Combination of Autocorrelation-Based Features and Projection Measure Technique for Speaker Identification

Kuo-Hwei Yuo, Tai-Hwei Hwang, and Hsiao-Chuan Wang, *Senior Member, IEEE*

***Abstract*—This paper presents a robust approach for speaker identification when the speech signal is corrupted by additive noise and channel distortion. Robust features are derived by assuming that the corrupting noise is stationary and the channel effect is fixed during an utterance. A two-step temporal filtering procedure on the autocorrelation sequence is proposed to minimize the effect of additive and convolutional noises. The first step applies a temporal filtering procedure in autocorrelation domain to remove the additive noise, and the second step is to perform the mean subtraction on the filtered autocorrelation sequence in logarithmic spectrum domain to remove the channel effect. No prior knowledge of noise characteristic is necessary. The additive noise can be a colored noise. Then the proposed robust feature is combined with the projection measure technique to gain further improvement in recognition accuracy. Experimental results show that the proposed method can significantly improve the performance of speaker identification task in noisy environment.**

***Index Terms*—Channel-normalization, relative autocorrelation sequence, projection measure, speaker identification.**

I. Introduction

THE environmental mismatch between training and testing data will drastically degrade the performance of speech or speaker recognition systems. The performance degradation is mainly due to the background noise and the channel distortion. Many techniques have been proposed to overcome this degradation problem [1], such as parallel model combination (PMC) [2], stochastic matching (SM) algorithm [3], [4], combining channel identification with power spectrum estimation [5], joint additive and convolutive bias compensation [6], and etc. Recently, several novel techniques for handset and channel compensation are proposed for the speaker recognition. Murthy *et al.* [7] presented both feature-based and model-based techniques to compensate for channel effects. In feature-based approach, they optimized the front-end processing, modified the filter bank computation, and thus introduced a new robust feature. In model-based approach, they utilized a local stereo database to estimate a speaker-independent transformation to other databases. Heck *et al.* [8] proposed a robust feature using a nonlinear artificial neural network to optimize the speaker recognition performance. The technique required neither stereo recordings nor labeling of handset types. Sönmez *et al.*. [9] introduced a normalization technique to compensate for the handset variation via a parameter interpolation extension of HNORM [10], [11] method. This technique was applied to a speaker tracking and detection system.

Although these techniques demonstrate the comparable performance, some weaknesses may restrict their practical applications. For examples, the PMC method needs a prior knowledge of the noise derived from nonspeech period and the SM algorithm needs time for iterative estimation of noise statistics from testing utterances. In this paper, we propose a feature that is inherently robust to noise. The feature is derived from the autocorrelation sequence of noise corrupted speech signal. A two-step temporal filtering algorithm is proposed to minimize the effect of additive noise and channel distortion. The filtering technique has been applied in several domains, such as using a high-pass filter in the subband domain [12] for the reduction of additive noise, using a band-pass IIR filter in the logarithmic subband domain [13] for the removal of convolutional noise, and using FIR filters in the DFT spectrum domain [14] for the dereverberation of speech signal. In this paper, the temporal filtering technique is applied to autocorrelation sequence domain and its logarithmic spectrum domain.

When a speech is corrupted by the additive noise, the noise component is additive to the speech not only in the power spectral domain, but also in the autocorrelation domain. Instead of subtracting the noise in power spectral domain, we remove the additive noise in autocorrelation domain based on the temporal trajectory filtering. This filtered sequence is named the Relative Autocorrelation Sequence (RAS). If the speech signal is also distorted by the channel, the RAS is transformed to logarithmic spectrum domain to remove the channel effect by mean subtraction operation. This new sequence is called the CHannel-Normalization Relative Autocorrelation Sequence (CHNRAS). We regard RAS and CHNRAS as alternative representations of speech signal. Then the mel-scale frequency cepstral coefficients (MFCCs) of RAS and CHNRAS are derived. These MFCCs are denoted as RAS-MFCC and CHNRAS-MFCC. A preliminary study on the application of RAS-MFCC has been

Manuscript received September 11, 2000; revised July 1, 2004. This work was supported in part by the National Science Council of Taiwan, R.O.C., under Contract NSC-89-2614-E-007-002. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Anath Sankar.

