ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

A STUDY ON MANDARIN BROADCAST NEWS SPEECH RECOGNITION

C. L. Chen, Y. R. Wang and S. H. Chen

Department of Communications Engineering,
National Chiao Tung University, Hsinchu
leonliang@ms22.url.com.tw, yrwang@cc.nctu.edu.tw

ABSTRACT

In this paper, a basic Mandarin broadcast news speech recognition system is constructed using the MATBN database. It considers the acoustic modeling for Mandarin base-syllables, particles, and paralinguistic phenomena. It also considers environment-dependent acoustic modeling for three recording environments: studio anchors, outdoor reporters, and outdoor interviewee. Moreover, it incorporates a bigram language model with adaptation using data in MATBN. Syllable recognition rates of 89.64, 84.42and 61.62% were achieved for the three environments of anchors, reporters and interviewees, respectively.

1. INTRODUCTION

Broadcast news is a major information source of our daily life. In the past few years, many researches have been done to apply the speech recognition technique to the broadcast news-related tasks, such as broadcast news information retrieval (IR) [1]. A good speech recognizer can help the IR system to preprocess the acoustic signals to generate the transcriptions of program contents. In this paper, a basic Mandarin broadcast news speech recognition system is constructed using the MATBN database collected in Academia Sinica.

The Academia Sinica in Taipei started to collect the broadcast news speech data since 2002. The database is called the Mandarin Chinese Broadcast News Corpus (MATBN) [2]. The transcriptions were manually annotated by using a sophisticated tool, called "*Transcriber*", developed by DGA and LDC. The tool can be used to transcribe many broadcast news speech characteristics and paralinguistic phenomena, and describe the broadcast news speech status hierarchically. The SGML (XML) structure tagging was chosen in *Transcriber*, so we need to write a XML-parser to preprocess the XML-type corpus for the extraction of usable sub-turn based data. Until July 2004, the transcription of 120-hour recording data has been completed. A part of it, the 40-hour data set collected in the first year, is firstly used to construct a preliminary speech recognition system. And then the whole120-hour database is used to develop the final speech recognition system.

There exists a variety of speaking and recording environments in MATBN. If we can model them separately, the acoustic models will be more precise. In this study, we separate them into three classes. One is for the studio anchors and weather anchors. Speeches in this class are of high SNR with less background noise. The other two are for the outdoor reporters and interviewees, respectively. Speeches in these two classes are of lower SNR with larger background noise or music. They are different only in the speaking style. Speeches in the second class are pronounced by well-trained reporters and hence are more clearly, while those in the third class are spontaneous speeches. Due to the fact that these three classes have substantially different speech characteristics, we train three different sets of HMM models for them.

Another problem to be conquered is the spontaneity of the database. Particles and many paralinguistic phenomena, such as breath and laugh, can be found in MATBN. They will hamper the recognition of speech recognizers developed in read speech. We solve the problem by creating some additional HMM models to represent them. Besides, a bigram language model with adaptation using data in MATBN is incorporated to help solving the ambiguity. All those works are exercised using the software HTK version 3.2.1 developed by the Cambridge University [3].

The remainder of this paper is organized as follows. In Section 2, the development of a preliminary speech recognition system is presented. The advancement to environment-dependent acoustic modeling is described in Section 3. Section 4 discusses the incorporation of language model. Some conclusions are given in Section 5.

2. A PRELIMINARY SPEECH RECOGNITION SYSTEM

In this section we present the development of a preliminary baseline speech recognition system using the first-year data set of MATBN. First, a brief introduction to the characteristics of MATBN is given. Then, the construction of HMM models is discussed. Lastly, the performance of the baseline system is examined.

2.1 MATBN Database

As mentioned above, the first-year, 40-hour broadcast news data set in MATBN is chosen to develop the baseline system. We first parse the XML transcription of the data set to extract usable data. In this study, we take all speech segments which contain clear sounds without mixing with background noise or music as usable data. There are in total 8.5-hour usable data. The data set is composed of 40% studio anchor speech, 22% outdoor reporter speech, and 38% outdoor interviewee speech. Table 1 shows the distribution of data types in the three environments. Here particle represents filler sounds like "um" and "a"; breath represents exhale and inhale sounds; garbage represents all paralinguistic sounds except breath; Min-Nan is a dialect of Chinese language commonly used in Taiwan. It can be found from Table 1, speech in MATBN is more or less like spontaneous speech.

