourse marker” tags are found in the interviewee speech. This is because the studio anchor speech and field reporter speech mostly consist of planned speech, while the interviewee speech mostly consists of spontaneous speech. It is interesting that the 9th most frequently used tag in the field reporter speech is “Formosan,” though the count number is not very high. This is because some field reporters were aborigines and they pronounced their own names in aboriginal languages. We also found that the interviewees spoke in dialects more often than the studio anchors and field reporters did. Moreover, in different types of speech, some speech segments contain English terms.

In addition to the 4,100 news stories, there are 581 headline sections (~5.5 hours), 652 advertising sections (~23.5 hours), 197 weather forecast sections (~10.3 hours), and 197 ending sections (~0.8 hours). All the headline sections and ending sections were carefully transcribed. The weather forecasts in 10 shows were also carefully transcribed, but the remaining 187 weather forecasts and all the advertising sections were simply annotated with time stamps without orthographic transcripts. The transcripts contain around 2.3 million Chinese characters in total.

Table 2. The frequency counts 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

Studio anchor speech		Field reporter speech		Interviewee speech		Overall	
Tags	Count	Tags	Count	Tags	Count	Tags	Count
Breathe	20780	Breathe	47125	pause	33987	breathe	89352
pause	5881	Pause	26070	breathe	21447	pause	65938
mispronunciation	780	mispronunciation	6845	particle	21153	mispronunciation	26134
Particle	327	English	992	mispronunciation	18509	particle	22428
English	306	Particle	948	discourse marker	5258	discourse marker	6090
discourse marker	232	discourse marker	600	restart	3695	restart	3911
Restart	160	Min-Nan	263	syllable contraction	1407	English	2626
repair	58	syllable contraction	219	English	1328	syllable contraction	1667
repetition	51	Formosan	79	repetition	1197	Min-Nan	1424
Syllable contraction	41	restart	56	Min-Nan	1131	repetition	1263

5. Corpus Webpage and Tool

Though the project is finished, we will continue to correct errors reported by users and post the most up-to-date versions of transcriptions on the corpus webpage, http://sovideo.iis.sinica.edu.tw/SLG/corpus/MATBN-corpus.htm. Currently, on this webpage, two tools are available for querying the corpus. Figure 3 shows a snapshot of the tool for querying sections. The tool shows statistical information, time stamps, and orthographic transcripts corresponding to sections such as news stories, headlines, endings, weather forecasts, etc. Figure 4 shows a snapshot of the tool for querying speakers. With this tool, users can easily find statistical information, time stamps, and orthographic transcripts corresponding to the speakers they specify.

6. Speech Recognition Evaluation

6.1 The development and evaluation sets

Ten one-hour shows were selected from the 198-hour carefully transcribed broadcast news database to evaluate the speech recognition performance. That is, we spot-checked about 5% (10/198*100%=5.05%) of the database. We divided the shows into a development set and an evaluation set. The development set consisted of five shows recorded on 2003/01/24, 2003/01/27, 2003/02/07, 2003/03/05, and 2003/03/06, while the evaluation set consisted of five shows recorded on 2003/01/28, 2003/01/29, 2003/02/11, 2003/03/07, and 2003/04/03. The basic guidelines for making selections were as follows: First, we wanted to include as many studio anchors as possible. Second, the test shows had to be broadcast after January 1st, 2003 so that we could use the newswire text before January 1st, 2003 to train the language models. In this section, we will report the speech recognition results obtained for the development set and evaluation set. These results are meant to serve as a reference for future

studies that use this corpus.

