Speaker-Independent Continuous Mandarin Speech Recognition Under Telephone Environments

Jia-lin Shen[1], Ying-chieh Tu[2], Po-yu Liang[2], Lin-shan Lee[1, 2]
1. Institute of Information Science, Academia Sinica
2. Department of Electrical Engineering, National Taiwan University
Taipei, Taiwan, R.O.C.
Tel. 886-2-27883799 ext. 2414, Fax. 886-2-27824814
Email : jlshen@iis.sinica.edu.tw

Abstract

This paper presents a study on speaker-independent continuous Mandarin speech recognition over the telephone. A comparison of several cepstral bias removal techniques such as cepstral mean subtraction (CMS), signal bias removal (SBR) and stochastic matching (SM) for telephone channel compensation was first investigated. Then some modifications and combinations of these techniques were developed for further improvement of the environmental robustness under telephone environments. To better estimate the contextual acoustics and co-articulation in spontaneous telephone speech, the between-syllable context-dependent phone-like units (such as triphones, biphones and demiphones) were used to train the speech models. In addition, the discriminative capabilities of the speech models were further enhanced using the minimum classification error (MCE) algorithms. Experimental results showed that the achieved recognition rates for Mandarin syllables were as high as 59.53%, which indicated a 27.81% of error rate reduction.

1. Introduction

During the past few years, interest has increased in developing spoken dialogue systems over the telephone [1]. Apparently, the recognition performance under telephone environments becomes crucial for a successful spoken dialogue system [2-3]. However, many problems arise from high-quality microphone to telephone networks such that the telephony based speech recognition is still very challenging. First, the speaker independence is highly desired in telephone environments. Secondly, the environmental variabilities become much more serious due to the channel distortions and the fairly high ambient background noise levels. Thirdly, the spontaneous speech over the telephone is very often ill-structured and co-articulated [4-5]. In this paper, some methods for overcoming these problems were developed and investigated.

As we know, the channel noise is usually convoluted with the speech signal in time domain, which becomes an additive term in the logarithmic spectral domain or cepstral domain. Therefore the channel noise can be compensated by subtracting a bias term from the noisy speech signal in cepstral domain (called cepstral bias removal). A comparative study of some widely used cepstral bias removal techniques such as cepstral mean subtraction (CMS)[6], signal bias removal (SBR)[7] and stochastic matching (SM)[8] were first investigated. Then some modifications and combinations were applied based on these techniques for further improvement of the environmental robustness under telephone environments. In order to better estimate the contextual acoustics and co-articulation in spontaneous telephone speech, the between-syllable context-dependent phone-like units (such as triphones, biphones and demiphones) are modeled. Moreover, the minimum classification error (MCE) algorithms are further used to enhance the discriminating ability of the speech models [9].

The baseline system is based on the context-dependent phone-like units (PLU) considering the within-syllable parts only and without any compensation, in which the average recognition rates for Mandarin syllables were 43.94%. The recognition accuracies can be immediately increased to 49.24% using the cepstral bias removal techniques for channel noise compensation and further improved to 58.56% when the between-syllable context-dependent phone models are used. Furthermore, the achieved recognition rates were improved to as high as 59.53% using the minimum classification error algorithms as the post processing, which indicated a 27.81% of error rate reduction as compared to the baseline system.

This paper is organized into 5 sections. Section 2 describes the baseline recognition system and the speech database used in the experiments. The cepstral bias removal techniques are described in section 3. In section 4, the experiments based on different types of between-syllable context-dependent phone models are performed and discussed. Section 5 finally gives the concluding remarks.

2. Baseline Recognition System

2.1 Speech Database

The speech database was produced by 59 male and 54 female speakers over the telephone provided by Telecommunication Laboratories, Taiwan, Republic of China. Each speaker produced 120 Mandarin sentences such that a total of 13,560 Mandarin sentences (5.87 hrs) are included in the speech database. The signal-to-noise ratios (SNR) of this database are distributed from 10 to 40 dB, in which 9.09%, 56.36% and 34.55% of this database locate in 10~20 dB, 20~30 dB and 30~40 dB, respectively. In the following experiments, 51 male and 49 female speakers were used to train the gender-dependent, speaker-independent models and the rest 8 male and 5 female speakers were used as the testing speakers.

