A SEGMENT-BASED C_0 ADAPTATION SCHEME FOR PMC-BASED NOISY MANDARIN SPEECH RECOGNITION

Wei-Tyng Hong and Sin-Horng Chen
Department of Communication Engineering,
National Chiao Tung University, Hsinchu, Taiwan.
E-Mail: schen@cc.nctu.edu.tw

ABSTRACT

A segment-based C_0 (the zero-th order of cepstral coefficient) adaptation scheme for PMC-based Mandarin speech recognition is proposed in this paper. It incorporates a new C_0 model of speech signal into the PMC method to improve the gain matching between the clean-speech HMM models and the current noise model. The C_0 model is constructed in the training phase by jointly modeling the normalized C_0 with other MFCC recognition features to form C_0-normalized HMM models. In the testing phase, it pre-segments the input utterance into syllable-like segments, performs C_0-denormaliztion operations to expand the C_0-normalized HMM models, and uses them in the PMC method. Compared with the conventional PMC method, the proposed method can achieve a much better noise compensation effect due to the use of more precise gain matching in the PMC model combination. Experimental results showed that the base-syllable accuracy rate was significantly upgraded for continuous noisy Mandarin speech recognition.

1. INTRODUCTION

Performances of speech recognizers, trained in clean speech databases, usually degrade seriously when operating in noisy environments. Many methods have been proposed in recent years [1] to make speech recognizers robust to various corrupting noises. Among them, the parallel model combination (PMC) method is a promising one [2,3]. The basic idea of the PMC method is to use noise-compensated HMM models, generated by combining the clean-speech HMM models with the current noise model, to recognize the input testing utterance. The mismatch on acoustic characteristics between the testing utterance and the reference HMM models can hence be compensated. One serious drawback of the conventional PMC method lies on the use of a rough gain factor for the entire testing utterance in the model-combination operation to adjust the energy level of the clean-speech HMM models in order to match the SNR of the testing utterance. Although some previous studies have reported that the recognition performance is not sensitive to the rough gain estimate [2], a precise gain estimate based on word- or syllable-like segment should be beneficial to the PMC method for the generation of more precise noise-compensated HMM models to deal with a continuous testing utterance in which energy level varies from syllable to syllable. This is obviously true because the model combination operation is performed in linear spectrum domain. A simple example to demonstrate the importance of gain factor in the model-combination operation is given as follows. Assume that a noisy speech signal $y(t)$ is generated from the combination of a clean speech signal $x(t)$ and a corrupting additive noise $n(t)$ by

$$y(t) = gx(t) + n(t)$$

where g is the volume of speech signal which controls the SNR level. Fig. 1(b) displays four plots of the 12 MFCC features of $y(t)$ obtained by combining the same $x(t)$ and $n(t)$ using four different values of g. It can be found from Fig. 1 that the resulting cepstrum of $y(t)$ is very sensitive to the value of g. It is resemble to the cepstrum of $x(t)$ for large g and to that of $n(t)$ for small g. So, the conventional PMC method, which uses a rough gain factor for an entire sentential utterance, will result in under noise-compensations (via using too large g) for consonant frames located in the ending part of the utterance and in over noise-compensations (via using too small g) for vowel frames in the beginning part.

To cue this drawback, we propose a segment-based C_0 adaptation scheme in this paper to improve the efficiency of the PMC method on recognizing noisy Mandarin speech. The new recognition method is referred to as the segment-based, C_0-adapted PMC (SCA-PMC) method. The segment-based C_0 adaptation scheme is implemented based on a new C_0 model of speech signal. Intuitively, a direct C_0 modeling method is to jointly model it with other MFCC recognition features (i.e., C_1-C_p) using a mixture Gaussian distribution for each HMM state. But this will make the resulting HMM model unsuitable for a testing environment with recording volume-control level different from that of the training environment. To avoid this defect, a new C_0 modeling method is proposed in this study. It modifies the direct C_0 modeling method by embedding a C_0 normalization operation in the segmental k-means training algorithm with a goal to generate a set of compact, C_0-normalized HMM models in the training phase. In the testing phase, it applies a C_0 de-normalization operation to expand these C_0-normalized HMM models so as to make their C_0's match with the time-varying volume of the input noisy speech signal. The C_0 normalization operation is realized by first subtracting the floor level of C_0, $C_{0,offset}$, from the C_0 value of each frame, and by then linearly scaling it with respect to its local maximum in the syllable-segment to which the current frame belongs. $C_{0,offset}$ is resulted from the background noise of the recording system in silence condition. Two schemes of modeling the normalized C_0 are suggested in this study. One is to model it with the averaged value for each HMM state. The other is to extend the first scheme by further jointly modeling the residue of the normalized C_0, with respect to its averaged value, with all other MFCC features using a mixture Gaussian distribution for each HMM state. The de-normalization operation is realized by first performing an RNN-based pre-segmentation [4,5] to divide the input testing utterance into syllable-like segments, and by then calculating local C_0 maxima for these syllable-like segments. Finally, all C_0 models are de-normalized with each of these local C_0 maxima and used in the PMC method to generate a set of

