MANDARIN TELEPHONE SPEECH RECOGNITION FOR AUTOMATIC TELEPHONE NUMBER DIRECTORY SERVICE

Yih-Ru Wang, and Sin-Horng Chen

Department of Communication Engineering,
National Chaio Tung University,
Hsinchu, Taiwan, Republic of China
E-mail : yrwang@cc.nctu.edu.tw, schen@cc.nctu.edu.tw

ABSTRACT

This paper discusses an HMM-based Mandarin telephone speech recognition method for implementing a prototype system of automatic telephone number directory service. It adopted the GPD/MCE training algorithm to train the HMM models for 100 final-dependent syllable initials and 40 syllable finals. The SBR method was used to compensate the speaker and channel effects. Besides, an RNN-based pre-classification scheme was employed to speed up the recognition search. A syllable recognition rate of 53.7% was achieved. This method was then used to implement an isolated-word recognizer for the prototype system to discriminate 1922 names of bank and insurance companies. Word recognition rates of 94.8% for top-1 and 97.9% for top-3 were achieved.

1. INTRODUCTION

In service providing systems through public telephone network such as information inquiry systems, the most natural and convenient way for a user to communicate with the system is by the speech. But, it is known that telephone speech recognition is still a difficult task. Channel distortion and background noise interference are two main factors that seriously degrade the recognition performance. The recognition rate of a telephone speech recognition system is usually far lower than that of a clean microphone speech recognition system.

For Mandarin telephone speech recognition, only a few studies have been reported in the past. In [1], IBM built an HMM-based telephone speech recognition system. The lexicon contained 44,000 words. It used a large telephone speech database, referred to as "Mandarin call home database", to train the HMM models. The database contains 7.4-hour spontaneous speech recorded from international telephone calls. The word and syllable recognition error rates were 70.5% and 58.7%, respectively. A comparative study using the Dragon system [2] with fast adaptation and speaker normalization for compensating the channel and speaker variations was also conducted. The word error rate was improved to 50% evaluated based on the same "Mandarin call home database". But, this performance is still far worse than that achieved in clean speech recognition.

Currently, the recognition rate achieved in large-vocabulary telephone speech recognition is still too low to be practically useful. But, for the special cases of small- and medium-size vocabulary, the telephone speech recognition performance is already good enough for developing real applications such as the above-mentioned service providing systems. In some small-vocabulary cases, very good recognition performance can be achieved even without considering the channel and noise compensations. Actually, several useful applications have been developed in the past few years. An English continuous digit string recognizer [3,4] was developed for AT&T employees to access the information of their credit card accounts. In that system, the vocabulary comprised only 11 words including 10 digits and "oh". In the GALAXY system [5], a multilingual speech recognition system was implemented for on-line accessing information, such as the weather's data of several cities, through data networks. It showed that a dialogue system, embedding with a speech recognizer which recognizes several thousand words, can be a very useful task-dependent information inquiry system.

In this paper, an HMM-based Mandarin telephone speech recognition method for a prototype of automatic telephone number directory system is proposed. The system is composed of a speaker-independent, isolated Mandarin word recognition sub-system and a Mandarin text-to-speech (TTS) sub-system. It uses the speech recognition sub-system to recognize the input speech and the TTS sub-system to generate all output response speeches. The whole system is implemented on PC under Windows 95 environment. It uses a telephone interface card supporting TAPI (Telephony API) driver and a 16-bit sound card as the I/O interface to telephone line. The vocabulary to be recognized in the prototype system comprises 1922 words which are names of banks and insurance companies and their shortened forms. The word length is in the range from 2 to 14 syllables.

The organization of the remaining part of the paper is stated as follows. Section 2 discusses the proposed HMM-based Mandarin telephone speech recognition method. The prototype system for automatic telephone number directory service is discussed in Section 3. Some conclusions and future works are given in the last section.