2.2 Front-end Processing

The telephone speech, which has a band of 150 Hz ~ 3.8 kHz, was sampled at an 8k Hz rate. After end-point detection is performed, 32 ms hamming window is applied every 10 ms with a pre-emphasis factor of 0.95. 14-order mel-frequency cepstral coefficients (MFCC) were derived from the power spectrum filtered by a set of 30 triangular band-pass filters. In addition, the first order derivatives of the 14 mel-frequency cepstral coefficients as well as the first and second order derivatives of the log short-time energy were also calculated to result in a feature vector of 30 dimensions for each frame [10].

2.3 Acoustic Modeling

The basic speech units used for recognition in this study are phone-like units (PLU) [11-12], in which a total of 34 context-independent (CI) PLU's are included. In fact, the most widely used units in the Mandarin speech recognition are the 22 Initial's and 40 Final's, where Initial means the initial consonant and Final means the vowel part but including possible media and nasal ending [10]. This is because of the mono-syllabic structure of the Mandarin Chinese, in which each Mandarin syllable can be decomposed into an Initial/Final format. One can note that each Initial is represented by one phoneme while each Final contains one to several phonemes. Accordingly, the numbers for the context-independent (CI) Initial/Final and PLU are 34 and 62, respectively. Also, when the right context dependency is considered, i.e., the speech units are regarded as different ones with respect to the beginning phonemes of the following units, the numbers for the right context dependent (RCD) Initial/Final and PLU can be expanded into 149 and 145, respectively. However, when the inter-syllable transitions are considered, the numbers for the RCD Initial/Final and PLU are immediately increased to 1269 and 480, respectively. Furthermore, if both the right and the left context dependencies are included, the numbers for Initial/Final and PLU will be further increased to 13,336 and 4605, respectively. One can find that the amount of Initial/Final units is nearly 3 times of that of phone-like units considering both the left and the right contextual effects. Because it is highly necessary to model the contextual acoustics and co-articulation in spontaneous telephone speech, we choose the PLU as the basic speech unit. The 3-state left-to-right continuous hidden Markov model (CHMM) [13] was trained for each PLU and the number of mixtures per state is dynamically determined by the amount of available training data with a maximum of 8 mixture components.

The block diagram of the training phase is shown in Fig. 1. The context-independent (CI) PLU based models are first obtained using the forward-backward algorithm, in which the initial model parameters were derived from uniform segmentation. Then the CI-PLU models were used as the initial seed models to derive the within-syllable CD-PLU models using the forward-backward algorithm. Furthermore, the between-syllable CD-PLU models can be trained using the within-syllable CD-PLU models as the initial models. Finally, the minimum classification error (MCE) algorithms are used for further enhancement of the discriminative capability of the between-syllable CD-PLU models.

2.4 Performance Baseline

This recognition process is based on the Viterbi search algorithm for obtaining the optimal Mandarin syllable sequence. Also, the recognition rates are evaluated as one minus substitution rates, insertion rates as well as deletion rates. In the baseline experiments, the within-syllable right-context-dependent (RCD) PLU's were used as the speech units. The average recognition rates for male and female testing speakers were 45.30% and 42.57% respectively as shown in the Table 1.

	male	female	average
Recognition rates(%)	45.30	42.57	43.94

Table 1 : The baseline experimental results using 145 within-syllable right-context-dependent phone-like units.

