Derivation of Robust Mel -Frequency Cepstral Coefficients Using a Weighting Discrete Cosine Transform

Wei-Wen Hung, Yu-Yung Huang and Chia-Hsiung Xsiao
Department of Electrical Engineering, Mingchi Institute of Technolog
Taishan, Taipei, Taiwan, 24306, Republic of China
E-mail : wwhung@ccsun.mit.edu.tw

ABSTRACT

Mel-frequency cepstral coefficient (MFCC) is one of the most popular speech features used in an automatic speech recognition system. In order to improve its discrimination capability and robustness in various environments, a weighting discrete cosine transform (WDCT) is proposed in this paper and incorporated into the derivation of the conventional mel-frequency cepstral coefficients. The weighting function used in the discrete cosine transform can be easily calculated from the log-spectral amplitudes of each speech frame and by which we can adequately explore the relative reliabilities among different critical band filters. Experimental results for recognition of continuous telephone speech indicate that the syllable recognition rates of the WDCT-MFCC are 3.91% and 2.16% higher than those of the conventional MFCC in the cases of with and without compensation of channel distortions, respectively. Those results verify the robustness and effectiveness of th proposed WDCT-MFCC. Moreover, comparisons of F-ratio measures between the conventional MFCC and WDCT-MFCC also conclude that the WDCT-MFCC has superior discrimination capability in modeling a speech recognizer.

***Key Words** : mel-frequency cepstral coefficient (MFCC), discrete cosine transform (DCT), critical band filter, F-ratio measure, discrimination capability.*

1. INTRODUCTION

When a speech recognition system trained in a well-defined environment is used in the real world applications, the acoustic mismatch between training and testing environments will degrade its recognition accuracy severely. This acoustic mismatch is mainly caused by a wide variety of distortion sources, such as ambient additive noise, channel effect and speaker's Lombard effect. During the past several decades, researchers focused their attentions in dealing with the mismatch problem and tried to narrow the mismatch gap. Up to the present time, there ar e many algorithms have been proposed and successfully applied for robust speech recognition. Generally speaking, the methods for handling noisy speech recognition could be roughl classified into the following approaches [2][13].

The first approach tries to minimize the distance measures between reference models and testing speech signals by adaptively adjusting speech signals in *feature space.* For example, Mansour and Juang [9] found that the norm of a cepstral vector is shrunk under noise contamination. Therefore, they used a first-order equalization method to adapt the cepstral mean of a reference model so that the shrinkage of speech features can be adequately compensated. Likewise, Carlson and Clement [1] also proposed a weighted projection measure (WPM) for recognition of noisy speech in the framework of continuous density hidden Marko model (CDHMM). In addition, the norm shrinkage of cepstral means will also lead to the reduction of HMM covariance matrices. Thus, Chien et al., [2] proposed a variance adapted and mean compensated likelihood measure (VA-MCLM) to adapt the mean vector and covariance matrix simultaneously.

The second approach estimates a transformation function in *model space* for transforming reference models into testin environment and thus the environmental mismatch gap can be effectively reduced. In the literature, there were a number of techniques compensating ambient noise effect in model space. Among them, one of the most promising techniques is the so-called parallel model combination (PMC). In the PMC algorithm, Varga and Moore [14] adapted the statistics of reference models to meet the testing conditions b optimally combining the reference models and noise model in linear spectral domain. In the later few years, several related works have been successively reported for improvin the performance of the PMC method. Flores and Young [4] integrated the spectral subtraction (SS) and PMC methods t seek for further improvement in recognition accuracy. In addition, Gales and Young [5] ext ended PMC scheme to include the effect of convolutional noise.

In the third approach, a more robust feature representation is developed in *signal space* so that the speech feature is invariant or less susceptible to environmental variations. In this approach, Mansour and Juang [10] proposed the short-time modified coherent (SMC) representation to instead of the linear predictive coding (LPC) technique for enhancing the signal-to-noise ratio (SNR) of noisy speech. Hernando and Nadeu [6] designed a one-sided auto-correlation linear predictive coding (OSALPC) technique for speech recognition in car environment. Moreover, Hung and Wan

[8] developed an adaptive signal limiter as a pre-processor to dynamically reduce the variability of speech features in mismatched conditions.

