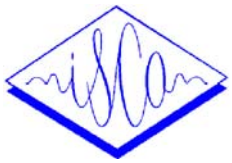

EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

ADAPTATION OF TIME DIFFERENTIATED CEPSTRUM FOR NOISY SPEECH RECOGNITION

Tai-Hwei Hwang, Lee-Min Lee and Hsiao-Chuan Wang*

Department of Electrical Engineering, National Tsing-Hua University, Hsinchu, Taiwan, ROC 30043
Email: hcwang@ee.nthu.edu.tw

*Department of Electrical Engineering, Mingchi Institute of Technology,
Taipei Hsien, Taiwan, ROC 243

ABSTRACT

In this paper, a noise compensation algorithm using the first order approximation of cepstral function is presented. The derivative term is replaced by the difference of cepstra for the adaptation of wide range variation of noise power. The differences of cepstral mean vectors between the clean and noisy version, termed as the deviation vector, are applied to adapt cepstrum and delta cepstrum. The experimental results show that using deviation vector to adapt the cepstral coefficients can gain a significant improvement over the method based on weighted projection measure. Further improvement can be made by jointly adapting the cepstrum and delta cepstrum.

1. INTRODUCTION

The performance of an automatic speech recognizer (ASR) degrades seriously when it was deployed in a noisy environment [1]. Many efforts have been plunged into the problem and can be categorized as follows. a) The clean reference patterns (speech models) are adapted to include the noise effect [2]. b) The more robust speech features are studied [4]. c) A robust distance measure is applied [5]. In [5], D. Mansour and B. H. Juang investigated the behavior of cepstral vector under the effect of additive white noise and found that the norm of cepstral vector shrank but the orientation was slightly affected. The mis-match due to additive noise is affected by two facts; one is the variety of noise and the other is the signal-noise ratio. For the adaptation of speech models, such as in [3], the spectrum of background noise can be obtained before recognition, and the combination ratio of models can be estimated by the power ratio of test utterance and background noise. The compensation process would be repeated during idle periods so that slowly changing noise or signal level could be tracked.

In our previous work, an adaptation method of cepstral coefficients was proposed based on a first order approximation of cepstral function [6]. The cepstrum coefficients can be effectively adapted to the noise environmental using the simple function. In this paper, the relationship of time-differentiated cepstrum and additive noise power is investigated. With a simplification on that relationship, an adaptation on the time-differentiated cepstrum can be easily performed.

2. ADAPTATION OF TIME-DIFFERENTIATED CEPSTRUM

2.1 Approximation of cepstral function of noise power

A cepstral vector, composed of cepstral coefficients of low quefrency except the term of zero quefrency, is generally adopted as a feature vector of speech. In our previous work, a cepstral vector of noisy speech was approximated by the first order Taylor series expansion arround the cepstrum of clean speech with respect to the noise power, *i.e.*,

$$\bar{c}(\eta) \cong \bar{c}(0) + \frac{\partial \bar{c}(\gamma)}{\partial \gamma}|_{\gamma=0}(\eta - 0) \qquad (1)$$

where $\bar{c}(\eta)$ is an approximation of cepstral vector of noisy speech, $\bar{c}(0)$ is its clean version. Since the interval of convergence is limited within a small range, the approximation error is getting large when increasing the noise power. Therefore, the performance

This work has been partially sponsored by the National Science Council, Taiwan, ROC, under contract number NSC-84-2213-E-007-050

improvement of using $\frac{\partial \bar{c}(\gamma)}{\partial \gamma}|_{\gamma=0}$ to adapt cepstrum coefficients is limited. To reduce the approximation error for a wide range variation of noise power, a cepstral difference of noisy speech and its clean version is adopted to replace the derivative term. In this case, equation (1) can be replaced by

$$\bar{c}(\alpha) \cong \bar{c}(0) + \alpha \cdot (\bar{c}(\gamma) - \bar{c}(0)) = \bar{c}(0) + \alpha \cdot \Delta\bar{c} \quad (2)$$

where $\bar{c}(\gamma)$ is a cepstral vector of noisy speech which is given by a pre-defined noise power,

$$\Delta\bar{c} = \bar{c}(\gamma) - \bar{c}(0) \quad (3)$$

is the replacement of first order derivative at zero noise power, and α is a scaling factor which accounts for the noise power level. Based on this approximation, cepstral coefficients of clean speech can be adapted to noisy condition by searching an optimal scaling factor α during the recognition phase.