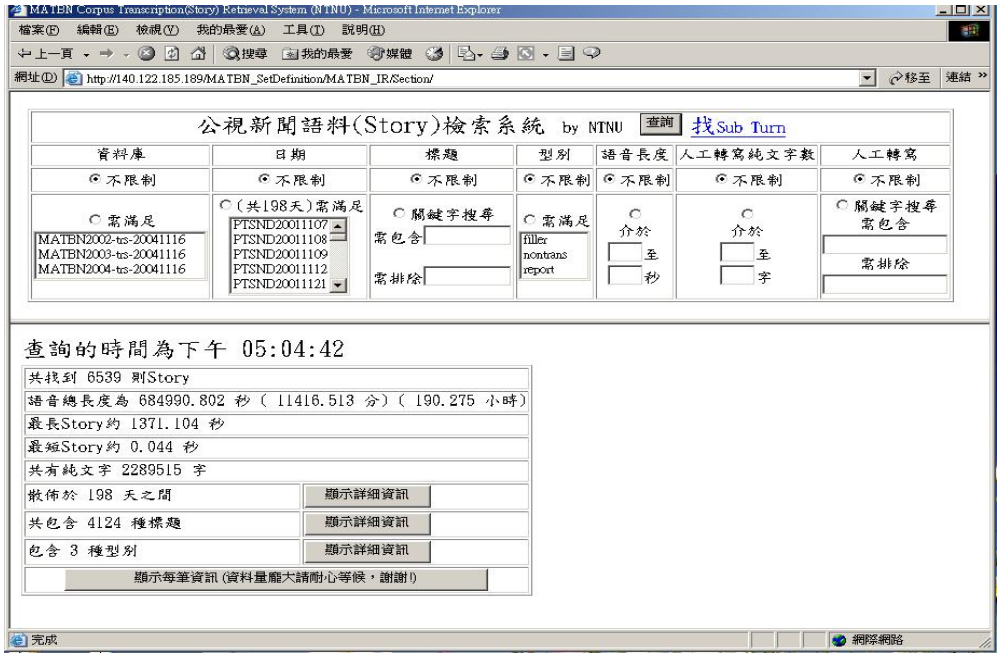

Figure 3. The tool for querying sections

Figure 4. The tool for querying speakers

6.2 The NTNU broadcast news transcription system

The NTNU broadcast news transcription system [Chen *et al.* 2004] was employed here for speech recognition evaluation. Some features of the speech recognizer are reviewed below.

6.2.1 Front-end signal processing

In the speech recognizer, spectral analysis is applied to a 20 ms frame of speech waveform every 10 ms. For each speech frame, 12 mel-frequency cepstral coefficients (MFCCs) and the logarithmic energy are extracted, and these coefficients and their first and second time derivatives are spliced together to form a 39-dimensional feature vector. In addition, to compensate for channel noise, utterance-based cepstral mean subtraction (CMS) is applied to the training and testing speech.

6.2.2 Acoustic modeling

The acoustic models were trained with a database of 16 hours of broadcast news speech collected from several radio stations located in Taipei. The broadcast news data were recorded using a wizard FM radio connected to a PC and digitized at a sampling rate of 16 kHz with 16 bit resolution. The data collection period was from December 1998 to July 1999. The training database is a combination of two corpora: The first corpus contains two hours of field reporter and interviewee speech, and four hours of studio anchor speech. The manual transcripts have been time-aligned to the phrasal level. The second corpus contains ten hours of studio anchor speech. Each audio file is a short news abstract (lasting 50 seconds on average) produced by a studio anchor. Unlike the first corpus, only the orthographic transcripts were available for each audio file; detailed time alignment was unavailable.

Due to the monosyllabic structure of the Chinese language, where each syllable can be decomposed into an INITIAL and a FINAL, the acoustic units used in the speech recognizer are intra-syllable right-context-dependent INITIAL/FINALs, including 112 context-dependent INITIALs and 38 context-independent FINALs. Each INITIAL or FINAL is represented by a Continuous Density Hidden Markov Model (CDHMM) with two to four states. The Gaussian mixture number per state ranges from 1 to 64, depending on the amount of corresponding training data available. In addition, the silence model is a 1-state CDHMM with 128 Gaussian mixtures trained with the non-speech segments. A total of 12,419 mixtures were obtained.

6.2.3 Lexicon and language modeling

The lexicon contains 71,694 words, including 66,290 words selected manually from the CKIP lexicon and 5,404 new words or compound words extracted automatically from the language model training corpus. The word-based unigram, bigram, and trigram language models were trained using the newswire text corpus, consisting of 170 million Chinese characters collected

from Central News Agency (CNA) in 2001 and 2002 (the Chinese Gigaword Corpus released by LDC). The language models were further processed with Katz backoff smoothing [Katz 1987] using the SRI Language Modeling Toolkit (SRILM)[4].

6.2.4 The speech recognizer

The speech recognizer was implemented with a left-to-right frame-synchronous tree search as well as a lexical prefix tree organization of the lexicon. At each speech frame, the so-called word-conditioned method was used to group path hypotheses that shared the same history of predecessor words into the same copies of the lexical tree, and to expand and recombine them according to the tree structure until a possible word ending was reached. At word boundaries, the path hypotheses among the tree copies that had the same search history were recombined and then propagated to the existing tree copies or used to start new ones if none yet existed. At each speech frame, a beam pruning technique, which considered the decoding scores of path hypotheses together with their unigram language model look-ahead and syllable-level acoustic look-ahead scores [Chen *et al.* 2004], was used to select the most promising path hypotheses. Moreover, if the word hypotheses ending at each speech frame had scores that were higher than a predefined threshold, then their associated decoding information, such as the word start and end frames, the identities of current and predecessor words, and the acoustic score, were kept in order to build a word graph for further language model rescoring. Once the word graph had been built, the Viterbi beam search with higher language model weighting [Wessel *et al.* 2001] was performed on it to generate the most likely word sequence. In this study, the word-based trigram language model was used in both the tree search procedure and the word graph rescoring procedure.

