FEATURE ADAPTATION USING DEVIATION VECTOR FOR ROBUST SPEECH RECOGNITION IN NOISY ENVIRONMENT

Tai-Hwei Hwang, Lee-Min Lee and Hsiao-Chuan Wang*
Department of Electrical Engineering, National Tsing-Hua University,
Hsinchu, Taiwan 30043
*Department of Electrical Engineering, Mingchi Institute of Technology,
Taipei Hsien, Taiwan 243

ABSTRACT

When a speech signal is contaminated by additive noise, its cepstral coefficients are assumed to be the functions of noise power. By using Taylor series expansion with respect to noise power, the cepstral vector can be approximated by a nominal vector plus the first derivative term. The nominal cepstrum corresponds to the clean speech signal and the first derivative term is a quantity to adapt the speech feature to noisy environment. A deviation vector is introduced to estimate the derivative term. The experiments show that the feature adaptation based on deviation vectors is superior to those projection based methods.

1. INTRODUCTION

Background noise is an inevitable source degrading the performances of the speech recognizers [1]. To overcome this problem, many efforts have been focused on developing the robust systems where the noise effect is minimized. Among these studies, D. Mansour and B. H. Juang investigated the behavior of cepstral vector under the effect of additive white noise [2]. In their research, they found that the norm of cepstral vector shrinks but the orientation is slightly affected when a clean speech is contaminated by white noise. With these discoveries, they proposed a series of robust distance measures which were termed as the projection measures. These projection measures were applied to the dynamic time warping (DTW) based speech recognizer and also to the hidden Markov model (HMM) based speech recognizer [5].

Since the orientation of a cepstral vector does not change drastically under the additive noise effect, we may adapt the clean cepstral coefficients along their changing direction to minimize the distortion. Two problems are raised in the adaptation scheme. One is how to find the changing direction of cepstral vector due to additive noise, and the other is how to determine an optimal scaling factor for adapting the clean cepstral vector. When a speech signal is contaminated by additive noise, its cepstral coefficients are assumed to be functions of noise power. By Taylor series expansion with respect to noise power, the cepstral vector can be approximated by a nominal vector plus the first derivative term [3]. The nominal cepstrum corresponds to the clean speech signal. The first derivative term is a quantity to adapt the speech feature to the noisy environment. This quantity is a product of noise power and a derivative vector of cepstrum with respect to noise power. The derivative vector implies the changing direction of cepstral vector due to additive noise in time domain.

To estimate the changing direction, two methods are applied in this paper. One is by using the transformation of LP coefficients [4], and the other is by taking difference between the cepstral vectors of clean speech and its noisy version. Once the changing direction for a cepstral vector is determined, the adaptation can be done by finding an optimal scaling factor during the recognition phase. In general, the scaling factor is relative to the signal to noise ratio (SNR). A maximum likelihood estimation of scaling factor was adopted in this study for simplicity.

This paper is organized as follows. In section 2, the adaptation model is defined and the methods for obtaining the deviation vector are also introduced. In section 3, the proposed adaptation method is introduced and its performance on the task of speech recognition is examined by DTW based and HMM based speech recognizers. Finally, a conclusion is made in section 4.

2. DEVIATION OF CEPSTRAL VECTOR

2.1. First order approximation

In the auto-regressive (AR) modeling of speech signal $x[n]$, the LP coefficients $\{a_i\}$ are determined by solving

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

the p linear equations, which can be written in the matrix form as

$$\mathbf{R}\mathbf{a}_x = \mathbf{r}, \qquad (1)$$

where

$$\mathbf{R} = \begin{bmatrix} r_{x,0} & r_{x,1} & \cdots & r_{x,p-1} \\ r_{x,1} & r_{x,0} & & r_{x,p-2} \\ \vdots & & \ddots & \\ r_{x,p-1} & r_{x,p-2} & & r_{x,0} \end{bmatrix}, \mathbf{a}_x = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix}, \mathbf{r} = \begin{bmatrix} r_{x,1} \\ r_{x,2} \\ \vdots \\ r_{x,p} \end{bmatrix},$$

and $r_{x,m}$ is m^{th} auto-correlation of $x[n]$, *i.e.*, $r_{x,m} = E(x[n]x[n+m])$. When the clean speech signal $x[n]$ is contaminated by an uncorrelated white noise $w[n]$, *i.e.*, $y[n]$ = $x[n]$ + $w[n]$, the auto-correlation of $y[n]$ should be

$$r_{y,m} = \begin{cases} r_{x,0} + \eta, \text{ for } m = 0 \\ r_{x,m}, \text{elsewhere} \end{cases}, \qquad (2)$$

where η is the noise power, *i.e.*, $\eta = r_{w,0}$. Thus, the LP coefficients of $y[n]$ are

