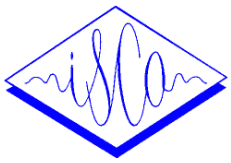

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

A STUDY ON ADAPTATIONS OF CEPSTRAL AND DELTA CEPSTRAL COEFFICIENTS FOR NOISY SPEECH RECOGNITION

Lee-Min Lee and Hsiao-Chuan Wang

Department of Electrical Engineering, National Tsing Hua University
Hsinchu, Taiwan

ABSTRACT

In this study, a family of coefficient adaptation methods for speech recognition under white noise environments is proposed. Based on the property of speech cepstral vector shrinking under white noise influence, the noisy speech reference cepstral vector can be approximated by a linear shrunk version of its clean counterpart. This approximation induces an affine transformation on delta cepstral vector to approximate its noisy version. Using these approximations, an adaptive HMM is proposed for noisy speech recognition. Three alternate adaptation schemes will be also investigated. The adaptation parameters can be determined by searching for optimal values such that the adapted reference is the closest to the test one. In addition, a bilinear function of log-signal-ratio is also proposed to determine the linear shrinking factor. The experimental results show that the proposed adaptation methods can compensate the noise effect.

I. INTRODUCTION

It is the fact that the automatic speech recognition systems developed in quiet environment may degrade drastically for noisy speech recognition. Since the environmental noise may not be known in advance, there is a need to adaptively compensate for the effect of noise. Several techniques have been proposed to compensate for the effect of noise [1–5]. In the study of noisy speech recognition, Mansour and Juang [1] observed two effects of additive white noise on the speech cepstral feature vector: (1) the cepstral vector norm shrink under noisy environment and (2) relative robustness of the cepstral vector orientation. Using these properties, they developed a family of distortion measures based upon projection operation for robust speech recognition. The projection measure can also be interpreted as a linear adaptation method which adapted clean reference cepstral vector to a shrunk version to approximate its behavior under noisy environment and then calculate the distance between the test cepstral vector and the optimal linearly adapted reference one.

In this paper, the concept of cepstral vector shrinking and linear cepstral adaptation is extended to include the adaptation of delta cepstral coefficients under noisy environments. Since the cepstral vector shrinks as noise increases, adaptation of the clean reference cepstral vector to a linear shrunk version is a reasonable approximation of its behavior under noise environments. This approximation induces an affine transformation on delta cepstral vector. Using these approximations, an adaptive HMM performing linear adaptation on cepstral vector and affine adaptation on delta cepstral vector is proposed for noisy speech recognition. In addition, three alternate adaptation schemes will be also investigated, namely linear adaptation on the cepstral vector only, linear adaptation on both cepstral and delta cepstral vector with the same linear shrinking factor, and linear adaptation on both cepstral and delta cepstral vector with two independent shrinking factors. The adaptation parameters can be determined by searching for optimal values such that the adapted reference is the closest to the test one, similar to the method proposed by Mansour and Juang [1]. In addition to this method, we propose a bilinear function (ratio of two linear polynomials) of log-signal-ratio to determine the linear shrinking factor for the noisy cepstral vector. For a correct reference template, it can be expected that appropriately adapted reference will be very close to the test. However, for an incorrect reference, the distance to the test could be large whatever adaptation is done. Experiments on multi-speaker (50 males and 50 females) isolated Mandarin digits recognition are conducted to check and compare the performances of various adaptation schemes. The experimental results show that linear adaptation on cepstral vector and affine adaptation on delta cepstral vector can obtain the best performance.

II. CEPSTRAL AND DELTA CEPSTRAL VECTOR ADAPTATION METHODS AND ADAPTIVE HMM

Let feature vector of reference model, $\mathbf{v}_r = [\mathbf{c}_r, \mathbf{d}_r]$,

This research has been partially supported by Telecommunication Laboratory of MOCT, and National Science Council, Taiwan, R.O.C.

