Computational Linguistics and Chinese Language Processing
Vol. 4, No. 1, February 1999, pp. 1-24
©Computational Linguistics Society of R.O.C.

Telephony Based Speaker-Independent Continuous Mandarin Syllable Recognition

Jia-Lin Shen*, Ying-Chieh Tu+, Po-Yu Liang+, Lin-Shan Lee+

Abstract

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

Keywords: Telephone speech recognition, Cepstral bias removal Between-syllable context-dependent phone models

1. Introduction

During the past few years, interest has increased in developing spoken dialogue systems for use over the telephone [1]. Apparently, good recognition performance under telephone environments is crucial for a successful spoken dialogue system [2-3]. However, many problems arise from high-quality microphone to telephone networks such that telephony based speech recognition is still a very challenging task for several reasons. First, speaker independence is highly desired in telephone environments. Secondly, the

* Institute of Information Science, Academia Sinica. E-mail: xshen@speech.ee.ntu.edu.tw

+ Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan, R.O.C.

environmental variabilities become much more serious due to channel distortions and the fairly high ambient background noise levels. Thirdly, 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.

It is well known that channel noise is usually convoluted with the speech signal in the time domain and becomes an additive term in the logarithmic spectral domain or cepstral domain. Therefore, channel noise can be compensated by subtracting a bias term from the noisy speech signal in the cepstral domain (called cepstral bias removal). Comparative studies on 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. Later, 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) were modeled. In addition, the minimum classification error (MCE) algorithms were further used to enhance the discriminating ability of the speech models [9-10].

Mandarin Chinese is a monosyllable-structured tonal language. There exist at least more than 100,000 commonly used words, and each word is composed of from one to several characters. Also, there exist more than 10,000 commonly used characters. However, all the Chinese characters are pronounced as monosyllables, and there exists a total of only 1345 different phonologically allowed syllables. Moreover, every syllable is assigned a tone, and the tone has lexical meaning. When the differences in tone are disregarded, these 1345 different "tonal syllables" are further reduced to 408 different "base syllables" (i.e., tone-independent syllable structures). Note that here, we use "tonal syllables" to indicate syllables with tones, but "base syllable" for those disregarding the tones. Because there exist only four lexical tones plus a neutral tone and the tones can be independently recognized using primarily pitch information, accurate recognition of all 408 Mandarin base syllables is believed to be the key problem in Mandarin speech recognition with a very large vocabulary [11].

Our baseline system is based on the context-dependent phone-like units (PLU), considering the within-syllable parts only and without any compensation, where the average recognition rate for Mandarin base syllables is 43.94%. The recognition accuracy 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 rate can be improved to as high as 59.53% using the minimum classification error algorithms in post processing, which results in a 27.81% of error rate reduction as compared to the baseline system.

This paper is organized into 6 sections. Section 2 describes the baseline recognition system and the speech database used in our experiments. The cepstral bias removal techniques are described in Section 3. In Section 4, the between-syllable context-dependent phone models are discussed. Then the minimum classification error (MCE) algorithms are described. In Section 5, experimental results are presented and discussed. Section 6 finally gives concluding remarks.

2. Baseline Recognition System

2.1 Speech Database

The speech database was produced by 59 male and 54 female speakers over a telephone provided by Telecommunication Laboratories, Taiwan, Republic of China. Each speaker produced 120 Mandarin sentences so that a total of 13,560 Mandarin sentences (5.87 hrs) were included in the speech database. The signal-to-noise ratios (SNR) of this database are distributed from 10 to 40 dB, and 9.09%, 56.36% and 34.55% of this database is located within 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 remaining 8 male and 5 female speakers were used as test speakers.

2.2 Front-end Processing

The telephone speech, which had a band of 150 Hz ~ 3.8 kHz, was sampled at an 8k Hz rate. After end-point detection was performed, a 32 ms hamming window was 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 calculated, resulting in a feature vector of 30 dimensions for each frame [11].