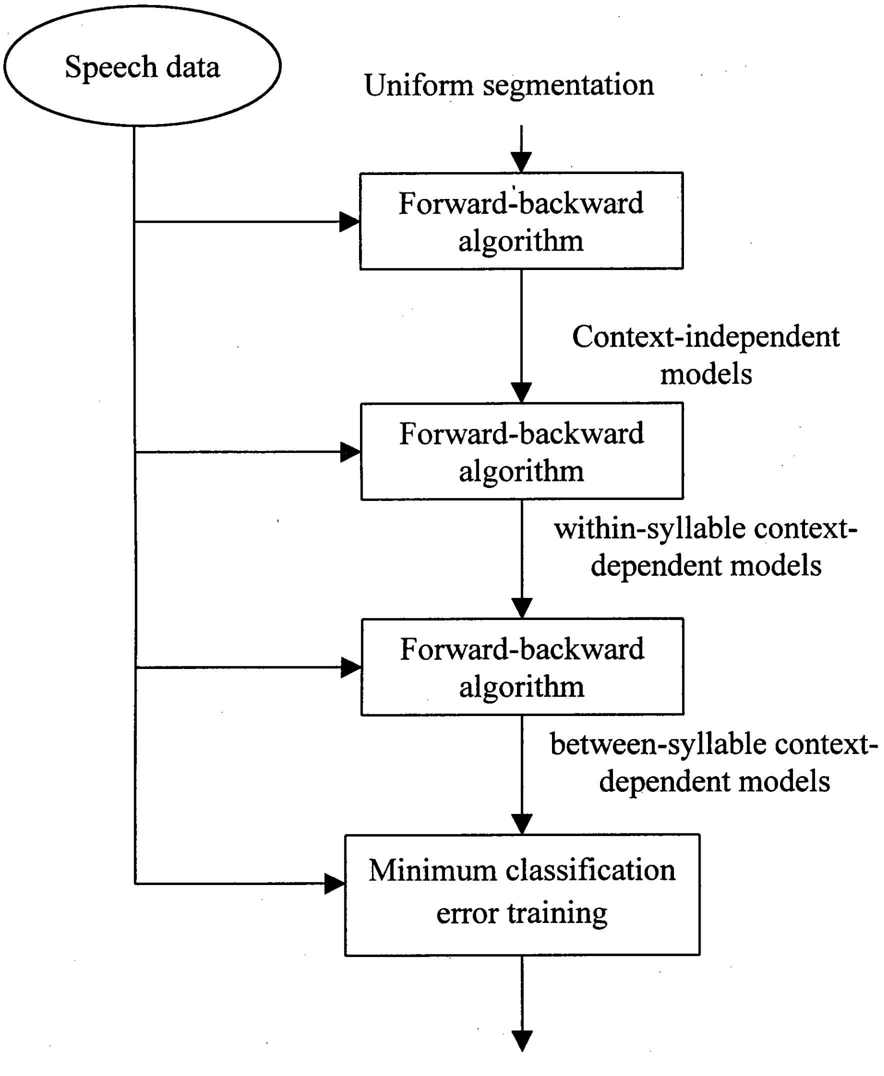
Speech data
Uniform segmentation
Forward-backward algorithm
Context-independent models
Forward-backward algorithm
within-syllable context-dependent models
Forward-backward algorithm
between-syllable context-dependent models
Minimum classification error training

Figure 1 : The block diagram of the training procedure.

3. Cepstral Bias Removal

As mentioned previously, the channel noise is convoluted with the clean speech signal in time domain and becomes additive in logarithmic spectral domain or cepstral domain. Therefore, the corrupted speech signal y can be represented by the bias transformation $y = x + h$, where y, x and h denote the cepstral representations for noisy speech, clean speech and channel noise, respectively. The cepstral bias removal techniques are thus developed to estimate the cepstral bias h and then subtract the bias from the noisy speech cepstral vectors. Three kinds of widely used cepstral bias removal techniques are discussed and improved in the following, including cepstral mean subtraction (CMS), signal bias removal (SBR) and stochastic matching (SM).

3.1 Cepstral Mean Subtraction (CMS)

In CMS [6], we make the assumptions that the cepstral mean of speech signal over a long time equals to zero such that the cepstral bias of channel noise can be estimated by long-time average of the noisy speech cepstral vectors.

$$h = \frac{1}{T}\sum_{t=1}^{T} y_t , \qquad (1)$$

where y_t means the noisy feature vector at frame t with a total of T frames. A few methods are investigated here for the estimation of the cepstral bias h in CMS, depending on the amount T of the speech data.

1. Global bias : A single bias vector is estimated with all of the available training speech data and shared by all of the training speakers.
2. Speaker-dependent bias : The bias vectors are estimated for each speaker separately such that a total of 100 bias vectors are obtained for all the 100 training speakers, respectively.
3. Sentence-dependent bias : Each sentence can obtain its individual bias vector for the compensation of the channel noise.

4. Sequential sentence-dependent bias : It is very often that the estimation of the cepstral bias is coarse on a sentence-by-sentence basis due to the insufficient length for an individual sentence. Therefore, the cepstral bias is sequentially obtained by the interpolation of the current estimate with the previous estimates.

The experimental results are shown in Table 2. One can find that the performance was even degraded using the global bias in CMS (43.03% vs. 43.94%). This is probably because the channel effects in telephone environments are almost constant for a given call but vary with calls such that a single bias can not represent the channel effect very well and even smears the speech signal characteristics. However, when the speaker-dependent cepstral biases are used, the average recognition rates can be improved from 43.94% to 48.86%, which indicates a 8.78% of error rate reduction. Also, the sentence-dependent bias estimation provides an average recognition rate of as high as 46.01%. It is apparent that the compensation due to the speaker-dependent bias outperforms that using the sentence-dependent bias. However, the sentence-dependent bias estimation is much more practical and feasible in real-world applications. Therefore, the sequential sentence-dependent bias estimation approach is developed to incrementally update the cepstral bias. It can be noted that comparable recognition rates with that using the speaker-dependent cepstral bias were achieved based on the sequential sentence-dependent bias estimation (48.74% vs. 48.86%).

	male	female	average
Global	44.53	41.52	43.03
Speaker-dependent	50.86	46.86	48.86
Sentence-dependent	48.78	43.23	46.01
Sequential sentence-dependent	51.01	46.46	48.74

Table 2. The experimental results using different cepstral bias estimation methods in cepstral mean subtraction (CMS).