noise-compensated HMM models to be used to recognize a part of the input speech surrounding the current syllable-like segment.

The remainder of the paper is organized as follows. Section 2 presents the proposed SCA-PMC noisy speech recognition method. The new C_0 model of speech signal is described in detail. Performance of the SCA-PMC method was examined by simulations on a continuous Mandarin speech recognition task and is discussed in Section 3. Some conclusions are given in the last section.

2. THE SCA-PMC METHOD

The proposed SCA-PMC noisy speech recognition method is composed of two phases: a training phase and a testing phase. The job of the training phase is to generate a set of C_0-normalized HMM models for speech signals. Fig. 2 shows a flow chart of the training algorithm. It differs from the conventional segmental k-means training algorithm mainly on inserting an additional C_0 normalization operation between the speech segmenting operation and the model updating operation. The C_0 normalization operation consists of two steps. The first one is to find the maximum of C_0 for each syllable segment. The second step is to normalize C_0 value of each frame by

$$C'_{0,t} = \frac{C_{0,t} - C_{0,offset}}{C_{0,\max} - C_{0,offset}}$$

where $C_{0,t}$ denotes the C_0 value of frame t, $C_{0,\max}$ is the C_0 maximum of the syllable segment which includes frame t, and $C_{0,offset}$ is the floor value of C_0 which represents the offset of the recording device in the silence condition. Two schemes of C_0 modeling for constructing the C_0-normalized HMM models are suggested in this study. One is to model $C'_{0,t}$ by its averaged value $\overline{C'}_{0,j}$ for state j of an HMM. The other is to combine the residue, $C'_{0,t} - \overline{C'}_{0,j}$, with other MFCC features (i.e., C_1-C_p) and to jointly model them by a mixture Gaussian distribution for each HMM state.

The job of the testing phase is to expand these C_0-normalized HMM models and to use them in the PMC method to recognize the input testing utterance. Fig. 3 displays a block diagram of the testing phase of the SCA-PMC method. It first pre-segment each input testing utterance into syllable-like segment by using an RNN-based broad-class discriminator and an finite-state-machine (FSM) [5]. The function of the RNN-based broad-class discriminator is to discriminate each input frame among three broad-classes of *final*, *initial*, and silence. Outputs of the RNN are then used to drive the FSM to divide the input utterance into segments of four types: *final*, *initial*, silence, and transition. It is noted that each *final*-segment coincides roughly with the *final* of a Mandarin syllable. We then find $C^y_{0,\max}$ of the noisy speech *y(t)* for each *final*-segment. Then, the $C^x_{0,\max}$ of the (clean) speech *x(t)* is estimated from $C^y_{0,\max}$ by using the noise masking method [6]. Then, the C_0 de-normalization operation is performed to expand these C_0-normalized HMM models by

$$\mu^{cep}_{x_{j,k},w}(m) = \begin{cases} C_{0,offset} + (C^x_{0,\max,w} - C_{0,offset})(\overline{C'}_{0,j} + \mu^{cep}_{\tilde{x}_{j,k}}(0)), & m = 0 \\ \mu^{cep}_{\tilde{x}_{j,k}}(\mathrm{m}), \quad m = 1, \cdots, p \end{cases}$$

and

$$\Sigma^{cep}_{x_{j,k},w}(m,n) = \begin{cases} (C^x_{0,\max,w} - C_{0,offset})^2 \Sigma^{cep}_{\tilde{x}_j}(0,0) \ \text{for} \ m = 0, n = 0 \\ \Sigma^{cep}_{\tilde{x}_{j,k}}(\mathrm{m,n}) \quad others \end{cases}$$