2.3 Acoustic Modeling

The basic speech units used for recognition in this study were phone-like units (PLU) [12-13], and a total of 34 context-independent (CI) PLUs was 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 endings [11]. This is because of the mono-syllabic structure of Mandarin Chinese, in which each Mandarin syllable can be decomposed into an Initial/Final format. One should note that each Initial is represented by one phoneme while each Final contains one to several phonemes. Therefore, the number of context-independent (CI) Initials/Finals and PLUs is 34 and 62, respectively. Also, when the right context dependency is considered, i.e., when the speech units are regarded as different with respect to the beginning phonemes of the following units, the number of right context dependent (RCD) Initials/Finals and PLUs can be expanded to 149 and 145, respectively. However, when inter-syllable context variations are considered, the number of RCD Initials/Finals and PLUs is immediately increased to 1269 and 480, respectively. Furthermore, if both the right and the left context dependencies are included, the number of Initials/Finals and PLUs is further increased to 13,336 and 4605, respectively. One can find that the number of Initial/Final units is nearly 3 times that of PLUs, considering both the left and the right contextual effects. Because it is very necessary to model contextual acoustics and co-articulations in spontaneous telephone speech, we use the PLU as the basic speech unit. Table 1 lists the 34 PLUs used in this paper and the corresponding phonetic alphabets for each Initial/Final of Mandarin Chinese in terms of the 34 PLUs. The 3-state left-to-right continuous density hidden Markov model (CHMM) [14] is trained for each PLU, and the number of mixtures per state is dynamically determined by the amount of training data with a maximum of 8 mixture components.

A 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 are derived from uniform segmentation. Then the CI-PLU models are used as initial seed models to estimate the within-syllable CD-PLU models using the forward-backward algorithm. Similarly, the between-syllable CD-PLU models are trained with the forward-backward algorithm using the within-syllable CD-PLU models as 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 Recognition Process

This recognition process is based on the one-pass Viterbi beam search algorithm with a fixed pruning threshold. The optimal Mandarin base syllable sequence is, therefore, decoded for each Mandarin utterance. Also, the recognition rates are evaluated as one minus substitution rates, insertion rates and deletion rates.

3. Cepstral Bias Removal

As mentioned previously, channel noise in a telephone environment is convoluted with the clean speech signal in the time domain and becomes additive in the logarithmic spectral domain or cepstral domain. Therefore, a corrupted noisy speech signal y can be represented by a bias transformation, i.e., $y = x + h$, 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 the estimated clean speech can be obtained by subtracting 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 assumption that the cepstral mean of a speech signal over a long period of time is equal to zero such that the cepstral bias of channel noise can be estimated by a long-term average of the noisy speech cepstral vectors :

$$h = \frac{1}{T}\sum_{t=1}^{T} y_t \quad , \tag{1}$$

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

1. Speaker-dependent bias : The bias vectors are estimated for each speaker separately such that a total of 100 bias vectors can be obtained for all 100 training speakers, respectively.

2. Sentence-dependent bias : Each sentence can obtain its individual bias vector for compensation of the channel noise.

3. Sequential sentence-dependent bias : It is very often the case 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 through interpolation of the current estimate with the previous estimates.

One can see that the estimation accuracy of the channel bias in CMS depends on both the phonetic variations and the speaker variations. In other words, the channel bias can be estimated more correctly by including sufficient phonetic variations in the noisy speech. However, the estimated channel bias may be smeared by speaker variations when it is obtained by all the speakers. In the database used in this paper, the data produced by a specific speaker is over the same handset so that the cepstral bias can be estimated more correctly using a large amount of speech data. This is why the speaker-dependent cepstral bias is utilized and compared. Then the sentence-dependent bias is used for practical speech recognition applications. Moreover, for the purpose of continuously improving the channel bias estimation to as good as the speaker-dependent bias, the sequential sentence-dependent bias is further developed.

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 $\{Y = y_1, y_2, \dots , y_T\}$:

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

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

where v_t designates the encoded codeword for the observation vector y_t at frame t. In this study, three kinds of codebooks are developed, including an *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 codebooks which are smaller to larger in size. This is because lower accuracy but higher robustness for the estimated channel bias can be provided using a smaller codebook. Therefore, we try to hierarchically update the channel bias by gradually increasing the size of the codebook. Instead of a data-driven codebook obtained by vector quantization methods, a phone-dependent codebook is used; i.e., the training data corresponding to the same context-independent PLU is clustered such that a total of 34 codewords can be obtained.

On the other hand, the channel bias can be estimated more accurately by including several neighborhood codewords. In other words, in the encoding process, a soft decision is used for estimating of the cepstral bias such that eq. (3) is expressed as follows.

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

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

Apparently, CMS is a special case of SBR with the codebook size set to 1. In fact, the SBR technique can be regarded as a blind equalization by minimizing the distortion between the speech data and a codebook, where the channel bias is estimated in a *minimum mean square error* sense while in CMS, the channel bias is estimated in a *maximum likelihood* manner. In fact, the robustness achieved using SBR is usually superior to that obtained using CMS when only short-duration speech signals are given, e.g., one Chinese sentence. This is probably because the poor phonetic variations in short-duration speech signals can be compensated by including a codebook in the SBR technique.