In this paper, a weighting discrete cosine transform (WDCT) is proposed and incorporated into the derivation of conventional mel-frequency cepstral coefficients (MFCCs). Our goal is to improve the discrimination capability and robustness of speech recognizer that uses the WDCT-MFCC as speech features. Experimental results show that no matter whether the underlying environment is compensated or not, the proposed WDCT-MFCC can always achieve higher syllable recognition rates and provide better robustness for telephone speech recognition. In order to make the description of the proposed WDCT-MFCC more detailed, this paper will be organized into the followin sections. The second section presents the mathematical formulation of the conventional m el-frequency cepstral coefficients. Following this, the third section describes how the weighting discrete cosine transform could be incorporated into the derivation of conventional MFCC. Experimental results that demonstrate the effectiveness and robustness of the WDCT-MFCC are shown in Section 4. Finally, in Section 5, we conclude and discuss our present and future works.

2. FORMULATION OF THE MFCC

Mel-frequency spectral coefficient (MFCC) is firstly introduced by Davis and Mermelstein [3] in 1980. The distinct advantage of the MFCC is that it takes the phenomena of mel pitch scale and critical band filter int account to make the derivation of speech features more meaningful. Assuming that $x_m(n)$ represents the frame of a speech signal where m is the frame index and n is the discrete time index within a frame. The derivation procedure of the MFCC can be briefly summarized as follows.

Step1. Calculate the power spectrum of each speech frame by using the discrete Fourier transform (DFT)

$$X_m(k) = \sum_{n=0}^{N-1} x_m(n)\cdot\exp\left(-j\cdot\frac{2\cdot\pi\cdot n\cdot k}{K}\right), \qquad (1)$$

where $0 \le m \le M-1,\ 0 \le k \le K-1,\ 0 \le n \le N-1.$

Step2. For the $i-th$ mel-scaled critical band filter $\psi_i(k)$, compute the accumulated log-spectral amplitude E_i from the output of critical band filter $\psi_i(k)$, i.e.,

$$E_i = \sum_{k=0}^{K-1} X_m(k)\cdot\psi_i(k), \qquad (2)$$

where i is the index o critical band filter and $0 \le i \le B-1$.

Step3. Calculate the mel-frequency cepstral coefficient c_j by means of the discrete cosine transform (DCT)

$$c_j = \sum_{i=0}^{B-1} \log(E_i)\cdot\cos\left[j\cdot\left(\frac{2\cdot i-1}{2}\right)\cdot\frac{\pi}{B}\right] \qquad (3)$$

where $0 \le j \le L-1$, and L is the desired length of MFCC.

3. MFCC BASED ON THE WEIGHTING DISCRETE COSINE TRANSFORM

It is well known that in time domain the segments of a speech signal with lower amplitude are more influenced b ambient noise and channel interference whereas the segments with higher amplitude are less influenced and bear more reliable information. This phenomenon should be als valid in spectral domain. That is, the larger the value of accumulated log-spectral amplitude, the more reliable the output of a critical band filter. Consequently, a weighting function is suggested to be incorporated into the discrete cosine transform and so that the relative reliabilities among various critical band filters could be adequately explored. Based upon above description, the derivation of conventional MFCC may be modified slightly and thus the Eq. (3) can be rewritten as [7]

$$\hat{c}_j = \sum_{i=0}^{B-1} w_i\cdot\log(E_i)\cdot\cos\left[j\cdot\left(\frac{2\cdot i-1}{2}\right)\cdot\frac{\pi}{B}\right] \qquad (4)$$

In above equation, the weighting function w_i is empiricall expressed as

$$w_i = \frac{\log(E_i)}{\sum_{ii=0}^{B-1}\log(E_{ii})}.$$

(5)

This modified MFCC is called the weighting discrete cosine transform MFCC and abbreviated as WDCT-MFCC.

4. EXPERIMENTS AND RESULTS

A multi-speaker continuous Mandarin speech recognition involving the MAT speech database [7] was conducted t demonstrate the discrimination capability and robustness of the proposed WDCT-MFCC. This MAT (Mandarin Across Taiwan) speech database provided by the Computational Linguistic Society of R.O.C. was collected over the public telephone network and each word comprised 1~23 Mandarin syllables. From the MAT database, we chose 8320 phonetically balanced Mandarin utterances (37784 syllables) spoken by 81 males and 79 females to train the right-context-dependent sub-syllable hidden Markov models (HMMs) of 410 Mandarin syllables. On the other hand, from the different set of the MAT database, 500 testing utterances (4754 syllables) spoken by 15 males and 15 females were used for recognition task. Also, each syllable model contained six to seven states in which the output observation distribution is characterized by an 8-mixture Gaussian continuous density hidden Markov model (CDHMM). The 12-order mel-frequency cepstral coefficients and its first-order time derivatives were used as the feature parameters.