$$\mathbf{a}_y = (\mathbf{R} + \eta\mathbf{I})^{-1}\mathbf{r}, \qquad (3)$$

where $\mathbf{I}$ is the identity matrix. It is obvious that the LP coefficients of noisy speech are functions of the noise power η. In other words, $\mathbf{a}_y$ should be expressed as $\mathbf{a}(\eta)$. From the formulation of LPC derived cepstral coefficients,

$$c_m = a_m + \sum_{k=1}^{m-1}\left(\frac{k}{m}\right)c_k a_{m-k}, \quad 1 \le m \le p \qquad (4)$$

the cepstral coefficients of noisy speech can be obtained through a non-linear transformation of $\mathbf{a}(\eta)$. Thus, the cepstral coefficients of noisy speech are also functions of noise power. Using the first-order Taylor series expansion around zero noise power, the cepstral coefficients of noisy speech can be approximated by

$$\mathbf{c}(\eta) \cong \mathbf{c}(0) + \frac{d\mathbf{c}(\eta)}{d\eta}\Big|_{\eta=0}(\eta - 0), \qquad (5)$$

where $\frac{d\mathbf{c}(\eta)}{d\eta}$ is the derivative of cepstrum with respect to noise power and can be derived by taking derivative of (4) with respect to η. Then we obtain

$$c_m^{(1)} = a_m^{(1)} + \sum_{k=1}^{m-1}\left(\frac{k}{m}\right)\left(c_k^{(1)}a_{m-k} + c_k a_{m-k}^{(1)}\right), \quad 1 \le m \le p \quad (6)$$

where $c_m^{(1)} = \frac{dc_m(\eta)}{d\eta}$ and $a_m^{(1)} = \frac{da_m(\eta)}{d\eta}$ are the m^{th} components of derivatives of cepstrum and LP coefficients, respectively. $\frac{d\mathbf{a}(\eta)}{d\eta}\big|_{\eta=0}$ can be obtained by taking derivative of (3) with respect to η and let $\eta = 0$, *i.e.*,

$$\frac{d\mathbf{a}(\eta)}{d\eta}\Big|_{\eta=0} = -\mathbf{R}^{-2}\mathbf{r} = -\mathbf{R}^{-1}\mathbf{a}(0). \qquad (7)$$

2.2. Alternative estimation of the first order derivative terms

The first order approximation of cepstral vector is the basis of our adaptation method. In this situation, $\mathbf{c}(\eta)$ can be looked as a tracking of cepstral vector of real noisy speech, and $\frac{d\mathbf{c}(\eta)}{d\eta}\big|_{\eta=0}$ can be considered as a changing direction caused by the additive noise. However, from the view point of curve tracking, $\mathbf{c}(\eta)$ might not be properly approximated along the direction of tangent such as obtained by the derivative. Thus, an alternative estimate of changing direction is proposed by the concept of a cut line, and can be defined by

$$\frac{d\mathbf{c}(\eta)}{d\eta}\Big|_{\eta=0} \cong \frac{\mathbf{c}(\eta^*) - \mathbf{c}(0)}{\eta^* - 0}, \qquad (8)$$

where $\mathbf{c}(0)$ is a cepstral vector of a segment of clean speech, $\{x[n]\}$, and $\mathbf{c}(\eta^*)$ is a cepstral vector of an artificially generated noisy version of $\{x[n]\}$. Given a noise power level η^*, $\mathbf{c}(\eta^*)$ can be derived from the noise contaminated auto-correlation domain using (2), (3) and (4).

Substituting (8) into (5), the proposed approximation can be also expressed as

$$\hat{\mathbf{c}} = \mathbf{c}(0) + \alpha\,\Delta\mathbf{c}(\eta^*), \qquad (9)$$

where $\alpha = \eta/\eta^*$ is an scaling factor and $\Delta\mathbf{c}(\eta^*)$ is defined by

$$\Delta\mathbf{c}(\eta^*) = \mathbf{c}(\eta^*) - \mathbf{c}(0), \qquad (10)$$

With definition (10), $\Delta\mathbf{c}(\eta^*)$ is termed as a deviation vector in this paper.

3. FEATURE ADAPTATION FOR SPEECH RECOGNITION

3.1. Adaptation schemes

A reference pattern is expressed by a cepstral vector and its associated deviation vector. The adaptation scheme of robust pattern matching can be done by finding a proper scaling factor α. The optimal scaling factor is the one which minimizes the cepstral distance between a test vector $\mathbf{c}_T$ and an adapted reference cepstral $\hat{\mathbf{c}}_R$, *i.e.*,

$$\alpha_{opt} = \arg\min_{\alpha}\{Dist(\mathbf{c}_T, \hat{\mathbf{c}}_R)\}, \qquad (11)$$

