INITIAL EXPERIMENTS ON RECOGNITION OF INTERNET-ACCESSIBLE COMPRESSED MANDARIN SPEECH

Wen-ping HSIEH, Berlin CHEN, Kuan-ting CHEN, and Hsin-min WANG

Institute of Information Science, Academia Sinica, Taipei
{swp, berlin, kenneth, whm}@iis.sinica.edu.tw

ABSTRACT

Massive quantities of spoken audio are becoming available on the web. For example, many radio and television stations are now broadcasting Internet-accessible contents. Automatic recognition of spoken audio that has been degraded by the compression schemes, which enable the delivery of streaming audio over the Internet, could be of great interest for indexing and retrieval purposes. Considering the characteristics and monosyllabic structure of the Chinese language, a syllable-based framework for retrieving Mandarin broadcast news has been investigated at Academia Sinica Taipei. This paper reports on our initial experiments on recognition of Internet-accessible Mandarin broadcast news in two data types - RealAudio and TrueSpeech.

1. INTRODUCTION

Automatic transcription of broadcast news speech is a very challenging task because of the diversity of acoustic environments and the vocabulary used, but it is an important task because of its extensive applications such as news retrieval and event tracking. There have been many previous or ongoing research projects in this area. In 1995, a new evaluation feature, Hub4, sponsored by DARPA was announced, whose evaluation set consists of the business-oriented broadcast material and Marketplace news program obtained by NIST. In 1996, NHK Science & Technical Research Laboratories launched broadcast news recognition projects in cooperation with several research institutes and universities in Japan, aiming at applications such as the automatic subtitling of TV programs. From 1997, TREC (Text REtrieval Conference) has included the spoken document retrieval (SDR) track in the annual evaluation, and the retrieval target is broadcast news. The more recent Topic Detection and Tracking Evaluation Project has included both newswire text and broadcast news in the evaluation since 1997. In the Internet era, radio stations are increasingly broadcasting over the Internet as more consumers begin using their PCs to listen to the radio. We could foresee the far increasing use of audio and video on the web in the near future. Because Internet-accessible speech is always in compressed format due to the restriction of transmission bandwidth, automatic recognition of such data could be of great interest for indexing and retrieval purposes. However, the poor recovery of speech signals inevitably results in poor machine recognition accuracy though the corrupting data compression process won't affect the perception of human ears. There are still not many reports on this very challenging task [1-3].

Considering the characteristics and monosyllabic structure of the Chinese language, a syllable-based framework for retrieving Mandarin broadcast news has been investigated at Academia Sinica Taipei [4]. This paper presents our recent experiments on recognition of Internet-accessible Mandarin broadcast news speech. The initial study aimed at finding out the specific problems related to this challenging task. Thus, we first tested on the acoustic models, language models, and speaker adaptation, all are basic methods established from our previous experiments on the high quality speech data. Two kinds of popular data types over the Internet - RealAudio and TrueSpeech are selected to get a first look at this problem. Both contain highly compressed speech. After we set up the basic environment for this work and put all these modules together, we found that the accuracy degradation due to the corrupting data compression process is still horrible. Substantial work to improve the recognition results as well as to investigate further problems is definitely necessary. From the signal point of view, the restoration of the corrupting speech signals should be helpful. However, the reduction in the acoustic mismatch between the training corpora and the testing corpora could be the most important issue at this stage. Our results bear a close parallel to the results reported in [2] on recognition of Cantonese RealAudio formatted broadcast news.

2. DATA COLLECTION

In this initial study, the test data contain two common data types over the Internet - RealAudio and TrueSpeech. The first collection consists of 29 RealAudio formatted Mandarin broadcast news stories (about 50.7min of speech materials). They were downloaded from the CTS, a local television station at Taipei, web site [5]. The CTS set was manually divided into the anchor part (8.2min), which was produced by news announcers under a studio environment, and the non-anchor part (42.5min), which was produced by field reporters and interviewees under diverse environments with background noise, background music, superposition of multiple speakers, and harsh sounds from machines. It was found that the speaking rates of the field reports are very diverse and such variations could be an important factor for the bad recognition results. A separate development set consists of 200 sentences of the anchor speech was used for tuning the recognition parameters whenever necessary. Both the development set and the anchor part of the above evaluation set are with similar acoustic conditions but were collected in

	Testing data		Training data
Broadcasting stations	CTS (RealAudio)	GOCOOL (TrueSpeech)	VOT et al (PCM raw data)
Environment	Studio/Field	Studio	Studio/Field
Database size: total/anchor/non-anchor (min)	50.7/8.2/42.5	24.7/24.7/0	361.5/235.0/126.5
Speaking rate: total/anchor/non-anchor (syl/sec)	4.96/5.25/4.62	5.24/5.24/-	5.44/5.26/5.78
Fidelity	Poor	Poor	Good
Compression rate	1:19.69	1:14.95	-