2. THE BASIC HMM-BASED MANDARIN TELEPHONE SPEECH RECOGNITION METHOD

2.1 The baseline HMM syllable recognizer

Mandarin Chinese is a tonal language. Each Chinese character is pronounced as a syllable accompanying with a tone. There are, in

This work was supported by Chinese Telecommunication Labs (CTL), Chinese Telephone Company and National Science Council(NSC 87-2213-E-009-056), Taiwan, ROC.

total, about 1300 syllables. If the tones are disregarded, there are only 411 phonologically allowed base-syllables. The phonetic structures of these 411 Mandarin base-syllables are very regular and relatively simple as compared with English. A base-syllable can be phonetically decomposed into an optional syllable *initial* and a syllable *final*. There are in total 22 initials and 39 finals. As considering the intra-syllable coarticulation, we choose 100 final-dependent initials and 40 finals as basic recognition units. Two left-to-right, gender-dependent HMM models are constructed for each recognition unit. The state numbers in these two types of initial and final sub-syllable HMM models are 3 and 5, respectively. In each state, a partitioned mixture Gaussian observation distributions with diagonal covariance matrices is used. The number of mixture component in each state depends on the number of training data. But, a maximum number of 10 mixture components is used in order to make a compromise between the recognition accuracy and the recognition speed. Besides, a single state HMM and a 3-state HMM are added to model the noise and the breath signals, respectively. The total number of mixture Gaussian components is 8145. The DP search with cumulative bounded state duration constraints is used in the recognition process.

A large continuous telephone-speech database was employed to train these HMM models. The database consists of 71,028 syllables and comprises many short phrases, bank and insurance company names, and short sentences. Its total length is about six hours. It was generated by 94 male and 101 female speakers. All speech signals have passed through the public telephone switching system and were digitally recorded in 8 kHz by using a Dialogic D/41D telephone card and a 16-bit Soundblaster card added on PC. The distribution of the signal-to-noise ratio (SNR) evaluated on per speaker basis is shown in Fig. 1. It can be seen from Fig. 1 that the variation on the speaker-based SNR in the database is very large. All speech signals were pre-processed to extract recognition features including 12 mel-cepstral coefficients, 12 delta mel-cepstral coefficients, and 1 delta energy parameters. The frame length is 240 samples and the frame shift is 80 samples. The signal-bias-removal (SBR) method proposed by [4] was then employed to suppress the speaker and channel variations. The numbers of codewords used in the SBR procedure for mel-cepstral coefficients, delta mel-cepstral coefficients, and delta energy were all empirically chosen to be 32. The system was first off-line tested on a task of recognizing 411 base-syllables using a testing database which consists of 14,512 syllables pronounced by 18 male and 14 female speakers. Syllable recognition rates of 52% and 48.3% were achieved respectively for the cases using and without using the SBR method. The sum of insertion and deletion error rates was as low as 2.5%.

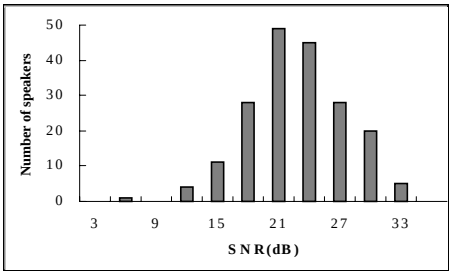

Fig. 1 The distribution of speaker-based SNR in the training database.

2.2 The GPD/MCE training of HMM models

In the above baseline syllable recognizer, all HMM models were trained by using the maximum likelihood (ML) criterion. They were then finer retrained by the GPD/MCE discriminative training algorithm [4]. The misclassification measure used in the GPD algorithm was defined as

$$d(X;\Lambda) = -g(X,W_0;\Lambda) + \ln\left(\frac{1}{r_m-1}\sum_{r=1}^{r_m}\exp\left[g(X,W_r;\Lambda)\cdot\eta\right]\right)^{\frac{1}{\eta}}$$

where $g(X,W;\Lambda)$ is the best likelihood score of the input X for the syllable sequence W, using the model Λ, W_0 is the correct syllable sequence,

$$W_r = \underset{W\neq W_1,\ldots,W_{r-1}}{\arg\max}\ g(X,q_r,W|\Lambda)$$

is the r^{th} best syllable sequence corresponding to W_r, and r_m is the total number of competing syllable sequences other than W_0. In this study, r_m is empirically set to be 20.