Dynamic features, which infer the time variation of spectra, have been shown to be important in speech recognition [8]. Thus, in addition to the cepstral coefficients, the time differential of cepstrum is adopted as a part of feature vector. In the case of additive noise, the time differentiated cepstral coefficients would be affected and have to be adapted for robust speech recognition. Taking time differential of (2), we obtain

$$\frac{\partial \bar{c}(\alpha,t)}{\partial t} \cong \frac{\partial \bar{c}(0,t)}{\partial t} + \frac{\partial \alpha}{\partial t} \cdot \Delta\bar{c} + \alpha \frac{\partial \Delta\bar{c}(t)}{\partial t}. \quad (4)$$

Assuming that the noise power is stationary, the time differential of scaling factor can be negligible and equation (4) can be rewritten as

$$\frac{\partial \bar{c}(\alpha,t)}{\partial t} \cong \frac{\partial \bar{c}(0,t)}{\partial t} + \alpha \frac{\partial \Delta\bar{c}(t)}{\partial t}. \quad (5)$$

In this expression, $\frac{\partial \bar{c}(0,t)}{\partial t}$ is the time differential of cepstral vector of clean speech and is adapted by $\alpha \frac{\partial \Delta\bar{c}(t)}{\partial t}$ to approximate the time differential of cepstral vector under noise effect.

The main procedures of the proposed method can be divided into two parts; one is to set up deviation vectors of the time differentiated of cepstrum in the before hand of recognition, and the other is to find the optimal scaling factors during the recognition phase. The deviation vectors $\Delta\bar{c}(t)$ can be obtained by taking the difference of clean cepstral vectors and its noisy version. The deviation vector of time differentiated cepstrum $\frac{\partial \Delta\bar{c}(t)}{\partial t}$ can be obtained by taking time differential of equation (3), and we obtaine

$$\frac{\partial \Delta\bar{c}(\gamma,t)}{\partial t} = \frac{\partial \bar{c}(\gamma,t)}{\partial t} - \frac{\partial \bar{c}(0,t)}{\partial t}. \quad (6)$$

In practical application, time differential of cepstrum is generally replaced by a polynomial approximation, which is termed as the delta cepstrum. Thus, the deviation vector of time differentiated can be obtained by taking the difference of delta cepstra of noisy speech and its clean version.

2.2 Optimal scaling factors for adaptation

Once the deviation vectors of cepstrum and of delta cepstrum are computed for each HMM state, the adaptation can be performed by searching for the optimal scaling factors. The scaling factor accounts for the noise power and can be obtained by the estimated power of background noise. Nevertheless, the estimation of noise power can be omitted by the direct estimation of scaling factor using a maximum likelihood procedure. The estimation of an optimal scaling factor could be done as follows. Let the feature vectors $\mathbf{v}$ of test utterence consist of cepstrum and delta cepstrum coefficients. Given the log likelihood function of a HMM state is modeled by a Gaussian distribution,

$$L(\mathbf{v}; \hat{\mathbf{u}}_i, \Sigma_i) =$$

$$\frac{-M}{2}\ln(2\pi) - \frac{1}{2}\ln(|\Sigma_i|) - \frac{1}{2}(\mathbf{v} - \hat{\mathbf{u}}_i)^T \Sigma_i^{-1} (\mathbf{v} - \hat{\mathbf{u}}_i), \quad (7)$$

where $\hat{\mathbf{u}}_i = \mathbf{u}_i(0) + \alpha \Delta\mathbf{u}_i$ is an adapted mean vector of state i, and M is the dimension of feature vector. The optimal scaling factor α is which maximizes the log likelihood function. Taking derivative of (7) with respect to α and setting the result to zero, the optimal adaptation factor can be obtained by

$$\alpha_{opt} = \frac{(\mathbf{v} - \mathbf{u}_i(0))^T \Sigma_i^{-1} \Delta\mathbf{u}_i}{\Delta\mathbf{u}_i^T \Sigma_i^{-1} \Delta\mathbf{u}_i}. \quad (8)$$

3. EXPERIMENTAL EVALUATION

3.1 Experimental setup

To show the effectiveness of the proposed method, a speaker-independent Mandarin digit recognition was conducted using an HMMs based recognizer. The HMMs of clean speech were trained by using clean speech data from 25 males and 25 females. Speech data from another 25 males and 25 females were used as the test data. There were 5 repetitions of each digit for each speaker. Sampling rate is 8kHz and no pre-emephasis applied on it. The feature vector is composed of 12 LPC derived cepstral coefficients and 12 delta cepstral coefficients. The covariance matrices of HMMs' states, Σ_i's, were all in the diagonal form for simplicity. There were two mixtures in each state. The segmental-k-means algorithm was applied to train the HMMs. The deviation vectors of states were trained by the segmentation of training data obtained from HMMs of clean speech. The procedure for obtaining the state deviation vectors is as follows. Using the HMMs of clean speech, a Viterbi decoding is applied to all of the training utterances to find their state sequences. According to the segmentation of state sequence, the state deviation vector can be estimated by taking the ensemble average of the corresponding deviation vectors, *i.e.*,

$$\Delta \mathbf{u}_i = \frac{1}{N_i} \sum_{k=1}^{N_i} \Delta \mathbf{v}_{i,k}(\eta^*), \qquad (9)$$

where $\Delta \mathbf{v}_{i,k}(\eta^*)$ is a deviation vector decoded to state i and N_i is the total number of $\Delta \mathbf{v}_{i,k}(\eta^*)$. The deviation vectors of training data can be obtained by using equation (3) on each training utterance. In which, the cepstral vectors of its noisy version are generated by a pre-defined global signal-noise ratio. This could be done in the auto-corrlation domain. Signal-noise ratios of 20dB and 15dB are adopted to generate the deviation vectors in each pass of experiment.