3.3 Stochastic Matching (SM)

In SM [8], the bias transformation function ($y = x + h$) is used to map corrupted input 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 [8]:

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

where $S^{(n)}$ denotes the state sequence at the n-th iteration while Λ_X means the speech models. This is because the reference space used in SM is the speech models instead of the codebook used in SBR, so that the state sequence for any speech utterance must be decoded first using the Viterbi search algorithm. Suppose Λ_X is modeled by means of Gaussian distributions; the cepstral bias can be estimated as follows.

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

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. Considering the formulations of the cepstral bias estimation in eqs. (3) and (6) based on SBR and SM separately, we find 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 means of Viterbi decoding in SM while in SBR it is obtained by vector quantization of a pre-training codebook. In addition, because the cepstral bias is iteratively updated in the recognition process in the SM technique, a better initial estimate of the bias can provide more improvement in performance. In other words, the SM method can be applied as post processing for further improvement of environmental robustness after CMS or SBR compensation is used. Block diagrams of the three cepstral bias removal techniques discussed in this section are shown in Fig. 2.

4. Context-Dependent Phonetic Models

4.1 Between-syllable Context-dependent Phone-like Units

In order to deal with the inter-syllable context variations so as to obtain further improvement in 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 [15-17]. In fact, there exists a trade-off between sensitivity and trainability for speech models based on different speech units with different contextual variations. In general, the sensitivity of a speech model increases when detailed speech units are used if they accurately represent the data. However, a limited amount of training data implies poor trainability when detailed speech models are used. As mentioned previously, the number of triphones is 4605 for the 34-phone set, which is more than 30 times the number for the within-syllable RCD phones used in the baseline system. It is noteworthy that nearly 2600 triphone units out of 4605 do not occur in our speech database mentioned in Section 2 Apparently, the trainability will become much poorer due to an insufficient amount of training data. In this study, we adopt two methods to increase the trainability using triphone models [18-19].

1. Back-off : When the number of occurrences 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 while considering left or right context dependency only.
2. Sharing : The triphone units are tied together by linguistic constraints.

● Biphone : Unlike triphone units that depend on both the right and the left context, biphone units only depend on a single context. Therefore, the right context-dependent (RCD) and left context-dependent (LCD) biphone units are used instead [20].

● 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 number of mixture components needed will not increase if the number of state per phone models is unchanged [21].

The structures of the context-dependent phone based hidden Markov models are shown in Fig. 3, including triphone, biphone and demiphone units. Moreover, both syllable internal context dependency (within-syllable) and cross syllable context dependency (between-syllable) are considered in this study.

4.2 Minimum Classification Error (MCE) Algorithm

To further improve the discriminative capability of the speech models, the minimum classification error (MCE) algorithm can be used as post-processing in the training procedure [9]. During MCE training, the model parameters are iteratively adjusted in a maximum discriminability manner such that the number of recognition errors can be minimized for the training speech database.

In MCE estimation, a loss function is first defined [10]:

$$L(y,\lambda) = \sum_t l(d(y_t,\lambda)), \tag{7}$$

where $d(y_t,\lambda)$ denotes the misclassification measure for the cepstral feature y_t observed in acoustic model λ at time t while $l(d(y_t,\lambda))$ is the individual loss, which is represented by the sigmoid function

$$l(d(y_t,\lambda)) = 1/(1+e^{-\beta d(y_t,\lambda)}), \tag{8}$$

where β is a pre-determined constant scaling value. The misclassification function can be defined as :

$$d(y_t,\lambda) = -\log\ p(y_t \mid \lambda_c) + \log\{\tfrac{1}{K} \sum_{j,j\neq c} p(y_t \mid \lambda_j)^{\eta}\}^{1/\eta} \quad , (9)$$

where *c* denotes the correct model index and parameter η is the weighting factor for adjusting the contribution of discriminant functions made by other models. In addition, *K* denotes the total number of mis-classified models used in MCE training. For simplicity, $d(y_t,\lambda)$ can be expressed as follows when η approaches ∞ :

$$d(y_t,\lambda) = -\log(\ p(y_t \mid \lambda_c) + \log(\ p(y_t \mid \lambda p)\ , \quad (10)$$

where λ_{p} designates the most mis-classified model. Accordingly, the parameters of the speech models can be re-estimated using the gradient descent scheme :

$$\lambda^{n+1} = \lambda^{n} - \varepsilon(n)\frac{\partial\ L(y,\lambda^{n})}{\partial\ \lambda^{n}} \quad , \quad (11)$$

where ε *(n)* is the step size used for adjustment with n being the iteration index. In this way, the parameters of the speech models can be refined by minimizing the recognition errors for all the training speech such that the discriminating capabilities of the speech models can be enhanced.