Table 1: Summarization of the broadcast news speech databases used in this paper.

different time frames. For simplicity, the recognition results of the development set will not be provided in the following experiments. The second collection consists of 20 Mandarin broadcast news stories (about 24.7min of speech materials) in TrueSpeech format. They were downloaded from the GOCOOL, an Internet-only broadcaster at Taipei, web site [6]. The GOCOOL set contains only the anchors' speech produced under a studio environment. However, it's worth mention that generally speaking these reporters are not professional news announcers and they are of the ages around twenty. Both collections contain highly compressed speech and the compression rates are about 1/20 and 1/15, respectively. All the data were converted to 8 kHz wave format for speech processing. Some interesting statistical information of these two databases is summarized in Table 1.

A different broadcast news database consisting of 453 stories (about 6.0 hours of speech materials) was used for training the speaker-independent HMMs. They were collected from several radio stations, such as Police Radio Station (PRS), UFO Station (UFO), Voice of Han (VOH), and Voice of Taipei (VOT), all located at Taipei, from December 1998 to July 1999. These data were recorded using a wizard FM radio connected to a PC, and digitized at a sampling rate of 16kHz with 16bit resolution. All the recordings were manually segmented into stories and sentences and transcribed. To match the sampling rate of the testing data, the training data were down sampled to 8kHz. The statistical information of the training database is also summarized in Table 1.

3. BASELINE SPEECH RECOGNITION EXPERIMENTS

First of all, we have performed some baseline speech recognition experiments on the anchor part of the testing databases using the baseline acoustic models. The speech recognizer performed only free syllable decoding without any grammar constraints in this section. In addition to the two compressed broadcast news databases described in Section 2, the other two non-compressed broadcast news databases, including the broadcast news database collected at our laboratory and the LDC Hub-4NE Mandarin broadcast news database, have also been evaluated and the experimental results will be provided for reference.

3.1 Signal Processing

As to the signal processing, each frame of the speech data is represented by a 39 dimensional feature vector, which consists of 12 MFCCs and the logarithmic energy, and their first and second time derivatives. Utterance-based cepstral mean subtraction is applied to all of the training and testing data. This is so far the most useful skill to eliminate channel effects with the least effort.

3.2 Acoustic Modeling

The Chinese language is well known for its monosyllabic structure. Each Chinese word is composed of one to several characters, and all the characters are monosyllabic. The total number of phonologically allowed syllables is only 1345 while these 1345 tonal syllables can be reduced to 416 base syllables (syllables independent of the tones) and 5 tones. Base syllable recognition is, thus, believed to be the first key problem in large vocabulary Mandarin speech recognition as well as in spoken document retrieval. However, although the base syllable is a very natural recognition unit for Mandarin Chinese due to the monosyllabic structure of the Chinese language, it suffers from inefficient utilization of the training data in the training phase and high computation requirements in the recognition phase. Thus, the acoustic units chosen in the baseline experiments are 112 context-dependent Initials and 38 context-independent Finals based on the special structure of Mandarin syllables [4]. Here, Initial means the initial consonant of the base syllable, and Final means the vowel (or diphthong) part but including optional medial or nasal ending. The 112 context-dependent Initials are obtained by considering the contextual dependency of the 22 Initials with respect to the beginning phonemes of their following Finals. Each Initial is represented by a HMM with 3 states while each Final is represented with 4 states. The Gaussian mixture number per state ranges from 2 to 16, depending on the amount of training data available. Therefore, every syllable unit is represented by a 7-state HMM. In addition, a 1-state HMM with 32 Gaussian mixtures trained using the non-speech segments is used to handle the silence and short pause.

3.3 Baseline Experiments

The experimental results for using the baseline acoustic models and free syllable decoding are shown in Table 2. The third row marked by 'BCC' represents the testing results for the non-compressed broadcast news database. This testing database consists of 757 recordings (about 10.2 hours of speech materials) collected from Broadcasting Corporation of China (BCC), which has been used for our previous spoken document retrieval experiments [4]. For the recognition of the BCC database, the acoustic models were trained from the same training database described in Section 2 except that the sampling rate was 16kHz. The last row marked by 'Hub-4NE' gives the testing result for the LDC Hub-4NE Mandarin broadcast news corpus. Note here we only used the VOA part of the corpus, in which 14 hours were used for training while the rest 1 hour for testing. The sampling rate of this database is also 16 kHz. It's worth mention that we got the similar accuracies for both the BCC and VOA tasks though some conditions between these two tasks are quite