3.2 Signal Bias Removal (SBR)

In SBR [7], a codebook Ω is first trained using all the available training data and the cepstral bias is obtained by maximizing the likelihood function $p(Y|\ h,\ \Omega)$, where Y means a set of noisy speech vectors $Y=\{y_1, y_2, \ldots, y_T\}$.

$$v_t = \arg\max_j p(y_t \mid h, \Omega_j), \tag{2}$$

$$h = \frac{1}{T}\sum_{t=1}^{T}(y_t - v_t) \tag{3}$$

where v_t designates the encoded codeword for the observation vector y_t at frame t. Apparently, CMS is a special case of SBR with the codebook size set to 1. In this study, three kinds of codebooks are developed, including *ad hoc* codebook, hierarchy codebook and phone-dependent codebook. In the *ad hoc* codebook, the codebook size is fixed and the codewords are trained using all the training speech based on the LBG algorithm, while in the hierarchy codebook, the codebook size is gradually increased such that the cpestral bias can be hierarchically updated using the codebook from smaller size to larger size. Instead of the data-driven codebook by vector quantization methods, the phone-dependent codebook is used, i.e., the training data corresponding to same context-independent PLU is clustered such that a total of 34 codewords cab be obtained.

On the other hand, in the encoding process, the soft decision is used for the estimation of the cepstral bias such that eq. (3) is expressed as below.

$$h = \frac{1}{T}\sum_{t=1}^{T}[\sum_{k=1}^{m} w_t^k (y_t - v_t^{\,k}) / \sum_{k=1}^{m} w_t^k] \tag{4}$$

where $v_t^{\,k}$ means the k-th nearest codeword for the observation vector y_t and $w_t^k = 1/\|y_t - v_t^{\,k}\|^2$ is the corresponding weighting factor.

Table 3 shows the experimental results using different types of codebook in SBR. It can be found that competitive recognition accuracies can be obtained using the *ad hoc* codebook with different sizes (46.79%, 46.48% and 46.26% for codebook size of 16, 32 and 64, respectively). In addition, when the hierarchy codebook is used where the codebook size is gradually increased from 16, 32 to 64, the recognition rates can be further improved to 47.35%. As shown in the last row of Table 3, the phone-dependent codebook can further provide slight improvement in recognition rates up to 47.50%. In Table 4, the encoding processes based on soft decision and hard decision are compared, in which the recognition rates can be further improved by 0.3%~0.5% using the soft decision for different types of codebook.

codebook type	codebook size	male	female	average
ad hoc	16	48.68	44.89	46.79
	32	48.73	44.22	46.48
	64	48.41	44.11	46.26
hierarchy	16,32,64	48.80	45.90	47.35
phone-dependent	34	49.33	45.67	47.50

Table 3. The experimental results using different types of codebook in signal bias removal (SBR).

codebook type	decision type	male	female	average
ad hoc(64)	hard	48.41	44.11	46.26
	soft	49.21	44.31	46.76
hierarchy	hard	48.80	45.90	47.35
	soft	49.41	45.96	47.69
phone-dependent	hard	49.33	45.67	47.50
	soft	49.81	46.04	47.93

Table 4. The comparative experimental results using hard decision and soft decision in encoding process in signal bias removal (SBR).

