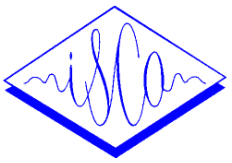

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

COMBINATION OF TEMPORAL TRAJECTORY FILTERING AND PROJECTION MEASURE FOR ROBUST SPEAKER IDENTIFICATION

Kuo-Hwei Yuo, Tai-Hwei Hwang, Hsiao-Chuan Wang*

Department of Electrical Engineering, National Tsing Hua University,
Hsinchu, Taiwan 300

*E000/Computer & Communication Labs, Industrial Technology Research Institute,
Chutong, Hsinchu, Taiwan 300

E-mail: hcwang@ee.nthu.edu.tw

ABSTRACT

This paper presents a method that combines the techniques of temporal trajectory filtering and projection measure for robust speaker identification. The proposed robust feature, called Relative Autocorrelation Sequence Mel-scale Frequency Cepstral Coefficients (RAS-MFCC), is derived based on filtering the temporal trajectories of short-time one-sided autocorrelation sequences. This filtering process can minimize the effect of additive noise in the noisy speech. Since the norm of RAS-MFCC shrinks due to noise corruption, the projection measure (PM) technique, which is effective in dealing with the norm shrinkage of cepstrum, can be applied for the distance measure of RAS-MFCCs. The combination of these two techniques is then applied to a task of speaker identification of 100 speakers. Our experiment shows that the use of RAS-MFCC feature achieves significant improvement in identification rate as comparing with the use of MFCC. The combination of RAS-MFCC feature with PM technique can further improve the recognition accuracy.

1. INTRODUCTION

The performance of a speech or speaker recognition system may drastically degrade when there is the mismatch between training and test environments. The mismatch between training and test environments is often due to the additive noises in noisy speech. Many techniques have been proposed to overcome this degradation problem [1]. The robustness of the recognizer can be accomplished in three ways: (1) using speech enhancement technique to increase the signal to noise ratio (SNR) [2], (2) extracting the robust parametric representation of speech signal to minimize the effect of noise to the speech [3-4], and (3) using model adaptation technique to dynamically adapt clean speech models to the noisy environment [5-7].

Although a variety of techniques can demonstrate the comparable performance, some weaknesses may limit their practical applications. For example, the spectral subtraction [2] and the parallel model combination (PMC) [5] need a priori knowledge of the noise characteristics.

With a concept similar to RASTA [3] approach, we propose a method to remove noise effect based on the idea of temporal filtering in autocorrelation domain [8]. The filtered sequences are named Relative Autocorrelation Sequences (RAS). We regard the RAS as another representation of speech. Then the mel-scale frequency cepstral coefficients (MFCC) are extracted from the RAS and denoted as RAS-MFCC. We have applied the RAS-MFCC in speech recognition [8]. This paper extends the application of RAS-MFCC to the task of speaker identification.

It is well known that the additive white noise reduces the norm of cepstral feature vector. The projection measure (PM) is an effective compensation technique for cepstral distance measure that takes into account the norm shrinkage [6-7]. The advantage of this technique is to adapt reference template to noisy environment without requiring explicit knowledge of noise. Since the norm of RAS-MFCC also shrinks due to noise corruption, the projection measure technique can be applied for the distance measure of RAS-MFCCs. The combination of RAS-MFCC feature with PM technique can further improve the recognition accuracy.

2. RELATIVE AUTOCORRELATION SEQUENCE (RAS)

Let m be the frame index and n be the sample index within a frame. The clean speech $x(m,n)$ corrupted by the additive noise $w(m,n)$ results in a noisy speech

$$y(m,n) = x(m,n) + w(m,n), \quad (1)$$

$m = 0, 1, 2, \ldots, M\text{-}1$, and $n = 0, 1, 2, \ldots, N\text{-}1$

where M denotes the total frames and N denotes the total samples in a frame.