5. Experimental Results

5.1 Baseline Performance

In the baseline experiments, the within-syllable right-context-dependent (RCD) PLUs were used as the speech units. The gender-dependent, speaker-independent speech models were trained without any compensation in telephone channel effects. The average recognition rates for male and female testing speakers were 45.30% and 42.57%, respectively, as shown in Table 2.

5.2 Cepstral Bias Removal

The experimental results obtained using CMS are shown in Table 3. One can find that when the speaker-dependent cepstral biases were used, the average recognition rates could be improved from 43.94% to 48.86%, which indicates an 8.78% error rate reduction. Also, the sentence-dependent bias estimation provides an average recognition rate of as high as 46.01%. This demonstrates the great effect of speaker variations on the estimation accuracy of the telephone channel bias. It is apparent that the compensation done using speaker-dependent bias outperforms that done using sentence-dependent bias due to increased coverage of phonetic variations. However, sentence-dependent bias estimation is much more practical and feasible in real-world applications. Furthermore, the sequential sentence-dependent bias estimation approach is used to incrementally update the cepstral bias. It can be seen that recognition rates comparable with that obtained using the speaker-dependent cepstral bias were achieved based on the sequential sentence-dependent bias estimate (48.74% vs. 48.86%). As a comparison, a single global bias was estimated by all speech database, leading to the recognition rate of 43.03%. Apparently, the results were even worse than that obtained without any compensation (43.94% as shown in Table 2). This is probably because the channel effects in telephone environments are almost constant for a given call made by a specific speaker but vary with different calls so that a single bias can not represent the channel effect very well and even smears the speech signal characteristics.

Then we performed experiments using the SBR technique, in which the cepstral bias is estimated sentence-by-sentence. Table 4 shows the experimental results obtained using different types of codebooks in SBR. It can be found that competitive recognition accuracy can be obtained using the *ad hoc* codebook with different sizes (46.79%, 46.48% and 46.26% for a codebook size of 16, 32 and 64, respectively). Note that SBR compensation based on a small codebook size even slightly outperforms that based on a large codebook size. 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%. Therefore, the combination of different codebook sizes leads to further improvement. In addition to the data-driven codebooks used in the above experiments, the phonetic codebook was also tested. As shown in the last row of Table 3, the CI phone-dependent codebook with a codebook size of 34 can further provide slight improvement in the recognition rates up to 47.50%. Furthermore, the encoding processes based on the soft and the hard decisions are compared in Table 5, which shows that the recognition rates can be further improved by 0.3%~0.5% using the soft decision for different types of codebooks.

Table 6 shows the experimental results obtained using SM approach. Note that although the recognition rates could be increased from 43.94% to 45.56%, the improvement was indeed the smallest compared to the CMS and SBR results also shown in Table 6. This is probably due to the mis-classified labeling of the observation vectors in the model matching process as indicated in Fig. 2 (c). That is, the corresponding distribution $(\mu_{n,m}, \Sigma_{n,m})$ in eq. (6) fot the feature vector y_t was probably incorrect. Note that better speech models could provide more correct labelling results; thus, better estimation of the cepstral bias could be obtained. In other words, we can use CMS or SBR to derive better labelling results and then apply SM to further improve the environmental robustness. As shown in the last two rows of Table 6, when SM was used as post-processing after CMS or SBR was used, the performance could be further improved. The recognition accuracy obtained using a combination of SBR and SM outperformed that obtained using SBR only (47.93% vs. 47.48%) or CMS only(49.24% vs. 48.86%). Although pre-processing using CMS or SBR could reduce the degree of incorrect labeling in the first model matching process, one may wonder what the result could be if the correct labeling was given. The average experimental result was 47.80% if CMS and SBR were not applied. Note that the result is slightly better than that obtained using SBR (47.48% vs. 47.80%), but worse than that obtained by combining SBR and SM (47.93% vs. 47.80%). In fact, it can be seen as a kind of SBR where the speech models are used as a codebook. Therefore, using SBR or CMS as pre-processing not only delivers more correct labeling, but also provides a good initial cepstral bias estimate.