be consisted of 12–dimensional cepstral coefficient vector $\mathbf{c}_r$ and 12–dimensional delta cepstral vector $\mathbf{d}_r$. Let $\mathbf{v}_t = [\mathbf{c}_t, \mathbf{d}_t]$ be a feature vector of a frame of testing utterance. In conventional HMM, to a state (or mixture) with mean vector $[\mathbf{c}_r, \mathbf{d}_r]$, the probability of producing $[\mathbf{c}_t, \mathbf{d}_t]$ output is given by

$$P_{\mathbf{v}_r}(\mathbf{v}_t) = K \cdot \exp\left\{-\frac{1}{2}\left(\|\mathbf{c}_t - \mathbf{c}_r\|_c^2 + \|\mathbf{d}_t - \mathbf{d}_r\|_d^2\right)\right\} \quad (1)$$

where K is the normalization factor and

$$\|\mathbf{c}_t - \mathbf{c}_r\|_c^2 = \sum_{i=1}^{12}\left(\frac{c_{t,i} - c_{r,i}}{\sigma_{ci}}\right)^2 \quad (2)$$

$$\|\mathbf{d}_t - \mathbf{d}_r\|_d^2 = \sum_{i=1}^{12}\left(\frac{d_{t,i} - d_{r,i}}{\sigma_{di}}\right)^2 \quad (3)$$

represent the weighted distances of the cepstral and delta cepstral vectors, respectively, where $\sigma_{c1}\cdots\sigma_{c12}$ and $\sigma_{d1}\cdots\sigma_{d12}$ represent the standard deviation of cepstral and delta cepstral coefficients, respectively. Note that, in the following discussions, we shall neglect the subscript in vector norm and inner product notations when there is no danger of causing confusion.

To compensate for the noise effect, the reference feature vector $[\mathbf{c}_r, \mathbf{d}_r]$ should be adapted to its noisy version $\left[\tilde{\mathbf{c}}_r, \tilde{\mathbf{d}}_r\right]$ according to the noise condition. This results in an adaptive HMM. Since the cepstral vector shrinks as the additive white noise increases, it is a good approximation to adapt the feature vector according to the noise level. Let $\mathbf{c}_r(t)$, $t = 1, 2, \cdots, T$ be a sequence of reference cepstral vector, then its noisy version can be approximated by

$$\tilde{\mathbf{c}}_r(t) = \lambda(t)\mathbf{c}_r(t)\ , \quad (4)$$

where $\lambda(t)$ is the linear shrinking factor at time t. The time derivative of Eq. (4) induces an affine transformation on the reference delta cepstral vector sequence to approximate its behavior under additive white noise environments.

$$\begin{aligned}\tilde{\mathbf{d}}_r(t) &= \delta\tilde{\mathbf{c}}_r(t) = \lambda(t)(\delta\mathbf{c}_r(t)) + (\delta\lambda(t))\mathbf{c}_r(t) \\ &\approx \lambda(t)\mathbf{d}_r(t) + (\delta\lambda(t))\mathbf{c}_r(t)\end{aligned} \quad (5)$$

Together with three alternate forms, four types of feature vector adaptation methods are investigated. These are

1. Linear adaptation on cepstral vector and affine adaptation on delta cepstral vector.
   $\tilde{\mathbf{v}}_r = \left[\tilde{\mathbf{c}}_r, \tilde{\mathbf{d}}_r\right] = [\lambda\mathbf{c}_r, \lambda\mathbf{d}_r + (\delta\lambda)\mathbf{c}_r]$
2. Linear adaptation on cepstral vector only.
   $\tilde{\mathbf{v}}_r = \left[\tilde{\mathbf{c}}_r, \tilde{\mathbf{d}}_r\right] = [\lambda\mathbf{c}_r, \mathbf{d}_r]$
3. Linear adaptation on both cepstral vector and delta cepstral vector with the same adaptation parameter.
   $\tilde{\mathbf{v}}_r = \left[\tilde{\mathbf{c}}_r, \tilde{\mathbf{d}}_r\right] = [\lambda\mathbf{c}_r, \lambda\mathbf{d}_r]$
4. Linear adaptation on both cepstral vector and delta cepstral vector with two independent parameters.
   $\tilde{\mathbf{v}}_r = \left[\tilde{\mathbf{c}}_r, \tilde{\mathbf{d}}_r\right] = [\lambda_1\mathbf{c}_r, \lambda_2\mathbf{d}_r]$

The adaptation parameters can be determined either by searching for optimal values such that the adapted reference is the closest to the test one, or by using a bilinear function of log-signal-ration estimated from the test utterance.