K.-H. Yuo is with the Chung-Shan Institute of Science and Technology, Tao-Yuan 325, Taiwan, R.O.C. (e-mail: gwohuei.you@msa.hinet.net).

T.-H. Hwang is with the Computer and Communication Laboratories, Industrial Technology Research Institute (ITRI), Taiwan, R.O.C. (e-mail: hthwei@itri.org.tw).

H.-C. Wang is with the Department of Electrical Engineering, National Tsing Hua University, Hsinchu 300, Taiwan, R.O.C. (e-mail: hcwang@ee.nthu.edu.tw).

Digital Object Identifier 10.1109/TSA.2005.848893

reported [15], [16] and compared with some robust features that are also derived from the autocorrelation domain [17], [18].

It is well known that the norm of cepstral vector of speech signal shrinks due to the additive noise. The projection measure is a robust compensation technique that takes into account the norm shrinkage [19], [20]. This technique can adapt the reference template to noisy environment without requiring the explicit knowledge of the noise. In the investigation of the properties of RAS-MFCC and CHNRAS-MFCC, we find that both of these two feature vectors also shrink due to additive noise. By taking the advantage of projection measure technique, we combine the proposed robust features, RAS-MFCC, and CHNRAS-MFCC, with the projection measure to further improve the recognition accuracy.

The remainder of the paper is organized as follows. The mathematical fundamentals of RAS and CHNRAS are described in Sections II and III, respectively. In Section IV, we investigate the properties of RAS-MFCC and CHNRAS-MFCC. The combination of RAS-MFCC or CHNRAS-MFCC with projection measure is described in Section V. In Section VI, a series of experiments on the task of speaker identification under various noisy environments is conducted to evaluate the performance of the proposed algorithm. Finally, a conclusion is given in Section VII.

II. Relative Autocorrelation Sequence (RAS)

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

$$y(m,n) = x(m,n) + w(m,n), \quad 0 \le m \le M-1, \quad 0 \le n \le N-1, \quad (1)$$

where M denotes the number of frames in an utterance and N denotes the number of samples in a frame.

If the noise is 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.,

$$r_{yy}(m,k) = r_{xx}(m,k) + r_{ww}(m,k), \quad 0 \le m \le M-1, \quad 0 \le k \le N-1, \quad (2)$$

where $r_{yy}(m,k), r_{xx}(m,k)$, and $r_{ww}(m,k)$ are the *one-sided autocorrelation sequences* of noisy speech, clean speech, and noise, respectively, and k is the autocorrelation sequence index. If the noise is stationary, the autocorrelation of noise in all frames can be assumed to be identical. Hence, the index m of $r_{ww}(m,k)$ can be dropped out, and (2) becomes

$$r_{yy}(m,k) = r_{xx}(m,k) + r_{ww}(k), \quad 0 \le m \le M-1, \quad 0 \le k \le N-1. \quad (3)$$

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

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

Applying the temporal filtering on both sides of (3), it comes out

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

where

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

The sequence, $\{\Delta r_{yy}(m,k)\}_{k=0}^{N-1}$, is named the Relative Autocorrelation Sequence (RAS) of noisy speech at the mth frame. (5) demonstrates that, in each frame, the RAS of noisy speech is equal to the RAS of clean speech. This implies that the RAS is a robust representation of speech signal within which the corruption of a stationary noise is removed. We consider the RAS an alternative representation of speech signal in time-domain that is robust to additive noise corruption.