Table 1 lists the inclusion rate of the top-N syllable lattice for both the ML- and MCE-trained HMM recognizers. The top-N syllable lattice was found by using the Viterbi-parallel-backtracking method [6] with the from-frame deviation equaling 5 frames. The top-1 syllable recognition rate raised from 52% to 53.7% when the HMM models were finer retrained by the GPD/MCE algorithm. This improvement is slightly worse than that achieved in [7], in which the syllable recognition rate was increased from 69.7% (HMM/ML) to 73.2% (HMM/MCE) tested on a multi-speaker, microphone-speech Putonghua database. This maybe result from the larger variability resided in our database caused by two additional factors of channel and background noise effects other than the speaker effect.

Table 1. Inclusion rates of HMM/ML and HMM/MCE.

Inclusion Rate (%)	HMM/ML		HMM/MCE	
	syllable	string	syllable	string
Top 1	52.0	8.5	53.7	9.3
Top 2	64.5	16.5	65.8	18.2
Top 3	70.8	23.2	71.9	24.7
Top 5	77.6	33.1	78.3	34.6
Top 10	83.8	45.1	84.2	45.8
syllable Ins.	1.7%		1.4%	
syllable Del.	0.8%		0.9%	

3. THE PROTOTYPE SYSTEM FOR AUTOMATIC TELEPHONE NUMBER DIRECTORY SERVICE

3.1 Isolated-Word Recognizer

Using the above HMM models, an isolated Mandarin word recognizer is implemented for a prototype system of automatic telephone number service. The vocabulary contains 1,922 names

of banks and insurance companies located in Taipei. Their word lengths lie in the range of 2 to 14 syllables. Words in the vocabulary is composed of 198 base-syllables out of 411 possible ones. These 198 base-syllables are formed by 78 final-dependent initials and 36 CI finals. Most words are ended with some common sub-words, such as /yin2 hang2/ (bank) and /bao3 xian3/ (insurance), which are generic names. Since these common sub-words carry little information about word discrimination, there exist many confusing word-pairs in the vocabulary. A word pronunciation tree with 5,361 syllable nodes was constructed to cover all these 1922 words and used to guide the recognition search.

In order to speed up the recognition process, a pre-classification scheme was used in the pre-processing stage [8]. The pre-classification scheme first employs a recurrent neural network (RNN) to discriminate each input frame among the 3 classes of silence, initial, and final. Outputs of the RNN are then used to drive a finite state machine (FSM) to classify the input frame into three stable states of silence (S), initial (I), and final (F), and one transient (T) state. Fig. 2 shows the architecture of the RNN. It is a two-layer network with all outputs of the hidden layer being fed back to itself as additional inputs. The inputs of the RNN consist of 26 features including the above-mentioned 25 speech recognition features and one additional log-energy feature. Fig. 3 shows the state transition diagram of the FSM. The FSM is designed to conform to the phonetic structure of Mandarin base-syllables. To drive the FSM, all the three outputs of the RNN are compared with two threshold values, TH_L and TH_H. While one output is higher than TH_H and the other two are all lower than TH_L, the FSM moves into one of the three stable states if it is a legal one. Otherwise, it goes to the T state. In the current system, TH_L and TH_H were empirically set to be 0.2 and 0.8, respectively. By using the pre-classification results, a fast search method [8] is designed and implemented to speed up the recognition process. Its basic idea is to set more restricted search constraints for the three stable states to prune all unnecessary path searches. The operations of the fast search method are explained as follows. When an input frame is classified as an I state, the search is restricted to stay only on all initial HMM states. For an S state, we only search all silence HMM states. For an F state, not only HMM states of all finals but also those of several short

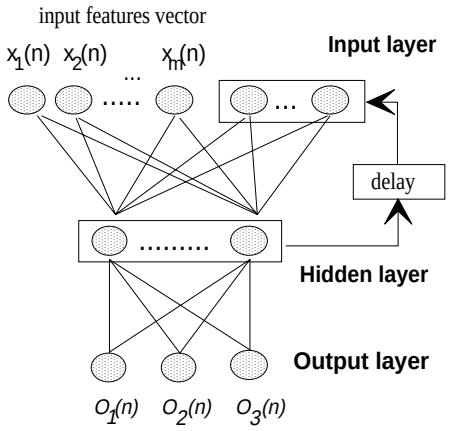

Figure 2. The architecture of the RNN for pre-classification.