3.3 Stochastic Matching (SM)

In SM [8], the bias transformation function ($y = x + h$) is used to map the input corrupted speech onto the acoustic space of speech models such that the recognition process can be performed in matched conditions. The cepstral bias h can then be estimated in a maximum likelihood manner.

$$\begin{aligned} S^{(n+1)} &= \arg\max_{S} p(Y, S^{(n)} \mid h^{(n)}, \Lambda_X) \\ h^{(n+1)} &= \arg\max_{h} p(Y, S^{(n+1)} \mid h^{(n)}, \Lambda_X) p(S^{(n+1)}) \end{aligned} \tag{5}$$

where $S^{(n)}$ denotes the state sequence at the n-th iteration while Λ_X means the speech models. Suppose Λ_X is modeled by Gaussian distributions, the cepstral bias can be estimated in the following.

$$h = \frac{\sum_{t=1}^{T} \sum_{n} \sum_{m} \gamma_t(n,m) \Sigma_{n,m}^{-1} (y_t - \mu_{n,m})}{\sum_{t=1}^{T} \sum_{n} \sum_{m} \gamma_t(n,m) \Sigma_{n,m}^{-1}} \tag{6}$$

where $(\mu_{n,m}, \Sigma_{n,m})$ denotes the mean vector and covariance matrix of the speech models at state n and mixture m while $\gamma_t(n,m)$ means the corresponding posterior probability observing the feature vector y_t at frame t. In comparison with the formulations of the cepstral bias estimation in eqs. (3) and (6) basded on SBR and SM separately, we found that similar forms can be obtained, i.e., the weighting average of the difference between the noisy feature vectors and the corresponding centroids in the acoustic space of training data. However, the corresponding centroid for each observation vector comes from the speech models by Viterbi decoding in SM while in SBR it is obtained by the vector quantization process of a training codebook. In addition, because the cepstral bias is iteratively updated in the recognition process in the SM method, better initial estimate of the bias can provide better improvement of the performance. In other words, the SM method can be applied as the post processing after the CMS or SBR compensation is used. The block diagrams of the three kinds of cepstral bias removal techniques discussed in this section are shown in Fig. 2.

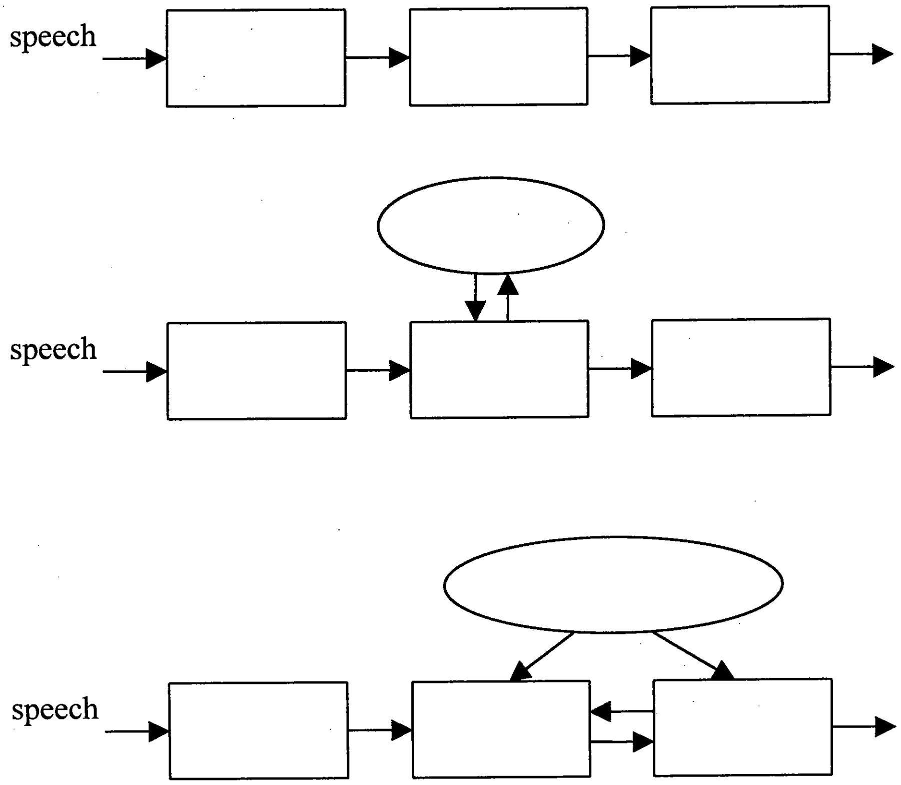

Table 5 shows the experimental results based on SM approach. Note that although the recognition rates can be increased from 43.94% to 45.56%, the improvements are indeed the least as compared to the CMS and SBR methods. This is probably due to the mis-classified labeling of the observation vectors in the model matching process. That is, the corresponding distribution $(\mu_{n,m}, \Sigma_{n,m})$ in eq. (6) for the feature vector y_t is probably incorrect. Therefore, better speech models can provide more correct labelling results and thus better estimation of the cepstral bias can be obtained. As shown in the last two rows of Table 5, when the SM is used as the post-processing after the CMS or SBR is performed, the performance can be further improved. The recognition accuracy using the combination of SBR and SM outperforms that using SBR only (47.93% vs. 47.48%) and so does CMS (49.24% vs. 48.86%).

	male	female	average	origin
SM	46.35	44.77	45.56	--
SBR+SM	49.62	46.23	47.93	47.48
CMS+SM	51.46	47.02	49.24	48.86

Table 5. The experimental results using different initial process in stochastic matching (SM).