III. Channel-Normalization Relative Autocorrelation Sequence (CHNRAS)

Assume that the clean speech $x(m,n)$ is corrupted by additive noise $w(m,n)$ and then distorted by a channel $h(n)$. Then the final noisy speech $y(m,n)$ becomes

$$y(m,n) = [x(m,n) + w(m,n)] \otimes h(n), \quad 0 \le m \le M-1, \quad 0 \le n \le N-1, \quad (6)$$

where "$\otimes$" denotes the convolution operation. Note that we assume the channel effect is fixed in an utterance, and thus $h(n)$ is independent of frame index m. If $x(m,n), w(m,n)$ and $h(n)$ are uncorrelated, the two-sided autocorrelation of the noisy speech is expressed as

$$r_{yy}(m,k) = [r_{xx}(m,k) + r_{ww}(m,k)] \otimes h(k) \otimes h(-k), \quad 0 \le m \le M-1, \quad 0 \le k \le 2N-1, \quad (7)$$

where $r_{yy}(m,k), r_{xx}(m,k)$ and $r_{ww}(m,k)$ are *two-sided autocorrelation sequences* of noisy speech, clean speech, and additive noise, respectively, and k is the autocorrelation sequence index within a frame. Because the additive noise $w(m,n)$ is assumed to be stationary, the frame index m of $r_{ww}(m,k)$ can be dropped out and (7) can be further expanded as

$$r_{yy}(m,k) = r_{xx}(m,k) \otimes h(k) \otimes h(-k) + r_{ww}(k) \otimes h(k) \otimes h(-k), \quad 0 \le m \le M-1, \quad 0 \le k \le 2N-1. \quad (8)$$

Since the two-sided autocorrelation sequences is symmetric, this $2N$-point $r_{yy}(m,k)$ is computed from N-point $y(m,n)$ by the following equation,

$$r;_{yy}(m,k) = \begin{cases} \sum_{j=0}^{N-1-k} y(m,j)y(m,j+k), & 0 \le k \le N-1 \\ 0, & k = N \\ r_{yy}(m,2N-k), & N+1 \le k \le 2N-1. \end{cases} \quad (9)$$

Applying the temporal filtering on both sides of (8), we obtain

$$\Delta r_{yy}(m,k) = \Delta r_{xx}(m,k) \otimes h(k) \otimes h(-k) \quad (10)$$

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

Note that we have removed the noise term $r_{ww}(k) \otimes h(k) \otimes h(-k)$ when we apply the temporal filtering.

Taking 2N-point DFT on both sides of (10) with respect to k, we obtain

$$S_{yy}^{\Delta}(m,f) = S_{xx}^{\Delta}(m,f) \cdot |H(f)|^2, \quad 0 \le f \le 2N-1 \quad (11)$$

where $S_{yy}^{\Delta}(m,f), S_{xx}^{\Delta}(m,f)$, and $H(f)$ denote the spectra of $\Delta r_{yy}(m,k), \Delta r_{xx}(m,k)$ and $h(k)$, respectively. Taking the logarithm of (11) yields

$$\log S_{yy}^{\Delta}(m,f) = \log S_{xx}^{\Delta}(m,f) + 2\log|H(f)|, \quad 0 \le f \le 2N-1. \quad (12)$$

The channel effect becomes an additive term in the logarithmic spectrum domain. Subtracting the mean of logarithmic spectrum in an utterance, we can remove the channel bias. Then taking exponential and $2N$-point inverse DFT with respect to f, we obtain

$$\overline{r_{yy}}(m,k) = \text{InverseDFT}\left\{\exp\left(\log S_{yy}^{\Delta}(m,f) - \frac{1}{M}\sum_{m=0}^{M-1}\log S_{yy}^{\Delta}(m,f)\right)\right\}, \quad 0 \le k \le 2N-1, \quad 0 \le f \le 2N-1, \quad 0 \le m \le M-1. \quad (13)$$

Note that the resulted $2N$-point $\overline{r_{yy}}(m,k)$ is a symmetrically two-sided sequence. The first N points of $\overline{r_{yy}}(m,k)$ are denoted by $\overline{r_{yy}^{+}}(m,k)$ and named the CHannel-Normalization Relative Autocorrelation Sequence (CHNRAS). We consider the CHNRAS an alternative representation of the original speech signal in time domain that is robust to channel and noise corruption.