5.3 Context Dependent Phonetic Models

In this subsection, we will investigate the recognition performance obtained 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 was used as the front-end robust processing. Also, an extra silence model was added to improve speech end-point detection. In the first two rows of Table 7, the recognition results obtained using the within-syllable left-context-dependent (LCD) and right-context-dependent (RCD) phonetic models are compared. It can be found that the right contextual effects influence the recognition accuracy more than the left contexts do (49.19% vs. 43.43%). Also, slight improvement was obtained with the addition of the silence model (48.86% vs. 49.19%). When the triphone based models were used, the recognition rates can be immediately improved to 56.80%, and the error rate was reduced by 14.98% with the expense of more than 30 times 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 was applied using corresponding between-syllable RCD PLUs to predict the unseen triphones. Instead, the sharing method was used to tie the triphone models together using linguistic constraints. It can be found from Fig. 4 that when the biphone and demiphone units were used to replace the triphone units, the number of mixture components needed was significantly reduced from 55,272 to 7,701 and 10,480, respectively. The recognition accuracy also improved from 56.80% to 58.56% and

57.04%, respectively, as shown in Table 7. In other words, the trainability as well as sensitivity could be improved by sharing the parameters of the triphone models based on the speech database used in this study.

As a comparison, the within-syllable and between-syllable RCD Initial/Final based models were trained, and the results are also shown at the bottom of Table 7. One can find that the recognition rate obtained using Initial/Final units was better than that obtained using the PLUs listed in the third row of Table 7 when within-syllable right context variations were considered only (50.74% vs. 49.19%). However, the recognition rate degraded and the number of mixture components needed greatly increased when between-syllable context dependency was included, as also shown in Table 7 (55.46% vs. 58.56%) and Fig. 4 (17015 vs. 7701). This indicated that the error rate was reduced by 7.48% using less than one-half of mixture components as compared with the results obtained using the between-syllable RCD PLU and Initial/Final based models.

Finally, when the minimum classification error (MCE) training algorithm was applied to the most successful between-syllable RCD PLU based models, the recognition rate could be further improved from 58.56% to 59.53% as shown in Table 8. In comparison with the baseline system listed in Table 2, the recognition rate increased from 43.94% to 59.53%, which indicates a 27.81% error rate reduction.

6. Conclusion

This paper has presented a study on speaker-independent continuous Mandarin syllable recognition under telephone environments. The widely used cepstral bias removal techniques were first compared and improved, including the cepstral mean subtraction (CMS), signal bias removal (SBR) and stochastic matching (SM) techniques. Then the context-dependent phonetic models (triphones, biphones and demiphones) were trained while considering both syllable internal context dependency and cross syllable context dependency. The minimum classification error (MCE) algorithms were further applied to enhance the discriminability of the speech models. Experimental results show that the achieved recognition rates could be improved from 43.94% to 59.53% as compared to the baseline system using the within-syllable RCD phone models.

References

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.

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

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.

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

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.

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

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.

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.

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.

W. Chou, B. H. Juang, C. H. Lee, "Segmental GPD Training Of HMM Based Speech Recognizer", in *Proc. Int. Conf. Acoustics, Speech, Signal Processing.*, pp. 473-476, 1992.

L.S. Lee, "Voice dictation of Mandarin Chinese", *IEEE Signal Processing Magazine*, Vol. 14, No. 4, pp. 63-101, July 1997.

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.

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

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

R.M. Schwartz, Y.L. Chow, S. Roucos, M. Krasner, J. Makhoul, "Improved hidden Markov modeling of phonemes for continuous speech recognition", in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, 1984.

K.F. Lee, "The SPHINX speech recognition system", in *Proc. Int. Conf. Acoustics, Speech, Signal*

Processing, 1989, pp. 445-448.

H.W. Hon, K.F. Lee, "Recent progress in robust vocabulary-independent speech recognition", *Proceeding DARPA Speech and Natural Lcnguage Processing Workshop*, Austin, pp. 170-177, 1995.

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

S.J. Young, "The general use of tying in phoneme-based HMM speech recognizers", in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, 1992, pp. 569-572.

M.Y. Hwang, H.W. Hon, K.F. Lee, "Modeling between-word coarticulation in continuous speech recognition", in *Proc. Eurospeech*, pp. 5-8, 1989.

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

Table 1. *(a). 34 phone-like units (PLUs). (b). The corresponding phonetic alphabets with respect to 22 Initials/40 Finals of Mandarin Chinese in terms of 34 phonemes.*

(a)

b	p	m	f	d	t	n	l	g	k	h	j	<	T	Z	C	S
R	z	c	s	a	o	e	E	i	u	U	r	M	N	#	Y	y

(b)