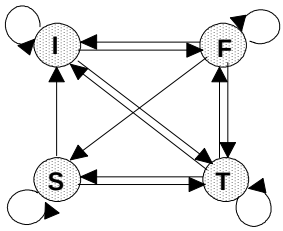

Figure 3. The state transition diagram of the FSM.

initials are searched. For a T state, a full search is performed. Further improvement of the fast search method has also been realized by incorporating with the beam search algorithm. Lastly, a robust end-point detection algorithm has also been implemented based on the pre-classification results of the input speech.

Performance of the isolated-word recognizer was examined by an off-line preliminary test using a database containing 327 testing words. The experimental results are shown in Table 2. It is noted that a full search was used in the HMM/ML recognizer. From Table 2, we find that the HMM/MCE+SBR recognition scheme has the best top-1 word recognition performance (94.8%) with the smallest beamwidth. So it is the best.

Table 2. Recognition rate for isolated-word recognizers.

Recognition scheme	Beam-width	Recognition rate (%)		
		Top 1	Top 3	Top 5
HMM/ML	full	87.2	93.0	93.3
HMM/ML+SBR	2700	93.6	97.9	97.9
HMM/MCE+SBR	2500	94.8	97.9	97.9

3.2 The Prototype System of Automatic Telephone Number Directory Service

An on-line telephone number directory prototype system was implemented on PC under Windows 95 operation system. The system uses the TAPI (Telephony API) interface to control the operations of telephone line. The system architecture is shown in Fig. 3. It is composed of two major parts - an isolated Mandarin word recognition sub-system and a Mandarin TTS sub-system. The speech recognition sub-system uses the HMM/MCE+SBR recognition method discussed above to recognize 1922 isolated words. It takes about 1.9 Mbytes to store parameters of all HMM models. Because of the use of the SBR method to normalize the speaker/channel variation, a two-pass search scheme was used in the recognition process. In the first pass, the SBR feature normalization and the RNN-based pre-classification were done. The recognition search was then performed in the second pass. The system response time is about 2-3 seconds for 1 second input speech evaluated using a PC with Pentium-200 CPU. The system uses the TTS sub-system to generate all output speech responses, including the welcome messages and the recognized telephone number message. The output speech are transformed to 8-kHz, μ-law format and fed to the telephone line through the Dialogic card. It is noted that all response messages can be updated very easily. The TTS sub-system uses an RNN to generate the required prosodic information, including syllable pitch contour, syllable energy level, syllable initial and final duration, and inter-syllable pause duration. The output synthetic speech is generated by a