If the noise is stationary and uncorrelated with the speech, it follows that the autocorrelation of the noisy speech $y(m,n)$ is the sum of autocorrelation of the clean speech $x(m,n)$ and autocorrelation of the noise $w(m,n)$, i.e.,

$$r_{yy}(m,k) = r_{xx}(m,k) + r_{ww}(m,k), \quad (2)$$

where $r_{yy}(m,k)$, $r_{xx}(m,k)$ and $r_{ww}(m,k)$ are the one-

sided autocorrelation sequences of the noisy speech, clean speech, and noise, respectively, and k is the autocorrelation sequence index. Moreover, if the noise is stationary, the autocorrelation sequences of noise in all frames can be assumed to be identical and $r_{ww}(m,k)$ will depend only on autocorrelation index k. Hence, we drop the index m of $r_{ww}(m,k)$, and Eq. (2) becomes

$$r_{yy}(m,k) = r_{xx}(m,k) + r_{ww}(k), \tag{3}$$

where the N-point $r_{yy}(m,k)$ is computed from N-point $y(m,n)$ using

$$r_{yy}(m,k) = \sum_{j=0}^{N-1-k} y(m,j)y(m,j+k),. \tag{4}$$

Taking difference of both sides of Eq. (3) with respect to frame index m for all k yields

$$\Delta r_{yy}(m,k) = \Delta r_{xx}(m,k), \quad 0 \le m \le M-1, 0 \le k \le N-1 \tag{5}$$

where

$\Delta r_{yy}(m,k) = r_{yy}(m+1,k) - r_{yy}(m-1,k)$, and

$\Delta r_{xx}(m,k) = r_{xx}(m+1,k) - r_{xx}(m-1,k)$.

The sequence, $\{\Delta r_{yy}(m,k)\}_{k=0}^{N-1}$, is named the Relative Autocorrelation Sequence (RAS) of noisy speech. Eq. (5) demonstrates that the relative autocorrelation sequence of noisy speech is equal to the relative autocorrelation sequence of clean speech. This implies that the RAS is a robust representation of speech within which the corruption of additive noise is inherently removed. We consider the RAS a new time-domain signal. We compute mel-frequency cepstral coefficients (MFCC) from the RAS sequence. The MFCC extracted from the RAS is called RAS-MFCC. This feature is robust to additive noise.

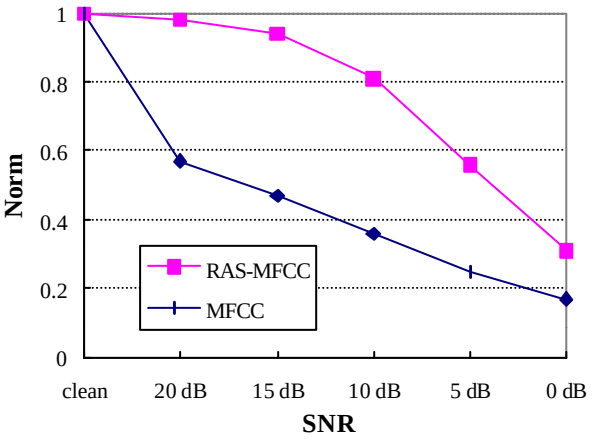

Figure 1: The cepstral norm shrinkage for MFCC, and RAS-MFCC in additive white noise.

3. GMM SPEAKER MODEL AND PROJECTION MEASURE (PM)

The additive white noise reduces the norm of cepstral feature vector. As a consequence of this result, the projection measure (PM) was proposed to adapt the reference templates to testing feature [6-7]. By investigating the property of RAS-MFCC, it shows that the norm of RAS-MFCC also shrinks due to the additive noise. Figure 1 illustrates the norms of the MFCC and RAS-MFCC in additive white noise at different SNR levels. Here, all norms are normalized to norms of clean speech. We can see that the norm shrinkage RAS-MFCC is somewhat moderate as comparing with MFCC. Because of this particular property, it is possible to apply the projection measure on the RAS-MFCC for robust speech of speaker recognition.