where $\mu^{cep}_{\tilde{x}_{j,k}}$ and $\Sigma^{cep}_{\tilde{x}_{j,k}}$ are the mean vector and the covariance matrix of the k-th mixture component in state j of an HMM model, $\mu^{cep}_{x_{j,k},w}$ and $\Sigma^{cep}_{x_{j,k},w}$ are the C_0-adapted versions of $\mu^{cep}_{\tilde{x}_{j,k}}$ and $\Sigma^{cep}_{\tilde{x}_{j,k}}$, $C^x_{0,\max,w}$ is the $C^x_{0,\max}$ value of the w-th *final*-segment, and p is the order of MFCC features. It is noted that, for the first C_0 modeling scheme, $\mu^{cep}_{\tilde{x}_{j,k}}(0)$ in above equation equals 0 for all mixture components. These C_0-adapted HMM models are then used in the PMC method to be combined with the current noise model to generate C_0-adapted, noise-compensated HMM models. The current noise model is estimated from the input testing utterance by using the method proposed previously in [4]. Due to the fact that the close-form solution to find a perfect model-combination operator does not exist yet, the log-normal approximation [2] is used in this study. Lastly, these HMM models are used in the recognition search to find the best recognized base-syllable sequence. It is noted that, the C_0 component of the noise-compensated HMM models is not taken as a recognition feature in the first C_0 modeling scheme, while it is taken as an additional recognition feature for the second C_0 modeling scheme.

Several advantages of the proposed SCA-PMC method can be found as compared with the conventional PMC method which uses a rough gain matching factor estimated for an entire testing utterance. Firstly, it can track both the local phonemic loudness variation (via the use of $\overline{C'}_{0,j}$) and the global intonational loudness variation (via the use of $C^x_{0,\max,w}$). This makes it has a better noise compensation effect. This also makes it insensitive to the volume adjustment of the recording device. Secondly, it can take C_0 as an additional recognition feature to assist in the recognition. Thirdly, the gain matching between the speech models and the noise model, required in the PMC model combination, is implicitly achieved by the proposed C_0 adaptation scheme. Lastly, the de-normalization factor $C_{0,\max}$ is always estimated from a frame with high SNR. This makes it be a reliable estimate.

3. EXPERIMENTAL RESULTS

Effectiveness of the proposed method was examined by simulation on a speaker-dependent continuous Mandarin base-syllable recognition task. A clean-speech database provided by the Chunghwa Telecommunication Laboratories was used. The database contained 452 sentential utterances and 200 paragraphic utterances recorded at a 20 kHz sampling rate. It consisted, in total, of 35231 syllables including 28197 training syllables and 7034 testing syllables. All speech signals were first pre-

processed for each of 20-ms Hamming-windowed frame with 10-ms shift. A set of recognition features including 13 MFCC (including C_0), 12 delta MFCC, and a delta log-energy was computed for each frame. All noisy speech databases used in the following tests were artificially generated by adding white noises of SPIB (Signal Processing Information Base) [7] with different SNRs into the above clean-speech database.

The HMM recognizer used 139 sub-syllable models as basic recognition units including 100 3-state right-context-dependent *initial* models and 39 5-state context-independent *final* models. Besides, a single-state utterance-dependent model was used for noise. In each state, a mixture Gaussian distribution with diagonal covariance matrices was used. The number of mixture in each state was variable and depended on the number of training samples, but a maximum number of 10 mixtures was set. A one-stage DP search with cumulative bounded-state-duration constraints was used to find the best recognized base-syllable sequence. The base-syllable accuracy rate was used to evaluate the recognition performance. The frequency-weighted SNR [8] is used in the assessment of the noise effect to speech recognition.

We start with checking the noise compensation effect of the C_0-adapted, noise-compensated HMM models produced by the proposed SCA-PMC method. Figs. 4(a) and 4(b) shows, respectively, the noise-compensated C_0 and C_1 contours, of the correct HMM mixture sequence, generated by the conventional PMC method. The corresponding C_0 and C_1 contours produced by the SCA-PMC method are shown in Figs. 4(c) and 4(d). Obviously, both C_0 and C_1 contours produced by the SCA-PMC method match much better to their corresponding counterparts of the noisy signal than these produced by the conventional PMC method. This confirmed that a better noise compensation effect can be achieved by the SCA-PMC method.