	Syllable Accuracy
CTS - the anchor part	31.13%
GOCOOL	41.53%
BCC (not compressed)	55.70%
Hub-4NE	54.36%

Table 2: The speech recognition results with the baseline speech recognizer only.

different. The major differences are: (1) The VOA database contains both the anchor reports and the field reports while the BCC database contains only the anchor reports. (2) The amount of the training data for the BCC task is less than half the amount of the training data for the VOA task. (3) For the BCC task, the training database was collected from the other radio stations while, for the VOA task, both the training data and the testing data were composed of only the VOA data. When we took a look at the results for the CTS and GOCOOL tasks, we immediately found a horrible accuracy drop compared to the non-compressed BCC task. The relatively poor results are obviously due to the corrupting data compression process and the mismatch between the training data and the testing data while the major reason for the accuracy difference between the CTS task and the GOCOOL task could be the different data compression algorithms with different compression rates. Reducing the mismatch between the training corpora and the testing corpora could be the most important issue at this stage, though, from the signal point of view, the restoration of the corrupting speech signals could be helpful to this challenging task as well. Using the speaker adaptation techniques to map the acoustic models into the testing conditions or applying the data compression algorithm to the training database and then using the compressed data to train the acoustic models might be able to reduce such a difference. Right now, we have only tested on the speaker adaptation techniques, as will be discussed in the following section, while the retraining of the acoustic models is in progress.

4. FURTHER SPEECH RECOGNITION EXPERIMENTS

To enhance the recognition accuracy, we have applied the syllable language models as the linguistic constraints to the search process and designed a set of sophisticated acoustic models, which took more contextual dependency into consideration. In addition, we have applied the MLLR speaker adaptation techniques to the recognizer.

4.1 Language Models

The language models were trained by a newswire text corpus consisting of 80 million Chinese characters collected from Central News Agency (CNA) in 1999. Word segmentation and phonetic labeling were performed for the training materials using a 62k-word lexicon. The syllable bi-gram language models were temporarily selected in this initial study for simplicity while the syllable tri-gram and word-based language models are currently in progress. As shown in Table 3, the accuracies for the CTS data and the GOCOOL data can be improved from 31.13% and 41.53% to 34.80% and 46.82%, respectively, by simply applying the syllable bi-gram language models to the recognizer.

	Baseline Initial-Final model with syllable bi-gram	Inter-syllable Initial-Final model	Inter-syllable Initial- Final model with syllable bi-gram
CTS - the anchor part	34.80%	34.53%	37.08%
GOCOOL	46.82%	44.48%	48.12%

Table 3: The speech recognition results with the syllable bi-gram language models and the inter-syllable Initial-Final models.

4.2 Inter-Syllable Context-Dependent Acoustic Models

To solve the problem of co-articulation in speech, we have to include more contextual dependency in acoustic modeling. Since the baseline acoustic models only took account of the intra-syllable contextual dependency, the most intuitive extension is to construct the inter-syllable context-dependent models, and such an extension will increase the number of Final units from 38 to 874 (38×23). Because such a large number of parameters can catch more characteristics of acoustic phenomena, the enlarged model set can definitely enhance the recognition accuracy if sufficient training data is available. Our training database is obviously far from sufficient because it consists of only 6 hours of speech materials. There have been many techniques for compensating for lack of training data, such as mixture tying. However, as shown in Table 3, the inter-syllable Initial-Final models improve the recognition accuracies for the two testing sets to 34.53% and 44.48%, respectively, though not any data sharing techniques were applied at this stage. The accuracies can be further improved to 37.08% and 48.12% by applying the syllable bi-gram language models to the recognizer.

4.3 Speaker Adaptation

The maximum likelihood linear regression (MLLR) approach has been shown to be effective for unsupervised adaptation. We used the recognition results obtained with the speaker independent models to perform batch MLLR self-adaptation. In the adaptation process, only the HMM mean vectors were adjusted via diagonal regression matrices. The number of regression classes was determined according to the previously constructed regression class trees [7], which were built in a top-down manner via binary-split vector quantization based on the Bhattacharyya distance between the mixture components. In order to achieve more precise clustering, the model space was first divided into 3 subspaces, including the subspace of the cepstral coefficients and the subspaces of the first and second time derivatives of the cepstral coefficients, respectively. Then, the 3 regression class trees were built in each subspace individually. Each of the stream-based clustered regression class trees has 10 layers with 512 leaf nodes, each of which representing a base regression class. A pre-determined threshold value was used in the adaptation process to decide whether or not a node had enough data for regression matrix estimation.

	Without MLLR	With MLLR
CTS - the anchor part	37.08%	37.12%
GOCOOL	48.12%	51.42%

Table 4: The speech recognition results with/without MLLR adaptation.