A Gaussian mixture density is a weighted sum of M component densities, and given by equations

$$p(x_t|\lambda) = \sum_{i=1}^{M} w_i N(x_t, v_i, \Lambda_i) = \sum_{i=1}^{M} \frac{w_i}{(2\pi)^{D/2}|\Lambda_i|^{1/2}} \exp\left\{-\frac{1}{2}\|x_t - v_i\|_{\Lambda_i^{-1}}\right\} \tag{6}$$

and

$$\|x_t - v_i\|_{\Lambda_i^{-1}} \triangleq (x_t - v_i)^{\mathrm{T}} \Lambda_i^{-1} (x_t - v_i) \tag{7}$$

where w_i, v_i, and Λ_i are weight, mean vector, and covariance matrix of the ith Gaussian distribution, respectively. When the testing utterance x_t is corrupted by noise, a modified scoring density based on the PM method is

$$P^{\mathrm{PM}}(x_t|\lambda) \triangleq \sum_{i=1}^{M} \frac{w_i}{(2\pi)^{D/2}|\Lambda_i|^{1/2}} \exp\left\{-\frac{1}{2}\|x_t - v_i\|_{\Lambda_i^{-1}}^{\mathrm{PM}}\right\}, \tag{8}$$

where

$$\|x_t - v_i\|_{\Lambda_i^{-1}}^{\mathrm{PM}} \triangleq \min_{\alpha} \|x_t - \alpha v_i\|_{\Lambda_i^{-1}}. \tag{9}$$

Here, α is a shrinkage factor that is dynamically determined during the distance measure of testing utterance and reference models.

3. EXPERIMENTS AND DISCUSSION

The database for the experiment of speaker identification is a 100-speaker Mandarin digit database. This database provided by 50 males and 50 females in five recording sessions contains 20000 (100 speakers × 5 sessions× 40 utterances) continuous Mandarin digit utterances. Each speaker in each recording session randomly selected an utterance table from 340 utterance tables and uttered 40 utterances in the chosen table. Five sessions were recorded over a half-month period. The database was recorded in the Chung-Hwa Telecommunication Laboratories in Taiwan. The utterances were recorded via high quality microphones in quiet environments. Since the

background noise and the channel distortion are minimized, the recorded speech is referred to the clean speech.

The database was partitioned into two parts. The first three sessions of each speaker were used for training the speaker model that was represented by a mixture of 64 Gaussian components with diagonal covariance matrix. The recorded speech of each speaker in each session is about 30 seconds and thus each speaker model is trained using 90 seconds of speech data. Note that each speaker model is trained on clean speech only. The 7-digit utterances in the remained two sessions were used for testing. Totally there are 1000 (5×2×100) test utterances. A 7-digit utterance is about two seconds in length.

The speech signal was sampled at a 10 kHz sampling rate, and weighted by a 25.6 ms Hamming window shifted every 12.8 ms. In computing the MFCC, a 20-channel filter bank with mel-scale frequency is applied. The log-energy outputs of the filter bank were transformed into a set of 14 cepstral coefficients. The 14 delta cepstral coefficients are computed in the span of five frames [9]. Thus each feature vector comprised 14 cepstral coefficients and 14 delta cepstral coefficients. In computing the RAS-MFCC, the one-side autocorrelation sequence is used instead of the original speech signal.

3.1. Testing on clean speech

This experiment is to evaluate the performance of MFCC and RAS-MFCC when the training data and the testing utterances are in matched environment. Two types of utterance lengths, 7 digits and 14 digits, are examined. We use 1000 7-digit utterances in the last two sessions and 500 14-digit utterances in the speaker identification test. The 14-digit utterances are the concatenation of two 7-digit utterances. The result is shown in Table 1. Both cases of using static features and using static plus dynamic features are evaluated. The former means the use of the MFCC or the RAS-MFCC. The latter means the use of MFCC plus delta MFCC, or RAS-MFCC plus delta RAS-MFCC. It shows more improvement in using short utterances than in using long utterances. When 14-digit utterances are used for testing, the add of dynamic features does not gain any improvement.