4. Between-syllable Context-dependent Phone Models

4.1 Between-syllable Context-dependent Phone-like Units

In order to deal with the inter-syllable context variations for further improvement of continuous Mandarin speech recognition, the between-syllable triphone models are used. In other words, each speech model represents a phone with specific left and right contexts [14-15]. As mentioned previously, the number of triphones is 4605 for the 34-phone set, which is more than 30 times of that for the 145 within-syllable RCD phones used in the baseline system. Apparently, the trainability will become poor due to the insufficient amount of training data. In this study, we adopt two ways to increase the trainability using the triphone models.

1. Back-off : When the occurrence of a triphone unit in the training database is less than a pre-defined threshold, this triphone is replaced by its corresponding context-independent phone unit or context-dependent biphone unit considering left or right context dependency only.

2. Sharing : The triphone units are tied together by the linguistic constraints.

- Biphone : Unlike the triphone units that depend on both the right and the left context, the biphone units only depend on single context. Therefore, the right context-dependent (RCD) and left context-dependent (LCD) biphone units are used instead.

- Demiphone : Each demiphone unit can be divided into two sections where the right part is dependent on the right context while the left part depends on the left context, separately. In this way, the needed number of mixture components will not be increased if the number of state per phone model is unchanged [16].

The structures of the between-syllable context-dependent phone based hidden Markov models are shown in Fig. 3, including triphone, biphone and demiphone units. To further improve the discriminative capability of the speech models, the minimum classification error (MCE) algorithm can be used as the post-processing in the training procedure [9]. During the MCE training, the model parameters are iteratively adjusted in a maximum discriminability manner such that the recognition errors can be minimized for the training speech database.

model	male	female	average
Intra-LCD phone	48.21	38.65	43.43
Intra-RCD phone	51.53	46.75	49.19
Triphone	58.92	54.68	56.80
Inter-RCD phone	60.52	56.59	58.56
Inter-demiphone	59.20	54.88	57.04
Intra-RCD Initial/Final	52.92	48.55	50.74
Inter-RCD Initial/Final	59.41	51.51	55.46

Table 6. The experimental results based on different types of context-dependent speech units (intra- denotes within-syllable while inter- denotes between-syllable).

4.2 Experiments

In this subsection, we investigate the recognition performance based on different types of context-dependent phone-like speech units. Here the cepstral mean subtraction (CMS) technique based on speaker-dependent cepstral bias estimation discussed previously is used as the front-end robust processing. Also, an extra silence model is added for the improvement of the speech end-point detection. As shown in first two rows of Table 6, the recognition results using within-syllable left context-dependent (LCD) and right context-dependent (RCD) are compared. It can be found that the right contextual effects are more influential on the recognition accuracy than that of left contexts (49.19% vs. 43.43%). Also, slight improvement can be obtained with the addition of the silence model as compared to the result shown in the second row of Table 2 (48.86% vs. 49.19%). Then, when the triphone based models are used, the recognition rates can be immediately improved to 56.80%, in which the error rates are reduced by 14.98% with the expense of more than 30 times of mixture components as shown in Fig. 4. It is noted that there exist around 2600 unseen triphones out of 4605. Here the back-off method is applied using between- syllable RCD PLU's to predict the unseen triphones. When the biphone and demiphone units are further used to tie the states of the triphone based models, the needed mixture components can be reduced from 55,272 to 7,701 and 10,480 respectively as also shown in Fig. 4. The recognition accuracies are also improved from 56.80% to 58.56% and 57.04% respectively as listed in Table 6. In other words, the trainability as well as the sensitivity can be increased by sharing the parameters of the triphone models. As a comparison, the within-syllable and between-syllable RCD Initial/Final based models are trained and the results are also shown in Table 6. One can find that although the recognition rates using Initial/Final units outperform that using PLU's considering within-syllable right context variations only (50.74% vs. 49.19%), the error rates and needed mixture components are greatly increased when the between-syllable context dependency is included as also shown in Table 6 and Fig. 4. It is indicated that the error rates are reduced by 6.96% using less than one half of mixture components compared with between-syllable RCD PLU and Initial/Final based models. Finally, when the minimum classification error (MCE) training algorithm is applied to the most successful between-syllable RCD PLU based models, the recognition rates can be further improved from 58.56% to 59.53% as shown in Table 7. In comparison with the baseline system listed in Table 1, the recognition rates are increased from 43.94% to 59.53%, which indicates a 27.81% of error rate reduction.