Table 1: Distributions of data types in three environments

Data type	Mandarin	Particle	Garbage	Breath	English	Min-Nan
Anchors	95.26%	0.15%	0.01%	4.28%	0.08%	0.01%
Reporters	95.35%	0.24%	0.01%	3.65%	0.36%	0.02%
Interviewee	94.11%	1.81%	0.09%	2.48%	0.13%	0.07%

2.2 Construction of HMM Models

To extract recognition features, a spectral analysis is applied to the speech waveform for every 32ms frame with 10ms frame shift. It extracts 38 recognition features including 12 MFCCs and their first and second derivatives, and delta and delta-delta log-energies. Here the window length for both delta and delta-delta feature calculations is set to 5 frames because of the relatively fast average speaking rate of 5 syllable/sec in MATBN. Besides, DC-bias removal and cepstrum mean subtraction (CMS) are also used in the frond-end pre-processing stage.

We divide the usable data set into two parts: a training set containing about 150K syllables (i.e., 9/10 of the usable data set) and a testing set containing about 19K syllables (1/10 of the usable data set). Before training the HMM models, a forced-alignment procedure is performed to obtain a set of initial HMM models. It uses the existing initial and final sub-syllable HMM models trained from the TCC300 read-speech database to form 411 base-syllable models and to represent particles which sound similarly. It also extracts several breath segments manually to train an initial breath HMM model. Besides, a global variance HMM model is generated and taken as garbage model to cover all other paralinguistic sounds [4,5], and English and Min-Nan speeches. After performing forced-alignment, we train initial models for all HMM models needed by the baseline system. The left-to-right HMM architecture is adopted. Table 2 shows the parameter setting of those HMM models. In the generation of initial HMM models by HTK, we choose the fixed boundary Baum-Welch re-estimation procedure rather than the Flat start.

Table 2: The parameter setting of HMM models in the baseline system

HMM model	Number	State no.	Mix. no./ state
INITIAL(Vowel)	100	3	1~16
FINAL(Consonant)	40	5	1~16
Particle	19	3	1~16
Breath	1	3	16
Silence	1	3	32
SP	1	1	32
Garbage	3	3	32

As shown in Table 2, there are 100 right-*final*-depended 3-state *initial* HMM models, 40 5-state *final* models, 19 3-state particle models, one 3-state breath model, one 3-state silence model, one 1-state short pause model, and three 3-state garbage models. The three garbage models include one for all paralinguistic sounds except breath, and two for English and Min-Nan speeches respectively. The short pause model is used to fill inter-syllable pauses and is tied with the middle state of the silence model.

After obtaining initial HMM models, we re-train all HMM models by using the well-known Baum-Welch parameter re-estimation algorithm. At last, 165 acoustic models which consist of in total 8,275 mixtures are obtained.

2.3 Performance Evaluation

The performance of the preliminary baseline speech recognition system was evaluated using the testing data set for the three environments. Here, free-grammar acoustic decoding was performed. For the purpose of clarity in performance comparison, we only calculated the recognition rate for 411 Mandarin base-syllables. The syllable accuracy rates were 72.63% for studio anchors, 61.55% for outdoor reporters, and 39.75% for outdoor interviewees. The performance is comparable to the SoVideo system trained and tested on the same MATBN database [1]. Based on these experimental results, several improvement methods were proposed and discussed in the next section.

3. ENVIRONMENT-DEPENDENT ACOUSTIC MODELING

To compensate the effect of inhomogeneous speech characteristics in the three environment classes, we train three different sets of HMM models for them. We extract usable data from the 120-hour data set recorded in the first two years. The resulting usable data set contains 420K syllables (or 24.2 hours in length). Nine tenths of the usable data set is used for the training and the remaining one tenth is used for the testing. The training data set consists of 175K, 104K, and 99K syllables for the three environments, respectively.

The training of environment-dependent HMM models starts from performing the forced-alignment to the training

data set using the HMM models of the baseline system. Then the initial HMM models for each environment are constructed. The parameter settings for the three sets of environment-dependent HMM models are basically the same as the baseline system except that the number of particle models in each environment is set according to the amount of available data. There are in total 4, 7, and 16 particle models created for these three environments, respectively. After applying the Baum-Welch algorithm, three sets of HMM models are generated.