In this study, L^2 norm is applied for the distance measure

$$Dist(\mathbf{c}_T,\hat{\mathbf{c}}_R) = \left|\mathbf{c}_R(0) + \alpha\,\Delta\mathbf{c}_R(\eta^*) - \mathbf{c}_T\right|^2, \qquad (12)$$

Taking derivative of $Dist(\mathbf{c}_T,\hat{\mathbf{c}}_R)$ with respect to α and setting the result to be zero, we have

$$\left\langle(\mathbf{c}_R(0) + \alpha\,\Delta\mathbf{c}_R(\eta^*)),\Delta\mathbf{c}_R(\eta^*)\right\rangle - \left\langle\Delta\mathbf{c}_R(\eta^*),\mathbf{c}_T\right\rangle = 0, \qquad (13)$$

where $\langle\cdot,\cdot\rangle$ is an inner product operator. Therefore, an optimal α is calculated by the following equation

$$\alpha_{opt} = \frac{\left\langle\Delta\mathbf{c}_R(\eta^*),(\mathbf{c}_T - \mathbf{c}_R(0))\right\rangle}{\left\langle\Delta\mathbf{c}_R(\eta^*),\Delta\mathbf{c}_R(\eta^*))\right\rangle}. \qquad (14)$$

The cepstral distance between clean speech and its noisy version increases as the added noise power increasing. An effective adaptation scheme should be able to minimize this distance. Here an experiment was conducted to show the cepstral distance between the adapted clean speech and its noisy version. The white Gaussian noise were added to the clean speech signal in time domain to get its noisy version. The SNR was estimated over whole test speech. The proposed adaptation schemes were performed in several cases. The one termed as Der was that using derivative vector for the adaptation. The others, Dev(X)'s, referred to that using deviation vectors which were generated by (10). In which X indicates SNR, in dB, which was required to produce $\mathbf{c}(\eta^*)$. The average Euclidean cepstral distance between the adapted clean speech and its noisy version are depicted in Figure 1. Obviously, the proposed adaptation scheme can minimize the distance significantly in both situations of using the derivative and deviation vectors. Besides, it also shows that the adaptation can perform better when using the deviation vector than using the derivative one.

3.2. Adaptation in speech recognition

3.2.1. For dynamic time warping (DTW) based speech recognizer

An experiment of isolated Mandarin digit recognition by using a speaker-dependent DTW based speech recognizer was performed to examine the performance of the proposed adaptation method. 12-order LPC derived cepstral coefficients were used as the feature vector. For comparison, two recognition systems with different strategies in distance measure were also examined. One was a baseline system which was with original Euclidean distance measure and the other one was with projected distance measure (PDM). The proposed adaptation schemes were performed in four case, as introduced in the previous section. There were 800 utterances from 10 males and 10 females used as the test patterns in the experiment. The comparison for different systems at various SNR are depicted in Figure 2. The proposed adaptation schemes, Der and Dev(X)'s, improve the accuracy of recognition significantly at all SNRs. However, due to the inherent drawback of curve tracking by a tangent line, the compensation effect is limited for Der.

3.2.2. For hidden Markov model (HMM) based speech recognizer

Another experiment was conducted in an HMM based speech recognizer. In this case, the state deviation vectors are needed for the adaptation scheme. The method for obtaining the state deviation vectors is as follows. Using the HMMs of clean speech, a Viterbi decoding procedure is applied to all of the training utterances to find their state sequences. According to the decoded state sequence, the state deviation vector can be estimated by taking the average of the corresponding deviation vectors, *i.e.*,

$$\Delta\mathbf{u}_i = \frac{1}{N_i}\sum_{k=1}^{N_i}\Delta\mathbf{c}_{i,k}(\eta^*), \qquad (15)$$

where $\Delta\mathbf{c}_{i,k}(\eta^*)$ was a deviation vector decoded to state i and N_i is the total number of $\Delta\mathbf{c}_{i,k}(\eta^*)$. During the recognition phase, the optimal adaptation factor, α_{opt}, is the one that maximizes the log likelihood function of a given state. If the log likelihood function of a state is modeled by a Gaussian distribution,

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

$$\frac{-M}{2}\ln(2\pi) - \frac{1}{2}\ln(|\Sigma_i|) - \frac{1}{2}(\mathbf{c}_t - \hat{\mathbf{u}}_i)^T\,\Sigma_i^{-1}(\mathbf{c}_t - \hat{\mathbf{u}}_i), \quad (16)$$

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 the state vector. Taking derivative of (16) with respect to α and setting the result to zero, we obtain the optimal adaptation factor given by

$$\alpha_{opt} = \frac{(\mathbf{c}_t - \mathbf{u}_i(0))^T\,\Sigma_i^{-1}\Delta\mathbf{u}_i}{\Delta\mathbf{u}_i{}^T\Sigma_i^{-1}\Delta\mathbf{u}_i}. \qquad (17)$$