ㄅ	ㄆ	ㄇ	ㄈ	ㄉ	ㄊ	ㄋ	ㄌ	ㄍ	ㄎ	ㄏ	ㄐ	ㄑ	ㄒ	ㄓ	ㄔ	ㄕ	ㄖ	ㄗ	ㄘ	ㄙ
b	p	m	f	d	t	n	l	g	k	h	j	<	T	ZY	CY	SY	RY	zy	cy	sy
ㄚ	ㄛ	ㄜ	ㄝ	ㄞ	ㄟ	ㄠ	ㄡ	ㄢ	ㄣ	ㄤ	ㄥ	ㄧ	ㄨ	ㄩ	ㄧㄚ	ㄧㄝ	ㄧㄞ	ㄧㄠ	ㄧㄡ	ㄧㄢ
#a	#o	#e	#E	#ai	#Ei	#au	#ou	#aM	#M	#aN	#N	#i	#u	#U	#ia	#iE	#iai	#iau	#iou	#iaM
ㄧㄣ	ㄧㄤ	ㄧㄥ	ㄧㄛ	ㄨㄚ	ㄨㄛ	ㄨㄞ	ㄨㄟ	ㄨㄢ	ㄨㄣ	ㄨㄤ	ㄨㄥ	ㄩㄝ	ㄩㄢ	ㄩㄣ	ㄩㄥ	ㄦ	NULL Initial	NULL Final 1	NULL Final 2	
#iM	#iaN	#iN	#io	#ua	#uo	#uai	#uEi	#uaM	#uM	#uaN	#ueN	#UE	#UaM	#UM	#ioN	r	#	Y	y	

Table 2. *The baseline experimental results obtained using intra-syllableright-context-dependent phone-like units.*

	male	female	average
Recognition rates	45.30	42.57	43.94

Table 3. *The experimental results obtained using different cepstral bias estimation methods in cepstral mean subtraction (CMS).*

	male	female	average
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 4. *The experimental results obtained using different types of codebooks in signal bias removal (SBR).*

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 5. *The comparative experimental results obtained using the hard decision and the soft decision in the encoding process 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 6. *The experimental results obtained using different initial processes in stochastic matching (SM).*

	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 7. *The experimental results obtained based on different types of context-dependent speech units (intra- denotes a within-syllable while inter- denotes a between-syllable).*

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 8. *The comparative results obtained using ML and MCE training based on between-syllable RCD phone models.*

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

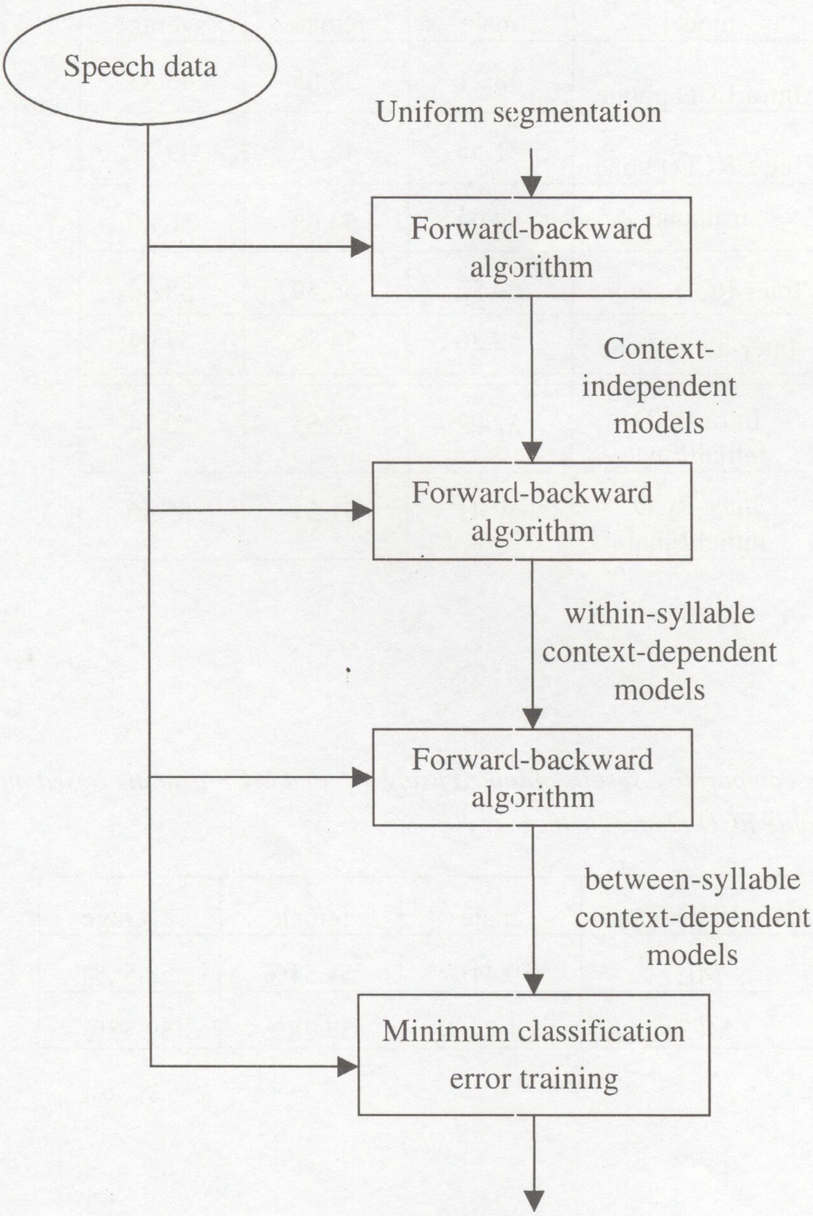
Speech data
Uniform segmentation
Forward-backward algorithm
Context-independent models
Forward-backward algorithm
within-syllable context-dependent models
Forward-backward algorithm
between-syllable context-dependent models
Minimum classification error training