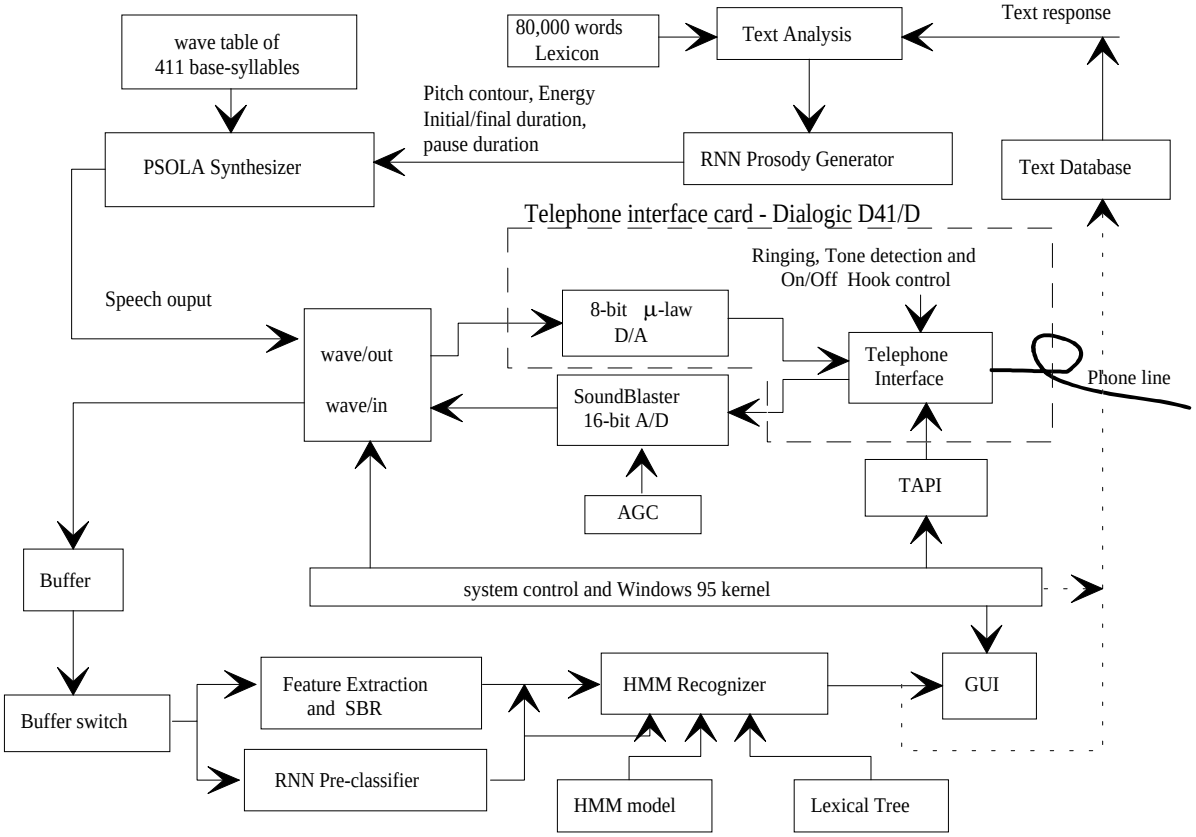

Figure 3. The system architecture of telephone number inquiry system.

PSOLA (pitch synchronous overlap-add) synthesizer. The acoustic inventory of synthesis units is composed of waveform templates of 411 base-syllables. The total memory size for the waveform table and the 80,000-word lexicon for text analyzer is about 3.1 Mbyte.

4. CONCLUSIONS AND FUTURE WORKS

The prototype system of automatic telephone number directory service has been successfully implemented under the support of the Chinese Telecommunication Labs (CTL) of Chinese Telephone company, Taiwan. All functions have been on-line demonstrated to work properly. Further improvements of the system on the recognition performance and speed are still on progress now. The vocabulary size will be expanded in the near future to include 12,000 company names on the yellow pages of a local telephone central office.

5. REFERENCES

[1] F. H. Liu, M. Picheny, P. Srinivasa, M. Monkowski, and J. Chen, “Speech Recognition on Mandarin Call Home: A Large Vocabulary, Conversational, and Telephone Speech Corpus“, ICASSP’96, vol.1, pp.157-160.

[2] J. Barnett, A. Corrada, G. Gao, L. Gillick, Y. Ito, S. Lowe, L. Manganaro, and B. Peskin, “Multilingual Speech Recognition at Dragon Systems“, ICSLP’96, pp.2191-2194.

[3] M. G. Rahim, and B.H. Juang, “Signal Bias Removal by Maximum Likelihood Estimation for Robust Telephone Speech Recognition“, IEEE Trans. On Speech and Audio Processing, vol.4, pp.19-30, Jan. 1996.

[4] B. H. Juang, W. Chou, and C. H. Lee, “ Minimum Classification Error Rate Methods for Speech Recognition, “, IEEE Trans. on Speech and Audio Processing, Vol. 5, No. 3, pp. 257-265, May, 1997.

[5] V. Zue, S. Seneff, J. Polifroni, H. Meng, and J. Glass, “Multilingual Human-Computer Interactions: From Information Access to Language Learning“, ICSLP’96, pp.2207-2210.

[6] E. F. Huang, and H. C. Wang, “ An efficient algorithm for syllable hypothesization in continuous Mandarin speech recognition, “ IEEE Trans. on Speech and Audio Processing, Vol. 2, No. 3, pp. 446-448, July, 1994.

[7] C. H. Lee, B. H. Juang, “ A Survey on Automatic Speech Recognition with an Illustrative Example on Continuous Speech Recognition of Mandarin, “, Journal of Computational Linguistics and Chinese Language Processing, vol 1, No. 1, pp. 1-36, August 1996.

[8] S. H. Chen, Y. F. Liao, S. M. Chiang, and S. Chang, ``An RNN-Based Pre-classification Method for Fast Continuous Mandarin Speech Recognition", to appear in IEEE Trans. on Speech and Audio Processing.

[9] S. H. Hwang, S. H. Chen, and Y. R. Wang, “A Mandarin Text-to-Speech System”, ICSLP’96, pp.1421-1424.