3.2. Testing on speech corrupted by additive noise

The testing speech is polluted by additive noise. This is a mismatched condition between the training and testing environments. The test data are 500 14-digit utterances. The feature vector is consisted of 14 static components and 14 dynamic components. In order to simulate various cases of noise corruption, the noise signal is artificially added to the test utterances. The white noise, the factory noise, the F16 cabin noise, and the babble noise are added to the clean speech in five SNR's. The white noise is artificially generated while the colored noises are extracted from NOISEX-92 database [10]. the combination of the projection measure with the RAS-MFCC is evaluated. The result is shown in Figure 2.

When the static and dynamic terms of MFCC are used to form the feature vectors, the corruption of white noise makes a bigger degradation than that of colored noises in the identification rate. The PM algorithm does give some improvement. When RAS-MFCC is used, the improvement is obvious in cases of SNR equal to 15 db and 10 dB (denoted by RAS in Figure 2). The biggest improvement is for the case of white noise corruption. It seems that the assumption of stationary and uncorrelated conditions on the additive noise in deriving the RAS has well described the property of white noise. When the PM algorithm is combined with the use of RAS-MFCC features, the performance can be further improved.

Table 1: Identification rate (%) for testing on clean speech.

	Static features		**Static+Dynamic features**	
Dimension of feature vector	14		28	
Length	7-digit	14-digit	7-digit	14-digit
MFCC	98.3	99.4	98.7	99.4
RAS-MFCC	95.8	98.0	96.2	98.0

4. ACKNOWLEDGEMENT

This research has been partially sponsored by the National Science Council of Taiwan, under contract number NSC-89-2614-E-007-002.

5. REFERENCES

1. Gong, Y., "Speech Recognition in Noisy Environments: A Survey," *Speech Communication*. Vol. 16, pp. 261-291, Apr. 1995.

2. Boll, S. F., "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech. Signal Processing*, vol. 27. pp. 113-120, Apr. 1979.

3. Hermansky, H. and Morgan, N., "RASTA Processing of Speech," *IEEE Trans. Speech Audio processing*, vol. 2, pp. 578-589, October, 1994.

4. Hernando, J. and Nadeu, C., "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.

5. Gales, M. J. F. and Young, S. J., "Robust Speech Recognition in Additive and Convolutional Noise Using Parallel Model Combination", *Computer Speech and Language*, pp. 289-307, Sep. 1995.

6. Mansour, D. and Juang, B. H., "A family of distortion measures based upon projection operation for robust speech recognition", *IEEE Trans. Acoust., Speech.*