***Figure 1** A block diagram of the training procedure.*

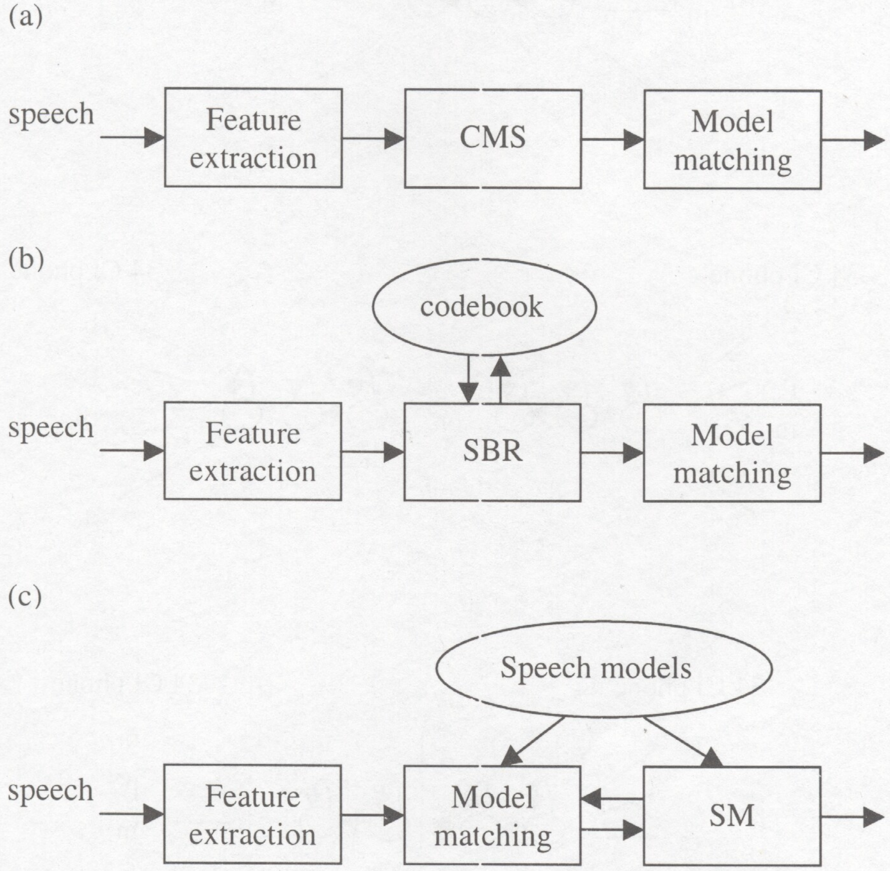

***Figure 2** Block diagrams of the three cepstral bias removal techniques. (a). cepstral mean subtraction (CMS), (b). signal bias removal (SBR), and (c). stochastic matching (SM).*