In the experiment, the HMMs of clean speech were trained by using clean speech data from 25 males and 25 females. The segmental-k-means algorithm was applied to train the HMMs. Speech data from another 25 males and 25 females were used as the test data. The training data and the test data were alternated for obtaining a confident result. 12-order LPC derived cepstral coefficients were used as the feature vector. The covariance matrices of HMMs' states, Σ_i's, were all in the diagonal form for simplicity. There were two mixtures on each state. An HMM based recognizer without adaptation was used as a baseline for comparison. In addition, a system with weighted projection measure (WPM) was also performed. The comparison are observed in Figure 3. Similar to the results of the DTW based speech recognizer, the proposed method can improve the robustness significantly. Besides, both of the results in Figure 2 and Figure 3 reveal that the recognition rate were not sensitive to the pseudo noise power η^* which was needed in the generation of the deviation vectors.

4. CONCLUSION

In this paper, we have addressed the compensation method for the additive white noise by using the derivative or deviation vectors. The derivative vectors are derived from the transformation of the first derivative of AR coefficients, while the deviation vectors are obtained from the difference of the clean cepstrum and its artificial generated noisy version. Both cases are examined by experiments. The performance of using deviation vectors seems superior to that of using derivative vectors. The proposed method is also out performed the projection method with similar computational cost.

REFERENCES

[1] B. H. Juang, "Speech recognition in adverse environments," Computer Speech and Language (1991) 5, pp. 275-294.

[2] D. Mansour and B. H. Juang, "A family of distortion measures based upon projection operation for robust speech recognition," IEEE trans. Acoust. Speech Signal Processing, vol. 37, pp. 1659-1671, November 1989.

[3] Guan, C., Chen, Y. and Wu, B., " Direct modification on LPC coefficients with application to speech recognition, " IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 107-110, 1993.

[4] Lee, L. M. and Wang, H. C., "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.

[5] 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.

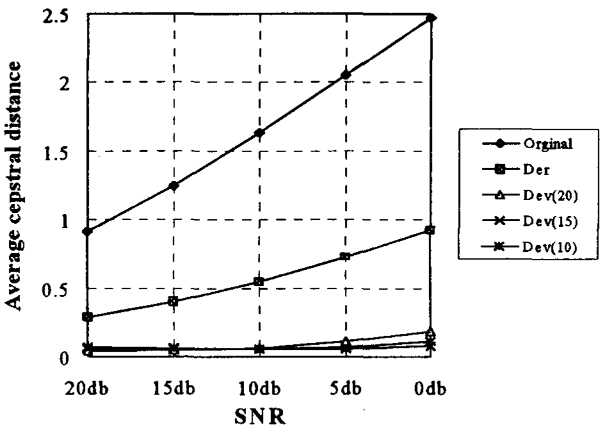

Figure 1 Averaged cepstral distance for various compensation methods

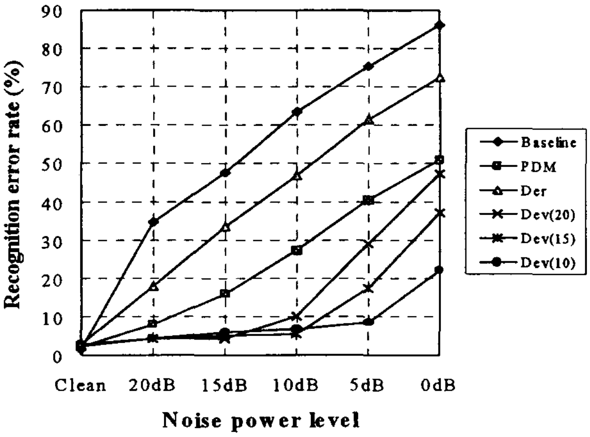

Figure 2, Error rate (%) for digit recognition with a speaker-dependent DTW based speech recognizer

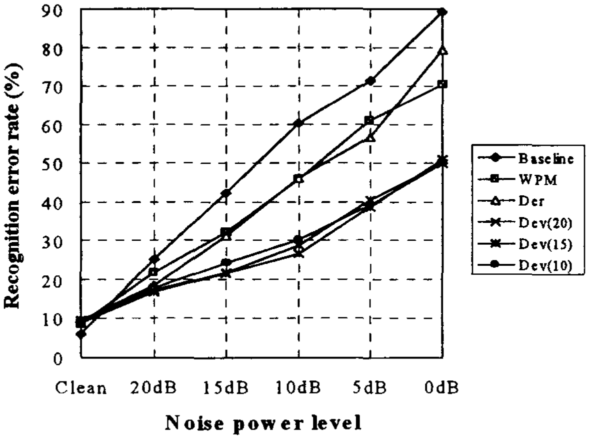

Figure 3, Error rate (%) for digit recognition with a speaker-independent HMM based speech recognizer