A. Determination of Adaptation Parameter by Using Bilinear Function of Log-signal-ratio

For the types 1 ,2 , and 3, the only adaptation parameter is the cepstral vector shrinking factor. Since the cepstral vector shrinks as noise increases, we propose a bilinear function of log-signal-ratio (the logarithmic power ratio of the clean speech to the noisy speech) to model the shrinking factor, i.e.,

$$\lambda = f(r) = \frac{mr + \alpha}{r + \alpha}\ , \quad (6)$$

where r represent the estimated log-signal-ratio according to the test utterance, m the lower bound of shrinking factor, and α approximately the half shrinking point when $m << 1$. Experimentally, the parameters m and α can be chosen as $m = 0.35$ and $\alpha = -0.005$, respectively. For type 1 adaptation, we can first apply the above equation to get the shrinking factor $\lambda(t)$ at each time index $t = 1, 2, \cdots, T$, and then calculate the delta shrinking factor.

B. Determination of Adaptation Parameter by Best Fitting of the Reference Feature Vector to the Test Speech

An alternate way to determine the adaptation parameters is by searching for optimal values such that the adapted reference feature vector is the closest to the test one. From Eq. (1), one can see that the distance between testing feature vector and reference feature vector is given by

$$J(\mathbf{v}_t, \tilde{\mathbf{v}}_r) = \|\mathbf{c}_t - \tilde{\mathbf{c}}_r\|^2 + \left\|\mathbf{d}_t - \tilde{\mathbf{d}}_r\right\|^2 \quad (7)$$

Since the adaptation types 2, 3, and 4 are linear transformations, Eq. (7) will be in a quadratic form. We can obtain the unique solution by taking the partial derivatives with respect to the adaptation parameters and setting to zeros. For type 2 adaptation, the distance is

$$J(\mathbf{v}_t, \tilde{\mathbf{v}}_r) = \|\mathbf{c}_t - \lambda\mathbf{c}_r\|^2 + \|\mathbf{d}_t - \mathbf{d}_r\|^2 \quad (8)$$

The optimal λ is obtained by solving the equation

$$\frac{dJ}{d\lambda} = \frac{d}{d\lambda}\|\mathbf{c}_t - \lambda\mathbf{c}_r\|^2 = \frac{d}{d\lambda}\langle\mathbf{c}_t - \lambda\mathbf{c}_r, \mathbf{c}_t - \lambda\mathbf{c}_r\rangle = 0 \quad (9)$$

to obtain

$$\lambda = \langle\mathbf{c}_t, \mathbf{c}_r\rangle / \langle\mathbf{c}_r, \mathbf{c}_r\rangle \quad (10)$$

Note that the solution obtained by the above formula can not be guaranteed to lie in the interval [0, 1]. Experimental results show that λ may become a small negative number at silence frames or when the speech signal is very weak.

Similarly, for type 3 adaptation,

$$\lambda = \frac{\langle \mathbf{c}_t,\mathbf{c}_r\rangle + \langle \mathbf{d}_t,\mathbf{d}_r\rangle}{\langle \mathbf{c}_r,\mathbf{c}_r\rangle + \langle \mathbf{d}_r,\mathbf{d}_r\rangle} = \frac{\langle \mathbf{v}_t,\mathbf{v}_r\rangle}{\langle \mathbf{v}_r,\mathbf{v}_r\rangle} \qquad (11)$$

For type 4,

$$\lambda_1 = \frac{\langle \mathbf{c}_t,\mathbf{c}_r\rangle}{\langle \mathbf{c}_r,\mathbf{c}_r\rangle} \ , \ \lambda_2 = \frac{\langle \mathbf{d}_t,\mathbf{d}_r\rangle}{\langle \mathbf{d}_r,\mathbf{d}_r\rangle} \qquad (12)$$

In type 1 adaptation, it requires formidable amount of computations to obtain the global optimal shrinking sequence, $\lambda(t)$, $t = 1,2,\cdots,T$. This is because that $\delta\lambda(t)$ must be calculated from a context window of several frames and the forward scoring procedure is not applicable. Hence, we use a scaled version of delta log energy, which is obtained by an average of delta log energy from the clean training data belonging to the same state, to approximate $\delta\lambda$. The scaling factor is chosen to be 0.075, empirically.