Performance of the environment-dependent acoustic modeling was then examined. Table.3 displays the syllable accuracy rates for the three environments. It can be found from Table 3 that the environment-dependent acoustic modeling outperformed the baseline system for all three testing environments. The performance improvements were more significant for the two environments of outdoor reporters and interviewees.

Table.3: The experimental results of the recognition tests using environment-dependent HMM models.

Testing environment	Baseline	Environment-dependent model using
Anchor	72.63%	74.82%
Reporter	61.55%	66.97%
Interviewee	39.75%	42.80%

4. THE INCORPORATION OF LANGUAGE MODEL

We then incorporate a bigram language model to the speech recognizer in order to discriminate words rather than base-syllables. We first select a lexicon containing about 60K words and train a general language model using a large text corpus. We then use the training data set of MATBN to adapt the general language model to make it better fit the testing environment. In the following, we discuss the constructions of the general and adapted language models and their uses in our speech recognition system in detail.

4.1 The General Language Model

A text corpus consisting of three sub-corpora is used to train the general bigram language model. The three sub-corpora are (1) Sinorama: a news magazine containing 11 million words; (2) NTCIR: an IR test bench covering several domains and containing 59 million words; and (3) Sinica corpus: a well-tagged corpus containing 5.8 million words. The total number of words is about 77 millions. A commercial software is first used to tokenize all texts of the first two sub-corpora into word/POS strings. Then 58,940 most high frequency words are selected to form the lexicon for our language modeling. The average length of words in the lexicon is 2.4 characters (or syllables).

We then calculate the bigram probabilities from the well-tagged corpus by

$$P(w_i \mid w_{i-1}) = \frac{Count(w_{i-1}, w_i)}{Count(w_{i-1})} \quad (1)$$

The total number of bigram probabilities calculated from the corpus is around 9.07 millions. To take care of infrequent words, the Good-tuning smoothing scheme provided by HTK is applied. For the case of $Count(.) = 0$, a back-off scheme [6] which uses (n-1)-gram probability to replace n-gram probability is employed.

Another problem to be taken care of is the OOV (Out Of Vocabulary) words. An OOV word is a word not existing in the lexicon. Two kinds of OOV words are needed to be processed. The first one is the Chinese words that are not included in the lexicon. In this study, they are degenerated to mono-syllable strings and taken care of in the language model by adding 411 base-syllables to the lexicon. Another is non-Chinese words or sounds, occurred in MATBN, including particles, paralinguistic sounds, English words, Min-Nan words, and so on. A word class referred to as Unknown word is used to represent all of them.

A problem occurred in the calculation of bigram probabilities as we expand the lexicon to include 411 base-syllables. There are too many data for the degenerated 411 base-syllables. Specifically, 5% data in the text corpus are for the degenerated 411 base-syllables and 95% for the 58,940 words of the original lexicon. A direct calculation of bigram probability will cause over-weighting to 411 base-syllables as incorporating the language model to the speech recognizer. In this study, we simply solve the problem by adding a term to consider the amount of constituent characters in each base-syllable class. For example, the class of base-syllable "ye" is formed by several Chinese characters {業, 頁, 葉, 夜, …}. If the last word in each transition is a degenerated base-syllable, then the new bigram probability is calculated by

$$P'(w_i \mid w_{i-1} = s_j) = P(w_i \mid w_{i-1} = s_j) * P(char \mid s_j), \quad (2)$$

where s_j is the degenerated base-syllable and $P(char \mid s_j)$ is the conditional probability in character level under the base-syllable s_j. In this study, we simply set the conditional probability to be the inverse of the amount c_j of the constituent characters in the base-syllable class s_j, i.e.,

$$P(char \mid s_j) = \frac{1}{c_j} \quad (3)$$

In the word-net, the logarithmic transition probability is used. So we can take care of a bigram probability involving a degenerated base-syllable by just adding a term to the original language score as shown in Fig. 1.

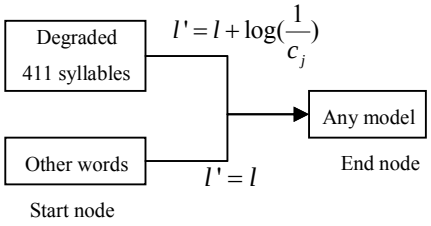

Fig. 1: The modified language score in the word-net

The experimental results of incorporating the general bigram language model to the speech recognizer discussed in Section 3 are displayed in Table 4. Here the weight for the language modeling score was set to be 5 times of the acoustic modeling score. It can be found from Table 4 that the syllable accuracy rates were greatly improved for all the three testing environments.