In our preliminary experiments, we used F-ratio parameter as a measure of discrimination capability for syllable models using different speech features. The F -ratio is a useful measure of separability among multiple clusters, and defined as [12]

$$F - ratio = \frac{\text{var}\, iance \quad of \quad means}{mean \quad of \quad \text{var}\, iances} \tag{6}$$

That is,

$$F - ratio = \frac{\frac{1}{U} \cdot \sum_{i=0}^{U-1} \left[\left(\frac{1}{U} \cdot \sum_{j=0}^{U-1} \mu_j \right) - \mu_i \right]^2}{\frac{1}{U} \cdot \sum_{i=0}^{U-1} \sigma_i^2}, \tag{7}$$

where μ_i and σ_i^2 represent the means and variances of the $i - th$ cluster and U is the number of clusters. I continuous speech recognition, the speech features corresponding to a' syllable model are generall characterized by using a hidden Markov model (HMM) which uses multi-variate Gaussian probability densit functions (pdf) to represent the distribution of those speech features. If the probability distributions of a syllable model are well separated, the corresponding measure of F-rati should be relatively high and better recognition accuracy can be accordingly obtained. In Fig. 1 and Fig. 2, we compared the F-ratio measures of male and female syllable models "ㄚ" using the MFCC and WDCT-MFCC as speech features, respectively. From those figures, we can observe that for most feature indices, the F-ratio measure of WDCT-MFCC is higher than that of the MFCC. This implies that the WDCT-MFCC is capable of providing better discrimination capability than that achieved by the MFCC.

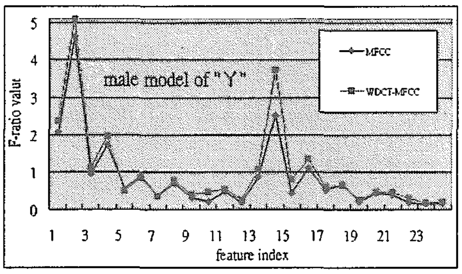

Figure 1. Comparison of F -ratio measures for male syllable model "ㄚ" using the MFCC and WDCT-MFCC as speech features.

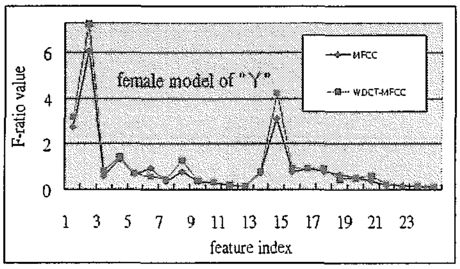

Figure 2. Comparison of F -ratio measures for female syllable models "ㄚ" using the MFCC and WDCT-MFCC as speech features.

To make the comparisons between the MFCC and WDCT-MFCC more convincing, we also conducted a series of recognition experiments for continuous telephone speech based on the MAT speech database to evaluate the robustness and effectiveness of the WDCT-MFCC. Experimental results for speech recognition using different methods over a real public telephone network were demonstrated in Table 1. In this table, the notation "CMS" is used to represent the cepstral mean subtraction [11] which is a simple, effective and widely used technique for compensating channel distortions. From the experimental results, we can find the following facts : (1) Comparing with the case of using conventional MFCC only, the insertion error rate (I.E.R), deletion error rate (D.E.R) and substitution error rate (S.E.R) when employing CMS technique drop by 0.42%, 0.08% and 3.41%, respectively. In addition, the syllable recognition rate (S.R.R) and utterance recognition rate (U.R.R) can also be improved about 4.11% and 0.4%, respectively. (2) In the environment with channel distortion, the S.R.R and U.R.R using the WDCT-MFCC are 2.16% and 1.1% higher than those using the conventional MFCC, respectively. (3) When the channel distortion is compensated b means of the CMN, the S.R.R and U.R.R using the WDCT-MFCC can be further improved by 3.91% and 2.6% with respect to those using the conventional MFCC, respectively.