6.3 Speech recognition experiments

We conducted speech recognition experiments on the development set and the evaluation set both separately and jointly. Table 3 provides some statistical information about the development set and evaluation set, and the speech recognition results obtained from them. The data corresponding to the development set, the evaluation set, and the union of the two sets are, respectively, depicted in Table 3 (a), (b), and (c). To facilitate the calculation of recognition rates, utterances containing foreign languages, dialects, or aboriginal languages were discarded prior to analysis. It is obvious from Table 3 that both the statistical information and speech recognition results for the development set and evaluation set are very similar to each other. The recognition accuracy for the interviewee speech is extremely poor, but the accuracy for the studio anchor speech and the field reporter speech is quite reasonable. This is

[4] The SRI Language Modeling Toolkit: http://www.speech.sri.com/projects/srilm/

obviously because most of the anchor speech and field reporter speech exhibits a high level of fluency, good pronunciation, and good acoustic quality, while most of the interviewee speech is of very low quality and intelligibility with background sounds of various types, and the speech itself sometimes contain mispronunciations, particles, repetitions, repairs, etc.

Table 3. Speech recognition evaluation results

(a) The 5-hour development set

	Number of speakers	Number of utterances	Speech length (sec.)	Syllable accuracy	Character accuracy	Word accuracy
Studio anchor	3	168	2438.709	79.18%	74.05%	65.74%
Field reporter	36	345	5608.170	65.43%	58.04%	48.74%
Interviewee	142	191	2463.509	26.57%	19.71%	15.55%
Overall	181	704	10510.388	60.57%	53.82%	45.34%

(b) The 5-hour evaluation set

	Number of speakers	Number of utterances	Speech length (sec.)	Syllable accuracy	Character accuracy	Word accuracy
Studio anchor	3	160	2347.630	78.98%	73.72%	66.42%
Field reporter	32	317	5215.351	67.42%	59.78%	50.27%
Interviewee	151	197	2300.656	26.75%	19.53%	14.66%
Overall	186	674	9863.637	61.46%	54.58%	46.06%

(c) Overall estimation (the union of the development and evaluation sets)

	Number of speakers	Number of utterances	Speech length (sec.)	Syllable accuracy	Character accuracy	Word accuracy
Studio anchor	3	328	4786.339	79.08%	73.89%	66.07%
Field reporter	45	662	10823.522	66.31%	58.88%	49.48%
Interviewee	279	388	4764.165	26.66%	19.62%	15.11%
Overall	327	1378	20374.026	61.00%	54.19%	45.69%

6.4 Discussion

Notice that, in a benchmark test, the model parameters optimally tuned for the development set are applied to the evaluation set directly. Further tuning conducted on the evaluation set is not allowed. In this study, we aimed to select development and evaluation data from the corpus so that users can experiment on it with the same setting. Therefore, we have only reported the baseline recognition accuracy of the development set and evaluation set. Furthermore, in this corpus, it is obvious that most of the anchor reporters and field reporters are female, while most of the interviewees are male. Therefore, it is difficult to design a benchmark test with a balanced number of female and male speakers using different types of

speech. However, this may simply reflect the real situation in Taiwan. In that case, we do not need to worry about the gender issue very much.

7. Concluding Remarks

Speech resources are crucially important for research and development in speech technology. But the development of a speech corpus used to be very tedious and time consuming. In August 2001, we started a 3-year project aimed at collecting a Mandarin Chinese broadcast news corpus in Taiwan. At the end of the project, we had labeled 198 one-hour news shows using the DGA&LDC Transcriber. In this paper, we have discussed the development and evaluation of the corpus. The speech corpus will soon be available through the Association for Computational Linguistics and Chinese Language Processing (ACLCLP). In the future, we will try to collect data from other TV broadcasters. We will also try to collect data from various programs.

Acknowledgements

This project was funded by the National Science Council of the Republic of China under grants NSC 90-2213-E-009-109, NSC-91-2219-E-009-039, and NSC-92-2213-E-009-021. The authors would like to thank the 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. Acknowledgements 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 experience with us, Ms. Kuan-jung Chen, Ms. Mei-li Chang, and Ms. Tzau-fang Yan for their hard work on transcription and annotation, Mr. Guo-hsien Wang and Mr. Yi-hsiang Chao for cloning speech data on the DAT to PC, Mr. Tzan-hwei Chen for developing the tool for automatically generating initial transcriptions, and all the colleagues from universities and research institutes that participated in this project. Acknowledgements also go to three anonymous reviewers for their helpful comments.