Table 4: Experimental results of incorporating the general language model into speech recognition

Testing environment	Environment-dependent model without LM	With general LM
Anchor	74.92%	87.80%
Reporter	67.00%	83.44%
Interviewee	40.85%	59.83%

4.2 The Adapted Language Model

One approach to improving the performance of language modeling is to adapt the general language model to the testing environment by using some task-dependent texts. In this study, we take the associated transcribed texts of the MATBN training data set as the adaptation data. The bigram probabilities are then modified by

$$P_{adap}(w_i \mid w_{i-1}) = \lambda P_{Gen}(w_i \mid w_{i-1}) + (1-\lambda) P_{MATBN}(w_i \mid w_{i-1}), \quad (4)$$

where λ is the adaptation weight, $p_{Gen}(\cdot)$ is the bigram probability of the general language model, and $p_{MATBN}(\cdot)$ is the bigram probability calculated from the MATBN training data set. The choice of λ is based on the minimum perplexity criterion. The perplexity of the general language model (i.e., λ=1) is 1453 while it is 751 for the MATBN training data (i.e., λ=0). A lowest perplexity of 551 can be found in the adapted language model for λ=0.4 [7]. With the optimum λ, the adapted LM outperformed the general LM. Comparisons are shown in Table 5 and 6.

Table 5: Comparison of the experimental results using the general LM and the adapted LM

Testing environment	With general LM	With adapted LM
Anchor	87.80%	89.64%
Reporter	83.44%	84.42%
Interviewee	59.83%	61.62%

Table 6: Comparisons of the word and character recognition rates using the general LM and the adapted LM

Different LM using	Anchor		Reporter		Interviewee	
	Word	Char.	Word	Char.	Word	Char.
General LM	68.48%	81.66%	58.97%	76.16%	35.46%	50.65%
Adapted LM	78.09%	86.41%	66.54%	78.63%	42.67%	54.08%

5. CONCLUSIONS

In this paper, the construction of a basic Mandarin broadcast news speech recognition system has been discussed. Spontaneity of the speech characteristics and environmental variation was properly considered in acoustic modeling. Language model adaptation was also applied using domain-specific data. The resulting system has reached high syllable recognition rates of 89.64, 84.42and 61.62% for the three environments of studio anchors, outdoor reporters, and outdoor interviewees, respectively. Significant performance improvement as compared with the baseline system was achieved.

ACKNOWLEDGEMENT

This work was supported by the NSC, Taiwan, ROC, under the project with contract NSC 92-2213-E-009-046 and MOE under the project contract A-93-E-FA06-4-4. The authors would like to thank Dr. Hsin-Min Wang, Institute of Information Science, Academia Sinica, for providing the MATBN database.

REFERENCES

[1] Hsin-Min Wang, Shi-Sian Cheng, and Yong-Cheng Chen, "The SoVideo Chinese Broadcast News Retrieval System", International Journal of Speech Technology 7, pp.189-202, 2004

[2] Hsin-Min Wang, "MATBN 2002: A Mandarin Chinese Broadcast News Corpus", ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR2003).

[3] S. Young, G.. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, *The HTK Book（for HTK Version 3.2.1）*

[4] Liu, D., L. Nguyen, S. Matsoukas, J. Davenport, F. Kubala, R. Schwartz, "Improvements in Spontaneous Speech Recognition", DARPA 1998 Broadcast News Transcription and Understanding Workshop, Leesburg VA, Feb. 1998

[5] Kazuyuki TAKAGI, Shuichi ITAHASHI, "Segmentation of Spoken Dialogue by Interjections, Disfluent Utterances and Pauses", Proc. of the ICSLP-96, pp.697--700

[6] Slava M. Katz, "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer", IEEE Trans. on Acoustic, Speech and Signal Processing, Vol. ASSP-35, No. 3, March 1987

[7] H. Meinedo, N. Souto, and J. Neto, "Speech recognition of broadcast news for the european portuguese language", Proc. of ASRU, 2001