IV. Properties of RAS-MFCC and CHNRAS-MFCC

Since RAS and CHNRAS are alternative time-domain representations of the original speech signal, we suggest that the features of RAS and CHNRAS are extracted in the same way as we extract features from the original speech signal. The mel-frequency cepstrum is a widely used feature in speech and speaker recognition. The mel-frequency cepstral coefficients (MFCC) of RAS and CHNRAS are denoted as RAS-MFCC and CHNRAS-MFCC, respectively. In order to know the properties of RAS-MFCC and CHNRAS-MFCC, we artificially add white noise to the clean speech and investigate the effect of additive noise to these features. Fig. 1 illustrates the norms of MFCC, RAS-MFCC, and CHNRAS-MFCC. It shows that the norms of RAS-MFCC and CHNRAS-MFCC shrink also. This implies that we can apply the projection measure technique to RAS-MFCC and CHNRAS-MFCC to adapt the reference model to the noisy condition. A preliminary study on combining RAS-MFCC with projection measure technique for speaker identification has been reported in [21].

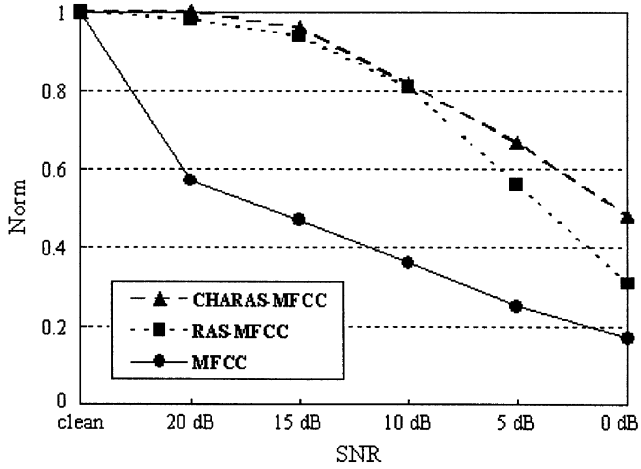

Fig. 1. Cepstral norm shrinkage of different MFCCs in white noise corruption.

The short-time modified coherence (SMC) [17] and one-sided autocorrelation LPC (OSALPC) [18] are other related robust features derived in autocorrelation domain. Both SMC and OSALPC are only robust to white noise, while RAS-MFCC is robust to any colored noise. A complete comparison among SMC, OSALPC, and RAS-MFCC can been found in [15].

V. Combination of RAS-MFCC and CHNRAS-MFCC With Projection Measure

For the traditional MFCC, projection measure technique has been proved to be effective for speech recognition [19], [20]. In this section, we perform a similar operation on RAS-MFCC and CHNRAS-MFCC to adapt the reference models by searching an optimal shrinking factor frame by frame. Each speaker model is expressed by a Gaussian mixture density that is a weighted sum of L component densities given by the equation

$$p(x_t \mid \lambda) = \sum_{i=1}^{L} w_i N(x_t, v_i, \Lambda_i) = \sum_{i=1}^{L} \frac{w_i}{(2\pi)^{D/2}|\Lambda_i|^{1/2}} \exp\left\{-\frac{1}{2}\|x_t - v_i\|_{\Lambda_i^{-1}}\right\} \quad (14)$$

where

$$\|x_t - v_i\|_{\Lambda_i^{-1}} \triangleq (x_t - v_i)^{\mathrm{T}} \Lambda_i^{-1} (x_t - v_i) \quad (15)$$

w_i, v_i, and Λ_i are weight, mean vector, and covariance matrix of the i-th Gaussian distribution, respectively. When the testing utterance x_t is corrupted by the noise, a modified scoring density based on the projection measure (PM) method is

$$P^{\mathrm{PM}}(x_t \mid \lambda) \triangleq \sum_{i=1}^{L} \frac{w_i}{(2\pi)^{D/2}|\Lambda_i|^{1/2}} \exp\left\{-\frac{1}{2}\|x_t - v_i\|^{\mathrm{PM}}_{\Lambda_i^{-1}}\right\} \tag{16}$$

where

$$\|x_t - v_i\|^{\mathrm{PM}}_{\Lambda_i^{-1}} \triangleq \min_{\alpha}\|x_t - \alpha v_i\|_{\Lambda_i^{-1}}. \tag{17}$$

Here, α is a shrinking factor that minimizes the distance between a testing frame and a speaker model. The task of speaker identification is to find a speaker model which gives the minimum overall distance.