III. EXPERIMENTS

A multispeaker (50 males and 50 females) speech recognition task for 10 isolated Mandarin digits is conducted to investigate the proposed noise adaptation methods. The results are also compared with a base line speech recognition system. The speech data is collected in a sound treated environment and sampled at 8KHz. They are referred as clean speech. There are three sessions of data collection. In each session, a speaker utters a set of 10 Mandarin digits. Two sessions are used for training and the other session is for testing. Endpoints are roughly detected so that each utterance still contains short periods of pre-silence and post-silence. The white Gaussian noise is artificially generated by computer and added to clean speech with specific SNR values. Each digit is modeled by a left to right HMM without jumps. Each HMM contains seven to nine states depending on its average duration and begins with an pre-silence state and ends with a post-silence state. The silence states for all digits are tied together, i.e., share the same statistic parameters. The model parameters are trained by the segmental k-means algorithm using clean speech data.

A. Determination of Adaptation Parameter by Using Bilinear Function of Log-signal-ratio

The frame with minimum power in a test utterance is considered as a noise frame. The noise power is calculated based on this frame. The speech power is calculated by subtracting the estimated noise power from the total noisy speech power. The shrinking factor for cepstral vector is obtained by applying the bilinear function to the log-signal-ratio as described in the previous section. The experimental results of the baseline system and adaptive systems of type 1, 2, 3 for single Gaussian state output probability distribution are listed in table 1.

	Base line	Type 1	Type 2	Type 3
clean	98.0	97.5	97.7	97.6
20dB	78.3	92.3	91.3	91.9
15dB	61.6	85.4	84.2	84.6
10dB	45.7	72.2	71.2	71.5
5dB	27.3	52.7	50.8	52.8

Table 1 Experimental results using single mixture and bilinear function of estimated log-signal-ratio as the adaptation parameter

From the above table, we can see that the recognition performance can be greatly improved by performing the linear adaptation on cepstral vector only (type 2). Further improvement can be obtained by applying linear adaptation on both cepstral vector and delta cepstral vector with the same adaptation parameter (type 3). Still further improvement can be achieved by linear adaptation on cepstral vector and affine adaptation on delta cepstral vector (type 1).

The experimental results for four Gaussian mixtures are listed in table 2.

	Base line	Type 1	Type 2	Type 3
clean	98.5	98.1	98 .3	98.1
20dB	83.4	92.8	93.8	92.9
15dB	68.4	85.7	86.8	85.6
10dB	47.3	74.3	73.2	73.5
5dB	27.6	56.6	52.6	56.1

Table 2 Experimental results using four mixtures and bilinear function of estimated log-signal-ratio as the adaptation parameter

We can see that linear adaptation on cepstral vector and affine adaptation on delta cepstral vector can obtain the best result at low SNR. However, when the SNR is high, linear adaptation on cepstral vector only (type 2) has the best performance. The reason can be explained as follows. The linear approximation model of cepstral vector adaptation will introduce some model errors. These errors will be not only propagated to the affine adaptation of delta cepstral vector but also amplified by the associated differential operation. Hence, the advantage of delta cepstral vector adaptation could be overwhelmed by the introduced errors. However, a comparison of type 2 and type 3 adaptations show that the translation part of delta cepstral vector adaptation does have some contribution. This suggests that a compromise on the delta cepstral vector shrinking factor may get a better result. Hence, we shall investigate the following modified adaptations.

M1. $\tilde{\mathbf{v}}_r = [\lambda\mathbf{c}_r, (0.5 + 0.5\lambda)\mathbf{d}_r + (\delta\lambda)\mathbf{c}_r]$

M3. $\tilde{\mathbf{v}}_r = [\lambda \mathbf{c}_r, (0.5 + 0.5\lambda)\mathbf{d}_r]$

The experimental results of the above modified adaptations (together with baseline system and type 2 adaptation) are listed in table 3. From the table 3, we can see that type M1 adaptation has the best performance.