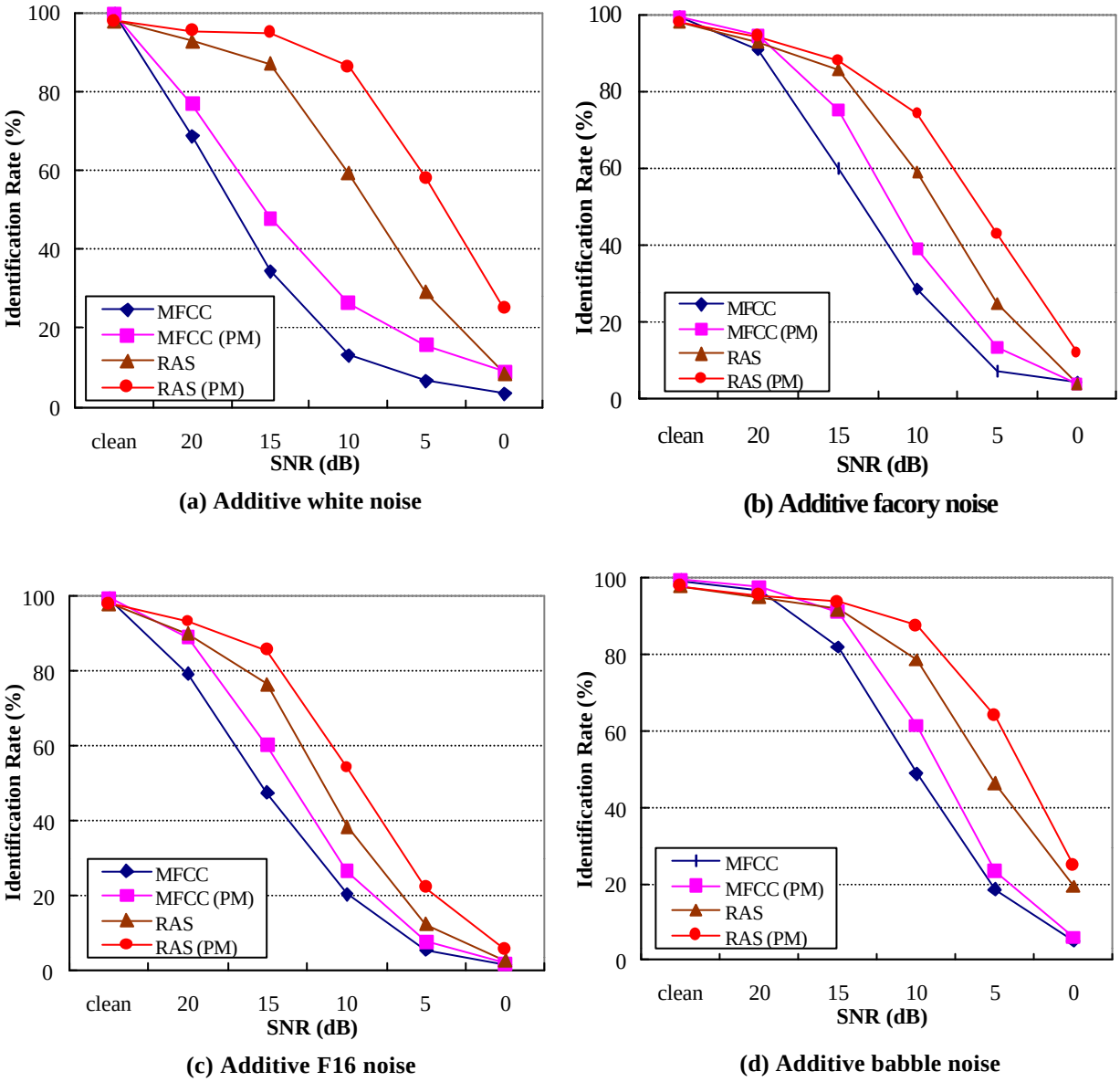

Figure 2: Identification rate (%) for testing on speech corrupted by noises

Signal Processing, vol. 37, no. 11, pp. 1659-1671, Nov. 1989.

7. Carlson, B. A. and Clements, M. A., "A projection-based likelihood measure for speech recognition in noise," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 2, no. 1, part 1, pp. 97-102, Jan. 1994.
8. You, K. H. and Wang, H. C., "Robust Features for Noisy Speech Recognition Based on Temporal Trajectory Filtering of Short-Time Autocorrelation Sequences", *Speech Communication* 28 (1999) 13-24.
9. Furui, S., "Speaker-independent isolated word recognition based on emphasized spectral dynamics", Proc. IEEE Internat. Conf Acoust.. Speech, Signal Process. ' 86, Tokyo, April 1986, pp. 1991-1994, 1986.
10. Varga, A. & Steeneken, H. J. M., "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication* 12, pp. 247-251, 1993.