References

Barras, C., E. Geoffrois, Z. B. Wu and M. Liberman, “Transcriber: Development and Use of a Tool for Assisting Speech Corpora Production,” *Speech Communication*, 33, 2001, pp. 5-22.

Chen, B., J. W. Kuo and W. H. Tsai, “Lightly Supervised and Data-driven Approaches to Mandarin Broadcast News Transcription,” *Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004.

Federico, M., D. Giordani and P. Coletti, "Development and Evaluation of an Italian Broadcast News Corpus," *Proceedings of the 2nd International Conference on Language Resources and Evaluation, 2000*.

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

Katz, S. M., "Estimation of Probabilities from Sparse Data for Other Language Component of a Speech Recognizer," *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(3), 1987, pp. 400-401.

Matsuoka, T., Y. Taguchi, K. Ohtsuki, S. Furui, and K. Shirai, "Toward Automatic Transcription of Japanese Broadcast News," *Proceedings of the 5th European Conference on Speech Communication and Technology*, 1997.

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

Tseng, S.-C., "Processing Spoken Mandarin Corpora," *Traitement automatique des langues, Special Issue: Spoken Corpus Processing*, 45(2), 2004, pp. 89-108.

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

Wang, H. M., S. S. Cheng and Y. C. Chen, "The SoVideo Mandarin Chinese Broadcast News Retrieval System," *International Journal of Speech Technology*, 7(2-3), 2004, pp. 189-202.

Wessel, F., R. Schluter, K. Macherey and H. Ney, "Confidence Measures for Large Vocabulary Continuous Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, 9(3), 2001, pp. 288-298.

Appendix

In the DGA&LDC Transcriber, the annotation tags are divided into 4 categories, namely, noise, pronounce, language, and lexical. We list them in Table A1. An event can be instantaneous (e.g., 我從來沒想過 <Event desc="laugh" type="noise" extent="instantaneous"/>會來參加這個勞工遊行，), sustained (e.g., 這個倒<Event desc="laugh" type="noise" extent="begin"/>沒什麼<Event desc="laugh" type="noise" extent="end"/>。), or associated with the previous character (e.g., 所以根據不同可能發<Event desc="mispronunciation hua1" type="pronounce" extent="previous"/>生的狀況，).

Table A1. The tags used in the corpus

(a) noise

description	example
breathe, clear throat, click, cough, cry, hiccup, laugh, noise, pause, sigh, silence, smack, sneeze, sniffle, swallow, yawn, etc.	我從來沒想過 <Event desc="laugh" type="noise" extent="instantaneous"/> 會來參加這個勞工遊行，
cough, cry, laugh, yawn, etc.	這個倒 <Event desc="laugh" type="noise" extent="begin"/> 沒什麼 <Event desc="laugh" type="noise" extent="end"/> 。
particle	所以各位 <Event desc="particle" type="noise" extent="begin"/> NE <Event desc="particle" type="noise" extent="end"/> 在我們各自的工作崗位上，
unrecognizable non-speech sound	我揣摩 <Event desc="unrecognizable non-speech sound" type="noise" extent="begin"/> ... <Event desc="unrecognizable non-speech sound" type="noise" extent="end"/> 笛子的

(b) pronounce

description	example
alternative	也就是待 <Event desc="alternative dai1" type="pronounce" extent="previous"/> 在家裡
mispronunciation	所以根據不同可能發 <Event desc="mispronunciation hua1" type="pronounce" extent="previous"/> 生的狀況，
stutter, syllable contraction, uncertain, unrecognizable speech sound, etc.	尤其是派系比較嚴重的 <Event desc="syllable contraction" type="pronounce" extent="begin"/> 這些 <Event desc="syllable contraction" type="pronounce" extent="end"/> 地方，
zhuyin	就連 <Event desc="zhuyin" type="pronounce" extent="begin"/> ㄅㄆㄇㄈ <Event desc="zhuyin" type="pronounce" extent="end"/> 他也是看不懂。

(c) lexical

description	example
abridged, cut, editing term, error, interrupted, discourse marker, new word, repair, repetition, restart, etc.	因爲肌瘤切除術我們曉得 <Event desc="repetition" type="lexical" extent="begin"/> 它的它的 <Event desc="repetition" type="lexical" extent="end"/> 一些缺點

(d) language

description	example
English, Formosan, Hakka, Japanese, Min-Nan, Unknown, etc.	經濟部長林信義今天已經率領 <Event desc="English" type="language" extent="begin"/> WTO <Event desc="English" type="language" extent="end"/> 代表團出發前往卡達。