	Base line	Type M1	Type 2	Type M3
clean	98.5	98.1	98 .3	98.3
20dB	83.4	93.8	93.8	93.5
15dB	68.4	87.0	86.8	86.7
10dB	47.3	74.6	73.2	73.9
5dB	27.6	55.8	52.6	55.9

Table 3 Experimental results using four mixtures with modified adaptation

B. Determination of Adaptation Parameters by the Best Fitting of the Reference Feature Vector to the Test Speech

For this class of adaptation, an estimation of the noise level from the test utterance is not required. Instead, the adaptation parameters are determined by using the prior knowledge of reference speech feature vector and the linear shrinking model as described in the previous section. The experimental results for one and four Gaussian mixtures are listed in table 4 and 5, respectively.

	Base line	Type 1	Type 2	Type 3	Type 4
clean	98.0	97.3	98.1	97.9	97.6
20dB	78.3	91.8	91.5	91.5	91.0
15dB	61.6	82.1	82.9	82.1	80.2
10dB	45.7	67.6	69.2	68.5	64.3
5dB	27.3	50.7	50.4	49.4	43.3

Table 4 Experimental results using single mixture and best fitting adaptation parameter

	Base line	Type 1	Type 2	Type 3	Type 4
clean	98.5	97.9	98.0	97.8	97.5
20dB	83.4	94.3	93.1	92.4	89.8
15dB	68.4	86.6	84.6	84.8	79.5
10dB	47.3	74.4	70.3	71.5	64.9
5dB	27.6	57.4	49.2	53.8	45.4

Table 5 Experimental results using four mixtures and best fitting adaptation parameter

From the above tables, we can see that the best performance is achieved by the one-pass type 1 adaptation at four mixtures. The poor performance of the type 4 adaptations account for the using of two independent linear shrinking factors (for the adaptation of cepstral and delta cepstral vectors respectively) so that some incorrect reference models will gain more chance of being adapted to well match the test speech. For example, a reference feature vector may be adapted to a match condition with large λ_1 and small λ_2, which is not reasonable.

IV. CONCLUSIONS

In this study, a family of cepstral and delta cepstral coefficient adaptation methods for noisy speech recognition is proposed and investigated. The cepstral vector shrinking property is used to derive the adaptation formula. The linear shrink approximation of noisy speech cepstral vector has led to linear adaptation of cepstral vector and affine adaptation of delta cepstral vector. The cepstral vector shrinking factor, which provides the mechanism of adaptation, can be determined either by searching for optimal values such that the adapted reference is the closest to the test one, or by using a bilinear function of log-signal-ratio estimated from the test utterance. Three alternate forms of adaptations are also investigated. The HMM speech recognition systems incorporated with the proposed feature vector adaptation methods are used to check and compare the recognition performances under various noisy environments. Experiments on multi-speaker isolated Mandarin digits recognition show that the proposed methods can compensate the effect of noise. Experiments also show that the best performance is achieved by linear adaptation of cepstral vector and affine adaptation of delta cepstral vector.

References

[1] D. Mansour and B. H. Juang. A family of distortion measures based upon projection operation for robust speech recognition. *IEEE Trans. ASSP*, Vol.37(No.11):pp.1659–1671, November 1989.

[2] A. Nadas, D. Nahamoo, and M. A. Picheny. Speech recognition using noise-adaptive prototypes. *IEEE Trans. ASSP*, Vol.37(No.10):pp.1495–1503, October 1989.

[3] A. Acero and R. M. Stern. Environmental robustness in automatic speech recognition. In *proc. ICASSP.*, pages 849–852, April 1990.

[4] A. P. Varga and R. K. Moore. Hidden markov model decomposition of speech and noise. In *proc. ICASSP.*, pages pp.845–848, 1991.

[5] L. M. Lee, J. K. Chen, and H. C. Wang. Noise adaptive cepstral coefficients and its application to noisy speech recognition. In *Proc. International Symposium on Speech, Image Processing & Neural Networks*, pages 347–350, April 1994.