Noisy test utterances were artificially produced by the waveform addition of white noise in time domain. For comparison, results of baseline system and weighted projection measure (WPM) based HMM recognizer [9] are also demonstrated. The baseline system is an HMM based recognizer without any adaptation on both cepstrum and delta cepstrum coefficients. The weighted projection measure is an adaptation method for the cepstral mean of HMM states but the delta cepstra are remained un-adapted. Dev(X) indicates the use of deviation vector for the cepstrum, not the delta cepstrum. Dev1(X) indicates the adaptation to both cepstrum and delta cepstrum. X indicates the signal-noise ratio in dB that is used in generating the deviation vectors.

3.2 Experimental results and discussion

Table 1 shows the error rates for the comparison of several adaptation schemes. The performance of adaptation using deviation vector for both Dev(X) and Dev1(X) are better than using weighted projection measure. However, the improvement on the adaptation of delta cepstrum is limited. The other experiments by using mel-cepstrum coefficients as the feature vectors also held to the similar results. As shown in Table 2, the adaptation performance on mel-cepstrum coefficients are more significant than that on LP derived cepstrum. For example, about 75% of error rate is gained for Dev1(20) of mel-cepstrum, while only 63% is obtained for LP derived cepstrum.

The results show no significant benifit by further adapting the delta cepstrum. One possible reason is that the omission of $\frac{\partial \eta}{\partial t} \cdot \Delta \vec{c}$ in equation (4) could be improper for the real application. The improvement becomes small when the deviation vector is generated by lower signal-noise ratio.

	20db	15db	10db	5db	0db
No adaptation	24.24	35.60	52.96	68.72	87.64
WPM	9.60	16.04	30.40	49.04	66.16
Dev(20)	9.28	12.24	20.68	35.32	63.60
Dev1(20)	8.84	11.12	19.36	33.48	60.04
Dev(15)	10.40	13.84	22.64	32.60	52.04
Dev1(15)	9.76	13.68	22.04	31.68	51.56

Table 1, Recognition error rates for various adaptation schemes by using cepstrum and delta cepstrum as the feature vector

	20db	15db	10db	5db	0db
No adaptation	18.28	25.60	36.32	63.44	84.92
WPM	10.84	18.36	29.76	47.48	64.68
Dev(20)	5.68	7.56	10.08	20.72	41.84
Dev1(20)	4.50	6.56	9.04	16.28	38.36
Dev(15)	5.88	7.64	9.92	20.56	39.76
Dev1(15)	5.88	7.40	8.88	14.88	33.40

Table 2, Recognition error rates for various adaptation schemes by using mel-cepstrum and delta mel-cepstrum as the feature vector

4. CONCLUSION

In this paper, the cepstral coefficients of low quefrency except the term of zero quefrency are approximated by a first order function of noise power. Taking the time differential on the noisy cepstrum, we formulate an adaptation scheme based on the deviation vectors. The deviation vectors are obtained from the difference of the clean cepstrum and its artificial generated noisy version. In fact, the deviation vector describes the changing direction of cepstrum and delta cepstrum due to additive noise. These deviation vectors were fixed for all the tests with different SNR levels. The performance of using deviation vectors is more significant than the projection method, while the amount of computation required for both methods are almost the same.

REFERENCES

[1] B. H. Juang, "Speech recognition in adverse environments," Comput. Speech Language, pp. 275-294, 1991.

[2] M. J. F, Gales and S. J. Young, "HMM recognition in noise using parallel model combination," EuroSpeech 93, pp. 837-840.

[3] M. J. F, Gales and S. J. Young, "Cepstral parameter compensation for HMM recognition in noise," Speech Comunication 12, 1993, pp. 231-239.

[4] H. Hermansky, N. Morgan, "RASTA processing of speech," IEEE Trans. Speech, Audio Processing, 2: 578-589, October, 1994.

[5] 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, No. 11, Nov. 1989.

[6] T. H. Hwang, L. M. Lee and H.C. Wang, "Feature adaptation using deviation vector for robust speech recognition in noisy environment", Int. Conf. Acoust., Speech, Signal Proceedings '97, pp. 1227-1231.

[7] L. M. Lee and H. C. Wang, "An extended Levinson-Durbin algorithm for the analysis of noisy auto regressive process," IEEE Signal Processing Letters, Vol. 3, No. 1, pp. 13-15, 1996.

[8] S. Furi, "Speaker independent isolated word recognition using dynamic features of speech spectrum," IEEE Trans. Acoustics, Speech, Signal Processing, ASSP-34 (1): 52-59, February 1986.

[9] B. A. Calson and M. A. Clements, "A projection-based likelihood measure for speech recognition in noise," IEEE Trans. Speech and Audio Processing, vol. 2, pp. 97-102, January 1994.