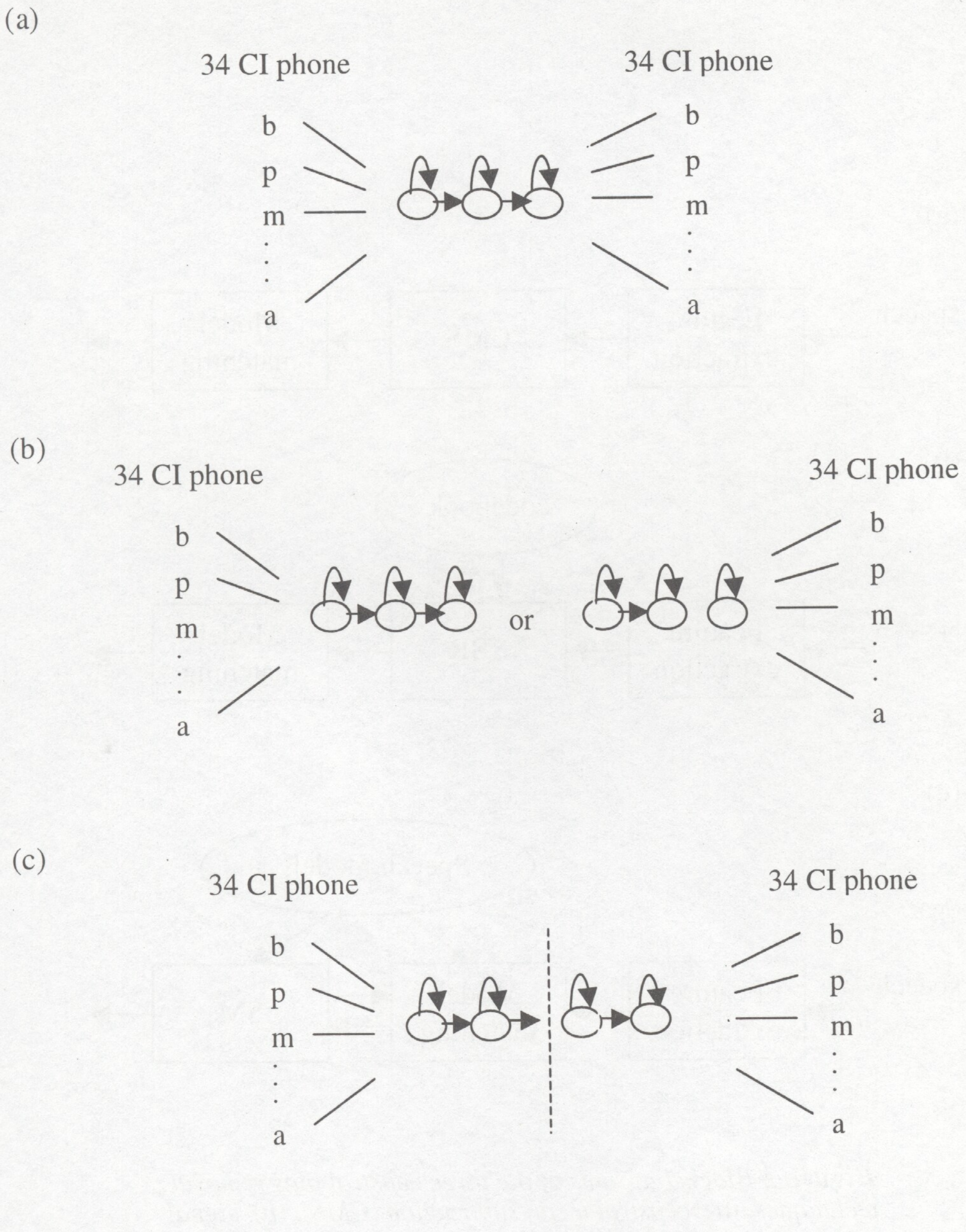

***Figure 3** The structures of the context-dependent phone-based hidden Markov models (HMM) : (a). triphone, (b). biphone and (c). demiphone.*

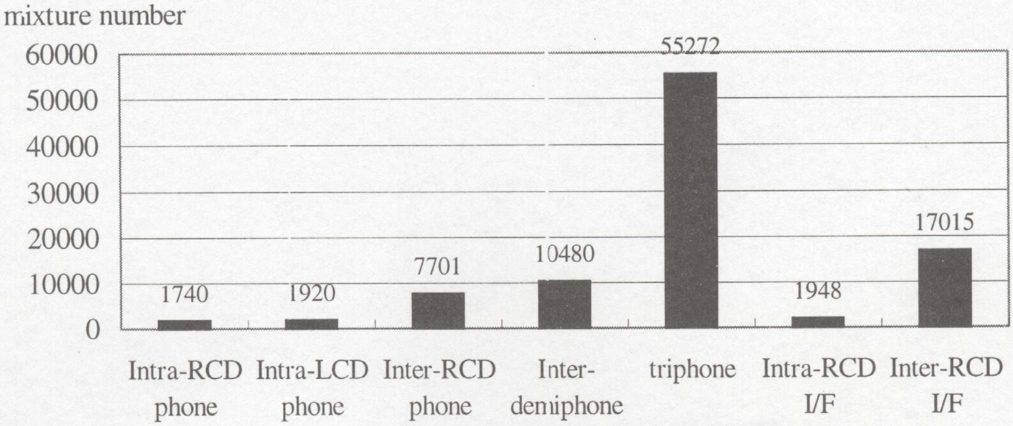

Figure 4 *The total number of mixture components for the acoustic models based on different types of speech units.*