Inter-RCD phone	male	female	average
ML	59.41	51.51	58.56
MCE	61.78	57.28	59.53

Table 7. The comparative results using ML and MCE training based on between-syllable RCD phone models.

3. Conclusion

This paper presents a study on speaker-independent continuous Mandarin speech recognition under telephone environments. The widely used cepstral bias removal techniques (CMS, SBR and SM) were first compared and improved. Then the between-syllable context-dependent phone models (triphones, biphones and demiphones) were trained. The minimum classification error (MCE) training algorithm was further applied. Experimental results showed that the achieved recognition rates can be improved from 43.94% to 59.53% as compared to the baseline system using within-syllable RCD phone models.

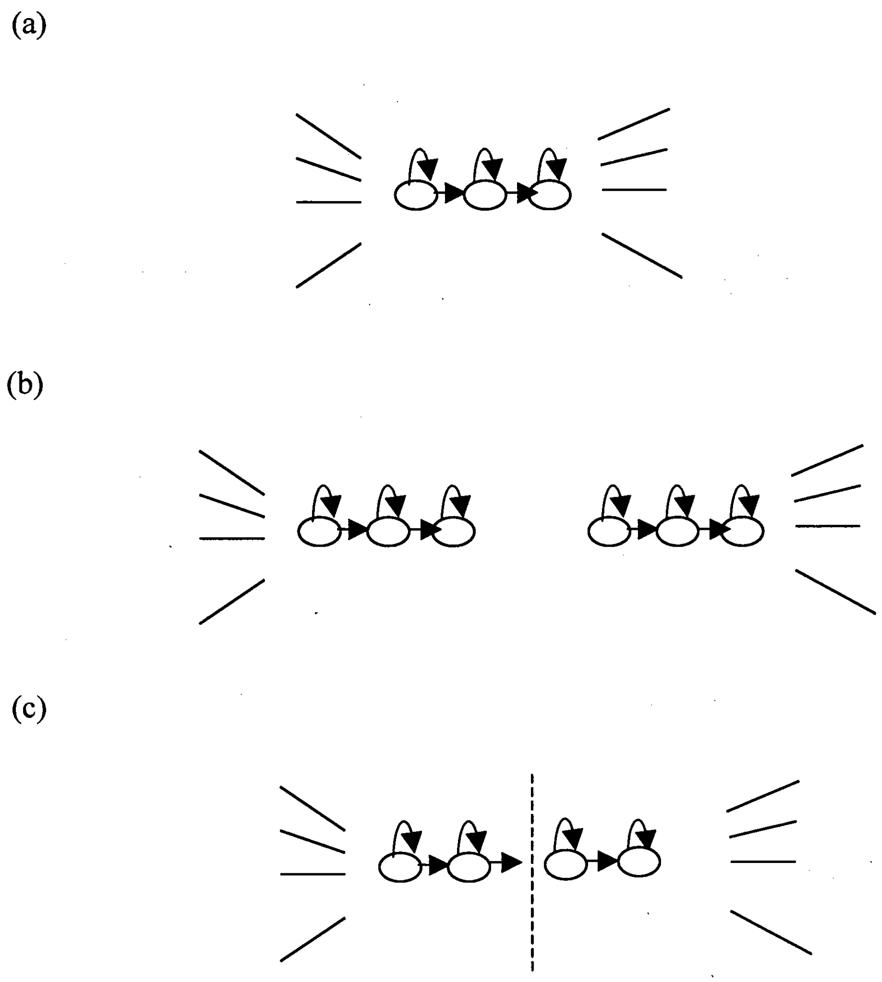

Figure 3. The structures of the between-syllable context-dependent phone based hidden Markov models (HMM) : (a). triphone, (b)., biphone and (c). demiphone.