VI. Experiments

The database for this study is a 100-speaker Mandarin digits database provided by Chung Hwa Telecommunication Laboratories, Taiwan, R.O.C. [15]. The speech data were collected from 50 males and 50 females in five recording sessions. Each speaker in each recording session randomly selected one of 340 prompting tables. According to each table, the speaker uttered 40 utterances. These 40 utterances contained seven sets of spoken materials. The first set included ten isolated digits in random order. The remaining sets were 2-digit through 7-digit strings with each set containing five utterances. Five sessions were recorded over about half-month period. The speech was recorded via high quality microphones in quiet environments. Since the effects of background noise and channel distortion are minimized, the speech in this database is referred to as the clean speech.

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. Thus a feature vector can be composed of 14 static components or 14 static plus 14 dynamic components. Both two cases will be evaluated. In computing RAS-MFCC and CHNRAS-MFCC, the one-side autocorrelation sequence and two-side autocorrelation sequence are used, respectively.

A. Speaker Models Based on Clean Enrollment

The speaker models are trained from the clean speech. Several experiments are conducted on clean speech testing to evaluate the effect of the training data size, the utterance length, and the combination of static and dynamic features. Following the clean speech testing, we evaluate the robust property with respect to additive noise and channel distortion.

The first three sessions of the clean speech database are used for training the reference model of each speaker. A recording session is 40 utterances, or about 30 s of speech data size. Totally the training data are about 90 s for each speaker.

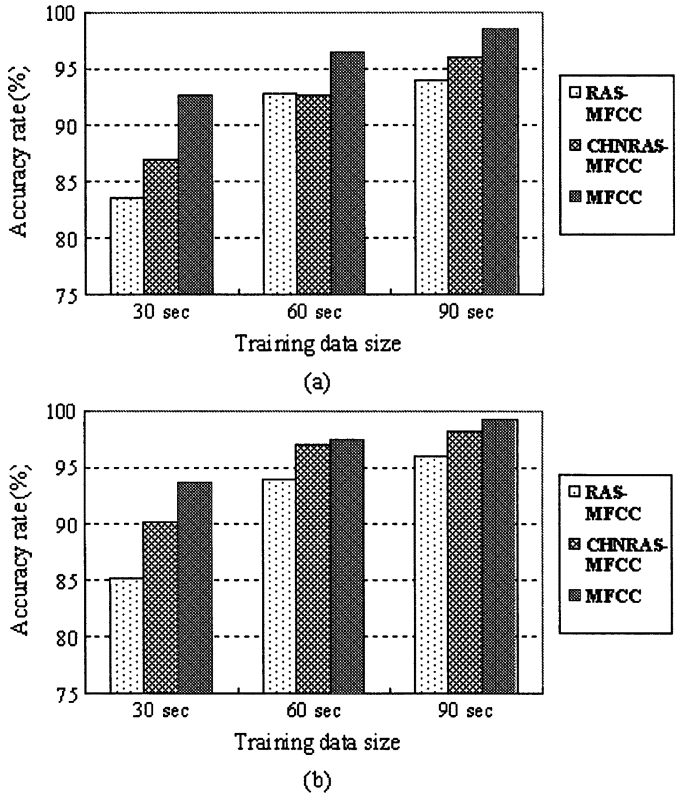

Fig. 2. Accuracy rates versus training data size for RAS-MFCC, CHNRAS-MFCC, and MFCC: (a) using static features only and (b) using static and dynamic features. (The speaker model is a Gaussian mixture with 32 component densities.).