Table 1. Experimental results for continuous speech recognition over a real public telephone network

	I.E.R	D.E.R	S.E.R	S.R.R	U.R.R
MFCC	8.94%	1.35%	48.71%	40.79%	4.4%
WDCT-MFCC	8.54%	1.37%	47.14%	42.95%	5.5%
MFCC+CMS	8.52%	1.27%	45.30%	44.90%	4.8%
WDCT-MFCC +CMS	7.82%	1.35%	42.02%	48.81%	7.4%

5. CONCLUSIONS AND FUTURE WORKS

In this paper, a slight modification is proposed for derivation of mel-frequency cepstral coefficients, i.e., incorporating a weighting function into the discrete cosine transform, t adequately reflect the relative reliabilities among various critical band filters employed in the conventional MFCC. Experimental results show that no matter whether the channel distortion of underlying environment is compensated or not, the proposed WDCT-MFCC indeed has superior robustness and discrimination capability comparing with those of the conventional MFCC. Moreover, due to its simple form, the weighting function used in the WDCT-MFCC demands much less computation time.

We are currently continuing the efforts towards the use of fuzzy membership weighting function instead of the proposed weighting function to further increase the discriminative ability and robustness of MFCC.

6. ACKNOWLEDGMENTS

We would like to express our deep gratitude to Dr. Hsia - Chuan Wang who is with the Department of Electrical Engineering at National Tsing Hua University, for his enthusiasm in providing the MAT speech database and many constructive suggestions that improve the quality of this paper. This research has been sponsored by the National Science Council, Taiwan, ROC under the contrast number NSC-88-2815-C-131-002-E [7] for undergraduate project.

7. REFERENCES

[1] B. A. Carlson and M. A. Clement, "A projection-based likelihood measure for speech recognition in noise," *IEEE Trans. on Speech and Audio Processing*. Vol. 2, pp. 97-102, 1994.

[2] J. T. Chien, "Speech recognition under telephone environments," Ph.D. Thesis. Department of Electrical Engineering, National Tsing Hua University, Taiwan, R.O.C, 1997.

[3] S. Davis and P. Mermelstein, "Comparison o parametric representations for monosyllable word recognition," *IEEE Trans. Acoust., Speech, Signal Processing,* Vol. 28, pp. 357-366, 1980.

[4] J. A. N. Flores and S. J. Young, "Continuous speech recognition in noise using spectral subtraction and HMM adaptation," *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP).* San Francisco. Vol. 1, pp. 409-412, 1992.

[5] M. J. F. Gales and S. J. Young, "Robust speec recognition in additive and convolutional noise using parallel model combination," *Computer Speech and Language,* Vol. 4, pp. 352-359, 1995.

[6] J. Hernando and C. Nadeu, "Linear Prediction of the one-sided autocorrelation sequence for noisy speech recognition," *IEEE Trans. Speech Audio Processing,* Vol. 5, No. 5, pp. 80-84, Jan. 1997.

[7] W. W. Hung, Y. Y. Huang and C. H. Xsiao, "Study on the segmentation p roblem for continuous speech recognition over telephone network," *Technical report (National Science Council Project for Undergraduate) sponsored by National Science Council, NSC-88-2815-C-131-002-E,* July 1999.

[8] W. W. Hung and H. C. Wang, "Smoothing hidden Markov models by using an adaptive signal limiter for noisy speech recognition," *Speech Communication*, Vol. 28, issue 3, pp. 243-260, July 1999.

[9] D. Mansour and B. H. Juang, "A family of distortion measures based upon projection operation for robust speech recognition," *IEEE Trans. on Acoustics, Speech, and Signal Processing,* Vol. 37, pp. 1659-1671, 1989.

[10]D. Mansour and B. H. Juang, "The short-time modified coherence representation and noisy speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing,* Vol. 37, pp. 795-804, June 1989.

[11]C. Mokbel, P. Paches, D. Jouvet and J. Monne, "Compensation of telephone line effect for robust speech recognition," *Int. Conf. Spoken Language Processing,* pp. 987-990, 1994.

[12]S. Nicholson, B. Milner and S. Cox, "Evaluating feature set performance using the F -ratio and J-measures," *Proceeding of European Conference on Speech Communication and Technology (EUROSPEECH),* vol. 1, pp. 413-416, Greece, Sept. 1997.

[13]A. Sankar and C. H. Lee, "A maximum-likelihood approach to stochastic matching for robust speech recognition," *IEEE Trans. on Speech and Audio Processing,* Vol. 4, pp. 190-202, 1996.

[14]A. P. Varga and R. K. Moore, "Hidden Markov model decomposition of speech and noise," *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP),* San Francisco. pp. 845-848, 1992.