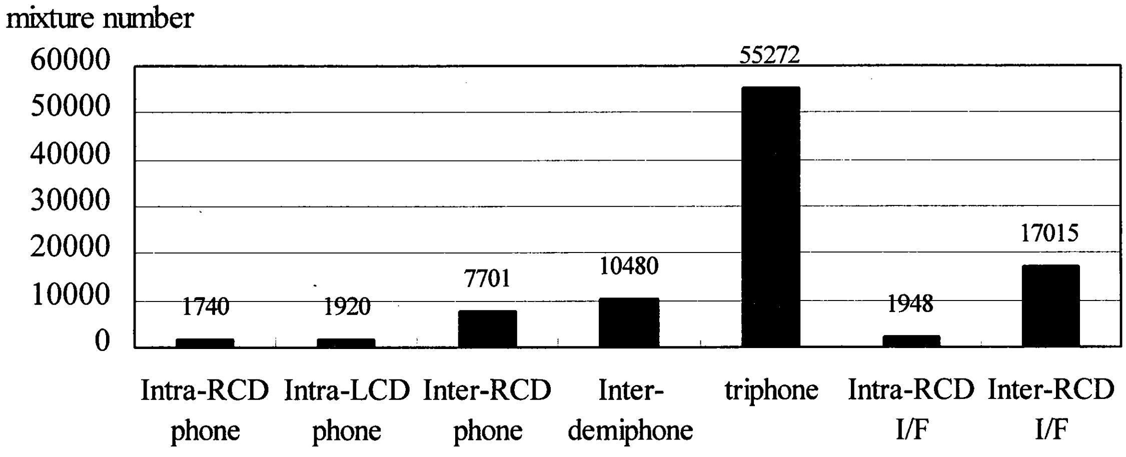

Figure 4. The total number of mixture components for the acoustic models based on different types of speech units.

References

1.R. Cole, *et.al.*, “The challenge of spoken language systems : research directions for the nineties”, *IEEE Trans. On Speech and Audio Processing*, Vol. 3, No. 1, Jan. 1995, pp. 1-21.

2.J. Takahashi, N. Sugarmura, T. Hirokawa, S. Sagayama & S. Furui, “Interactive voice technology development for telecommunication applications”, *Speech Communication*, 17:pp. 287-301, 1995.

3.D. Johnson, “Telephony based speech technology – from laboratory visions to customer applications”, *Journal of Speech Technology*, Vol. 2, No. 2, Dec. 1997, pp. 89-100.

4.C. Mokbel, D. Jouvet & J. Monne, “Deconvolution of telephone line effects for speech recognition”, *Speech Communication*, Vol. 19, pp. 185-196.

5.P.J. Moreno and R.M. Stern, “Source of degradation of speech recognition in the telephone network”, in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, 1994, pp. 109-112.

6.S. Furui, “Cepstral analysis technique for automatic speaker verification”, *IEEE Trans. Acoust. Speech, Signal Processing*, ASSP-29, Apr. 1981, pp. 254-272.

7.M.G. Rahim & B.H. Juang, “Signal bias removal by maximum likelihood estimation for robust telephone speech recognition”, *IEEE Trans. on Speech and Audio Processing*, Vol. 4, No. 1, pp. 19-30, Jan. 1996.

8.A. Sankar & C.H. Lee, “A maximum likelihood approach to stochastic matching for robust speech recognition”, *IEEE Trans. on Speech and Audio Processing*, Vol. 4, No. 3, pp. 190-202, May, 1996.

9.B.H. Juang, W. Chou, 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.

10. L.S. Lee, "Voice dictation of Mandarin Chinese", *IEEE Signal Processing Magazine*, Vol.

14, No. 4, pp. 63-101, July 1997.

11. R.Y. Lyu, H.M. Wang & L.S. Lee, "A comparison of different units applied to isolated/continuous large vocabulary Mandarin speech recognition", in *Proc. Int. Conf. Computer Processing of Oriental Language*, May 1994, pp. 211-214.

12. C.H. Lee & B.H. Juang, "A survey on automatic speech recognition with sn illustrative example on continuous speech recognition of Mandarin", *Computational Linguistics and Chinese Language Processing*, Vol. 1, No. 1, Aug. 1996, pp. 1-36.

13. L.R. Labiner, "A tutorial on hidden Markov models and selected applications in speech recognition", *Proc. IEEE,* 77(2) : 257-286, Feb. 1989.

14. K.F. Lee, "The SPHINX speech recognition system", in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, 1989, pp. 445-448.

15. J. J. Odell, "The use of context in large vocabulary speech recognition", *Ph.D. dissertation*, Queen's college, UK, Mar. 1995.

16. J.B. Marino, A. Nogueiras, A. Bonafonte, "The dimiphones : an efficient subword units for continuous speech recognition", *Int. Conf. Eurospeech*, pp. 1215-1218, 1997.