TABLE I
Accuracy Rate (%) Versus Utterance Length for Clean Testing Speech (Clean Enrollment)

	Static features		Static+Dynamic features	
Utterance Length	7-digit	14-digit	7-digit	14-digit
MFCC	98.3	99.4	98.7	99.4
RAS-MFCC	95.8	98	96.2	98
CHNRAS-MFCC	96.8	99	98.2	99.8

TABLE II
Accuracy Rate (%) Versus Utterance Length for Testing Speech Corrupted by Channel Distortion (Clean Enrollment)

	Static features		Static+Dynamic features	
Utterance Length	7-digit	14-digit	7-digit	14-digit
MFCC	36.9	37.4	61.4	65.4
MFCC (CMN)	95	98.4	96.8	99
CHNRAS-MFCC	86.2	91.2	93.1	97.8

1) Clean Speech Testing for Different Training Data Sizes: The 7-digit utterance, which is about 2 s in length, extracted from the fourth session in the speech database is used

TABLE III
ACCURACY RATE (%) FOR TESTING SPEECH CORRUPTED BY ADDITIVE NOISE: (CLEAN ENROLLMENT): (a) *WHITE* NOISE; (b) *FACTORY* NOISE; (c) *F16* NOISE; AND (d) *BABBLE* NOISE

(a)

SNR (dB)	clean	20	15	10	5	0
MFCC	99.4	68.8	34.4	13.2	6.6	3.4
MFCC (PM)	99.6	76.8	47.8	26.4	15.6	9
RAS-MFCC	98	92.8	87	59.2	29.2	8.4
RAS-MFCC (PM)	98	95.4	95	86.4	58	25
CHNRAS-MFCC	99.8	92.8	68	27.8	7.2	3
CHNRAS-MFCC (PM)	99.6	94	75.2	44.8	17	7.4

(b)

SNR (dB)	clean	20	15	10	5	0
MFCC	99.4	91	60	28.8	7.2	4.2
MFCC (PM)	99.6	94.6	75.2	39.2	13.4	4
RAS-MFCC	98	93	85.6	59.2	24.8	3.8
RAS-MFCC (PM)	98	94.4	88	74.2	43	12
CHNRAS-MFCC	99.8	99.4	96.6	81.4	40.4	8.6
CHNRAS-MFCC (PM)	99.6	99.4	98.6	93.6	77.4	38.4

(c)

SNR (dB)	clean	20	15	10	5	0
MFCC	99.4	79.4	47.6	20.4	5.4	1.4
MFCC (PM)	99.6	89	60.2	26.6	7.6	1.8
RAS-MFCC	98	90	76.6	38.4	12.4	2.6
RAS-MFCC (PM)	98	93.4	85.6	54.2	22.2	5.6
CHNRAS-MFCC	99.8	98.6	96.4	79.8	42.2	8
CHNRAS-MFCC (PM)	99.6	99.4	98	91.6	71.2	36.8

(d)

SNR (dB)	clean	20	15	10	5	0
MFCC	99.4	97	82	48.8	18.6	5
MFCC (PM)	99.6	97.6	91.2	61.4	23.4	6
RAS-MFCC	98	95	91.8	78.6	46.2	19.4
RAS-MFCC (PM)	98	95.6	93.8	87.6	64	24.8
CHNRAS-MFCC	99.8	99	98	88.4	57.6	15.6
CHNRAS-MFCC (PM)	99.6	99.4	98.6	90.2	66.8	29.2

as the testing utterance. Totally there are 500 testing utterances. In this experiment, the speaker models are trained using 30, 60, and 90 s of speech data for each speaker in order to evaluate the effect of training data size. Each model is a Gaussian mixture of 32 component densities. The RAS-MFCC, CHNRAS-MFCC, and MFCC are the speech features for this evaluation. Fig. 2 shows the performance of speaker identification. Part (a) shows the result of using static features only, and part (b) shows the result of using the combination of static and dynamic features. It is obvious that the performance is improved as the training data increases. MFCC is better than RAS-MFCC and CHNRAS-MFCC in all cases. Comparing with RAS-MFCC and CHNRAS-MFCC, MFCC is more effective in clean speech testing when the data size is small. Part (b) shows that the inclusion of dynamic features is always better than that of using static feature only. When the training data size is 90 s, all the features can reach the high performance of above 95% accuracy.