We then examine the performance of the proposed SCA-PMC method for noisy Mandarin base-syllable recognition. The following recognition schemes were tested:

(1) The 'Match' scheme: the HMM method trained and tested under a matched condition. Its performances are taken as benchmarks.
(2) The 'PMC' scheme: the conventional PMC method.
(3) The 'SCA-PMC-1' scheme: the proposed SCA-PMC method using the first C_0 modeling scheme.
(4) The 'PMC/LC' scheme: a modified version of 'PMC' with board-class based likelihood compensation [4].
(5) The 'SCA-PMC-1/LC' scheme: an extended version of 'SCA-PMC-1' with board-class based likelihood compensation.
(6) The 'SCA-PMC-2/LC' scheme: the proposed SCA-PMC method using the second C_0 modeling scheme and with board-class based likelihood compensation.

Table 1 shows the experimental results of these recognition schemes under white noise at 6dB, 18dB and 30dB. It is can be found from Table 1 that all schemes with the SCA C_0-adaptation scheme outperform significantly over their counterparts without SCA. This confirms the effectiveness of using the proposed C_0 modeling method in the PMC method. We also find that 'SCA-PMC-1/LC' and 'SCA-PMC-2/LC' have the best accuracy rates. This shows that SCA and the likelihood compensation scheme can be simultaneously used in the PMC method. The effectiveness of using likelihood compensation to assist in noisy speech recognition can be seen by comparing the accuracy rates of 'PMC/LC' with these of 'PMC'. The details about the likelihood compensation can be found in [4]. Lastly, the accuracy rates of 'SCA-PMC-2/LC' at both 18dB and 30dB are slightly better than these of 'Match'.

Table 1. The experimental results of noisy Mandarin base-syllable recognition under white noise.

Recognition Scheme	6dB	18dB	30dB
Match	*64.9*	*74.3*	*79.7*
PMC	*41.9*	*66.0*	*70.5*
SCA-PMC-1	*50.2*	*70.0*	*73.4*
PMC/LC	*51.0*	*71.0*	*73.4*
SCA-PMC-1/LC	*55.9*	*73.6*	*79.2*
SCA-PMC-2/LC	*55.7*	*75.6*	*80.0*

5. SUMMARY

A modified PMC method for noisy Mandarin speech recognition has been proposed in this paper. It incorporates a new C_0 model of speech signal into the PMC method for improving its recognition efficiency. Experimental results have confirmed that the proposed method has a much better noise compensation effect and hence outperforms the conventional PMC method.

6. ACKNOWLEDGMENT

This work was supported by the National Science Council of Taiwan under Contract No. NSC87-2213-E-009-056. The database is provided by the Chunghwa Telecommunication Laboratories.

7. REFERENCES

[1] Yifan Gong, "Speech recognition in noisy environments: A survey", *Speech Communication, Vol. 16,* pp. 261-291, 1995.
[2] M. J. F. Gales and S. J. Young, "Cepstral parameter compensation for HMM recognition in noise," *Speech Communication*, Vol. 12, pp. 231-240, 1993.
[3] M. J. F. Gales and S. J. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Trans. Speech and Audio Processing*, Vol. 5**,** pp. 352-359, 1996.
[4] W. T. Hong and S. H. Chen, "A robust RNN-based pre-classification for Noisy Mandarin speech recognition," *EuroSpeech 97*, Vol. 3, pp. 1083-1086, 1997.
[5] S. H. Chen *et. al*., "An RNN-based preclassification method for fast continuous Mandarin speech recognition," *IEEE Trans. Speech and Audio Processing*, Vol. 6, pp. 86-90, 1998.

[6] D. H. Klatt, "A digital filter bank for spectral matching," *ICASSP 76*, pp. 573-576, 1976.
[7] The web page of *Signal Processing Information Base*: http://spib.rice.edu/spib/select_noise.html
[8] J. Tribolet *et. al.*, "A study of complexity and quality of speech waveform coders," *ICASSP 78*, pp. 586-590, 1978.