As shown in Table 4, the accuracy for the CTS data is improved little with the MLLR techniques for unsupervised story-based self-adaptation; on the other hand, the accuracy for the GOCOOL data can be further improved to 51.42%. Experimental results show that the effect of adaptation highly depends on the initial model accuracy and, of course, the story length. As mentioned in Section 2, the length of the anchor segment in the CTS data is much shorter than that of the story in the GOCOOL data. Furthermore, the optimal number of regression classes is difficult to determine. Therefore, we are also working on more robust speaker adaptation techniques.

5. PRELIMINARY EXPERIMENTS ON THE CTS NON-ANCHOR PART

In the above sections, we mainly focused on recognition of the anchor speech of the testing data. In this section, we will describe some preliminary results for recognition of the non-anchor part of the CTS database. The experimental results are shown in Table 5. The accuracy is only 22.64% for free syllable decoding, but can be improved to 25.37% using the syllable bi-gram language models and further improved to 25.86% using both the inter-syllable context-dependent acoustic models and the syllable bi-gram language models. As shown in Table 5, these results are actually much worse than those of the anchor part. This is because of the various background noises and frequently changed interviewees and acoustic conditions, such as environments.

We have also applied our newly developed automatic segmentation [8] tool to the 29 CTS test stories. Experiment results show that the average recognition accuracy is only 20.91% based on automatic segmentation compared to 27.87%, the average recognition accuracy based on manual segmentation as shown in Table 5. Right now, this automatic segmentation tool only detects the change in acoustic conditions and does not exclude the silence or noise segments. This should be the major reason for accuracy degradation. In order to apply the speaker adaptation techniques to the CTS test stories, the segments with the similar acoustic conditions should be further clustered together.

6. CONCLUSION AND FUTURE WORK

This paper presented our recent experiments on automatic recognition of Internet-accessible Mandarin broadcast news speech. The corrupting data compression process has made recognition of such data a very challenging task. Right now the recognition accuracy is obviously very poor. Even with some improvements achieved by applying the language models, the sophisticated acoustic models and the speaker adaptation techniques to the recognizer, the accuracy is still worse than the accuracy of the non-compressed speech, which was obtained by using the baseline acoustic models in free syllable decoding. Substantial work to improve the recognition results is definitely necessary. Reducing the mismatch between the training corpora and the testing corpora could be the most important issue at this stage. The other ongoing studies include robust acoustic features, automatic segmentation, and speaker adaptation. Furthermore, we have been working on using the speech recognition techniques to index 36 hours of contents from 1689 news stories collected from the CTS web site for news retrieval.

	Baseline Initial-Final model	Baseline Initial-Final model with syllable bi-gram	Inter-syllable Initial-Final model	Inter-syllable Initial-Final model with syllable bi-gram
Anchor	31.13%	34.80%	34.53%	37.08%
Non-anchor	22.64%	25.37%	25.03%	25.86%
Average	24.38%	27.30%	26.98%	27.87%

Table 5: The speech recognition results for the CTS testing database.

7. ACKNOWLEDGEMENTS

This work was partially supported by National Science Council under the grant No. NSC 89-2213-E-001-026. The authors would like to thank Mr. Jeih-weih HUNG for his valuable assistance and comments. The authors would also like to thank the Chinese Television System for providing the TV news.

8. REFERENCES

[1] Peter Scheytt, Petra Geutner, Alex Waibeil. "Serbo-Croatian LVCSR on the Dictation and Broadcast News Domain". *IEEE International Conference on Acoustics, Speech, and Signal Processing.* 1998.

[2] Meng Helen M., Lo W. K., Li Yuk Chi, and Ching P. C. "Multi-scale Audio Indexing for Chinese Spoken Document Retrieval". *International Conference on Spoken Language Processing*. 2000.

[3] The SpeechBot white paper, available at http://speechbot.research.compaq.com/

[4] Chen B., Wang H. M., and Lee L. S. "Retrieval of Broadcast News Speech In Mandarin Chinese Collected In Taiwan Using Syllable-Level Statistical Characteristics". *IEEE International Conference on Acoustics, Speech, and Signal Processing*. 2000.

[5] The CTS (Chinese Television System) web site http://www.cts.com.tw

[6] The GOCOOL web site http://www.gocool.com.tw

[7] Gales, M. J. F. "The Generation and Use of Regression Class Trees for MLLR Adaptation". *Technical Report CUED/F-INFENG/TR263*, Cambridge University, Cambridge, U.K. August 1996.

[8] Hung J. W., Wang H. M., and Lee L. S. "Automatic Metric-based Speech Segmentation for Broadcast News via Principal Component Analysis". *International Conference on Spoken Language Processing*. 2000.