2) Clean Speech Testing for Different Testing Utterance Lengths: The purpose of this experiment is to evaluate the effect of testing utterance length. The training data size is 90-second. The speaker model is a Gaussian mixture of 64 component densities. The 7-digit utterances in fourth and fifth sessions are used for testing. Two types of utterance lengths, 7-digit, and 14-digit, are used in this testing. The 14-digit utterance is a concatenation of two 7-digit utterances, so that the length is about 4 s. Totally, there are 1000 7-digit utterances and 500 14-digit utterances for the testing. Both feature vectors of 14 static components and 14 static plus 14 dynamic components are evaluated. The performance is shown in Table I. It shows that the accuracy rates of using long testing utterances, i.e., 14-digit utterances, for all features are comparable.

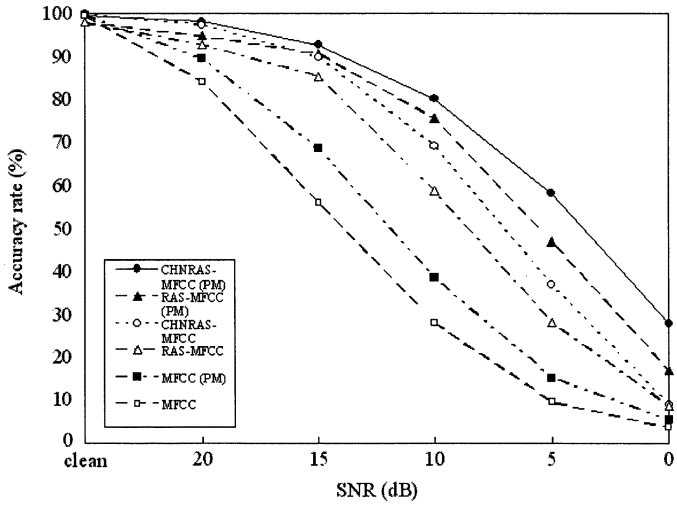

Fig. 3. Accuracy rate (%) for testing speech corrupted by additive noise (clean enrollment).

The performance of RAS-MFCC is a little bit less than that of MFCC and CHNRAS-MFCC. The inclusion of dynamic components yields some improvement.

3) Noisy Speech Testing—Corrupted by Channel Distortion: The testing speech is polluted by the channel distortion. The purpose of this experiment is to evaluate the performance of CHNRAS-MFCC. For comparison, the MFCC with cepstral mean normalization (CMN) [22] is included. The channel effect is simulated by convoluting the clean speech with a channel filter that is randomly selected from 41 channel filters of 49-order [23]. The result is shown in Table II. Since each speaker model is trained using the clean speech, the

TABLE IV
ACCURACY RATE (%) FOR TESTING ON SPEECH CORRUPTED BY ADDITIVE NOISE AND CHANNEL DISTORTION: (CLEAN ENROLLMENT): (a) *WHITE* NOISE AND CHANNEL DISTORTION; (b) *FACTORY* NOISE AND CHANNEL DISTORTION; (c) *F16* NOISE AND CHANNEL DISTORTION; AND (d) *BABBLE* NOISE AND CHANNEL DISTORTION

(a)

SNR (dB)	clean	20	15	10	5	0
MFCC	65.4	10.4	4.2	3	1	1.2
MFCC (PM)	77	24.8	12.6	6.8	4.2	1.8
MFCC (CMN)	99	75	40.4	15.2	7.2	3.6
MFCC (PM+CMN)	99.2	79.6	50.2	21.8	11	5.6
CHNRAS-MFCC	97.8	92.6	70.2	29.2	7	2.4
CHNRAS-MFCC (PM)	98.4	93.4	75.2	44.8	17.8	7.6