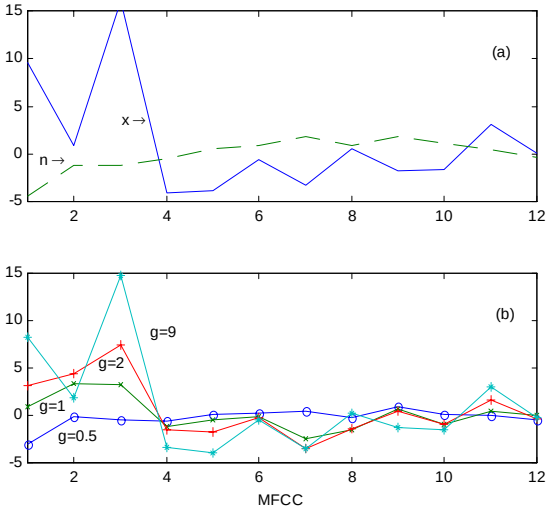

Fig. 1(a) The MFCC features of clean speech *x* and noise *n*. (b) The plots of the MFCC features of noisy signal *y* obtained by combining the same *x* and *n* using four different values of *g*.

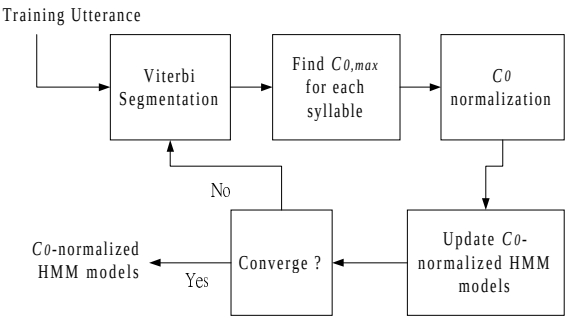

Fig. 2. A flow chart of the training phase of the SCA-PMC method.

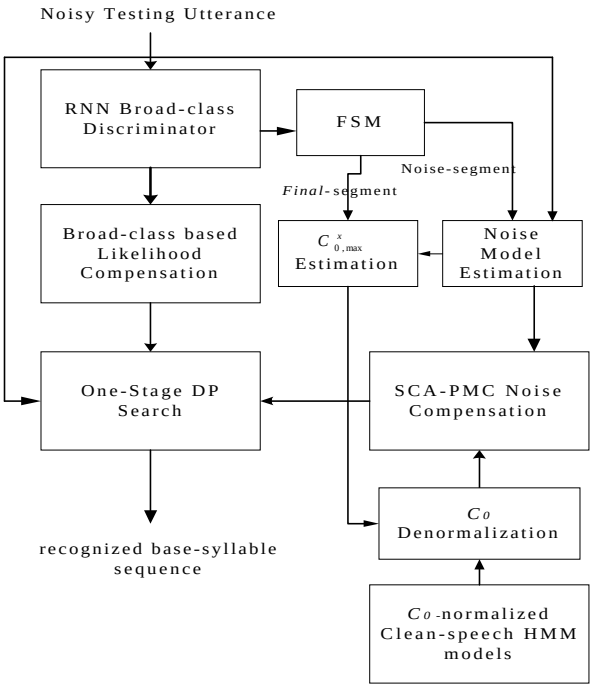

Fig. 3. A block diagram of the testing phase of the SCA-PMC method.

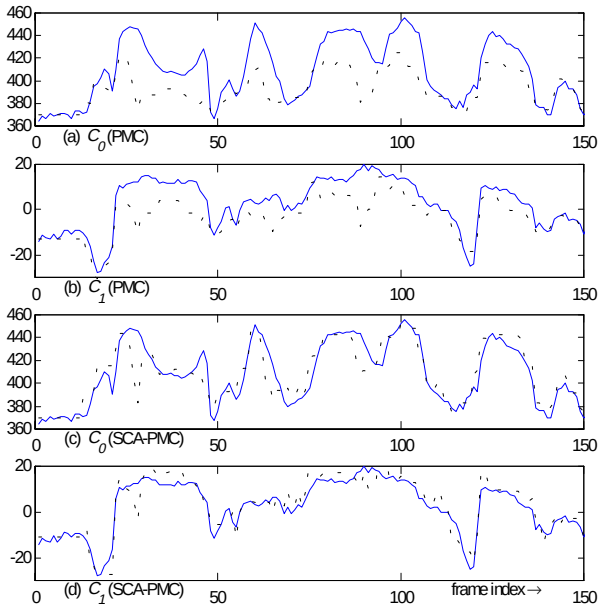

Fig. 4. The noise-compensated C_0 and C_1 contours produced by (a)(b) the conventional PMC and (c)(d) SCA-PMC. The solid lines represent contours of the original noisy speech and the dotted lines denote contours of correct mixture sequence determined by Viterbi segmentation.