(b)

SNR (dB)	clean	20	15	10	5	0
MFCC	65.4	34	22.4	9.6	3	1.6
MFCC (PM)	77	54.4	36.8	17.8	3.2	1.4
MFCC (CMN)	99	93.8	77.6	42.8	15.4	4
MFCC (PM+CMN)	99.2	96.6	92	69.8	29.2	7
CHNRAS-MFCC	97.8	99.4	97.4	85.4	42.4	8.6
CHNRAS-MFCC (PM)	98.4	98.8	98	91.2	72.4	39.4

(c)

SNR (dB)	clean	20	15	10	5	0
MFCC	65.4	29.6	17.6	7.6	2.4	1.4
MFCC (PM)	77	46.8	30.8	17.2	4.8	1.2
MFCC (CMN)	99	93.2	79.2	45.2	14.4	2.8
MFCC (PM+CMN)	99.2	94	85.2	58.2	22.4	4.4
CHNRAS-MFCC	97.8	99.2	97.8	84.8	46.8	10
CHNRAS-MFCC (PM)	98.4	98.8	97.2	90.2	69.6	38

(d)

SNR (dB)	clean	20	15	10	5	0
MFCC	65.4	39.2	27.6	14.6	7.2	3.8
MFCC (PM)	77	51.8	40	24.2	9.8	2.8
MFCC (CMN)	99	93.6	83.8	51.2	20.6	5.4
MFCC (PM+CMN)	99.2	97	92.4	75.6	42.4	10.6
CHNRAS-MFCC	97.8	96.8	94.8	83.4	53	13.2
CHNRAS-MFCC (PM)	98.4	96.2	92.8	81.6	58	24.8

performance of MFCC degrades significantly if no channel compensation is made to the testing utterance. When the MFCC is compensated by CMN and augmented with its dynamic term, the MFCC achieve the best recognition accuracy. The performance is very close to the case of clean speech testing. When the dynamic features are included, the performance of CHNRAS-MFCC is less than that of MFCC with CMN by 1.2% in long testing utterances and 3.7% in short testing utterances. This experiment shows that the cepstral mean normalization of MFCC can overcome the channel distortion effectively. The CHNRAS-MFCC needs dynamic components and long testing utterances to obtain comparable performance.

4) Noisy Speech Testing—Corrupted by Additive Noise: The testing speech is polluted by the additive noise. Both RAS-MFCC and CHNRAS-MFCC are evaluated and compared with the traditional MFCC compensated by projection measure (PM). This experiment uses long test utterances, i.e., 500 14-digit utterances. The dynamic features are included. The testing utterances are generated by adding the artificial noises in five SNR levels. The white noise is generated by using a random number generation program, and other colored noises, i.e., factory noise, F16 noise, and babble noise, are extracted from NOISEX-92 database [24]. The techniques of combining the projection measure with RAS-MFCC and CHNRAS-MFCC are evaluated. The result is summarized in Table III(a)–(d). For the case of white noise corruption, i.e., in Table III(a), the performance of MFCC degrades most significantly among all features. Although the MFCC with projection measure can make some improvement, its performance is still worse than RAS-MFCC and CHNRAS-MFCC. It is obvious that RAS-MFCC and CHNRAS-MFCC are quite robust to the additive noises. Their performance can be further improved by combining with the projection measure. The RAS-MFCC with the projection measure achieves better performance than the CHNRAS-MFCC with the projection measure only in the case of corrupting by white noise.

In Table III(b), (c),, and (d), the testing speech was corrupted by factory, F16, and babble noises, respectively. Fig. 3 plots the average accuracy rates shown in Table III. The performance of MFCC degrades significantly. The best performance comes from CHNRAS-MFCC combined with projection measure. Generally, the CHNRAS-MFCC (PM) is better than RAS-MFCC (PM). This is due to a mean normalization in frequency-domain during the derivation of CHNRAS-MFCC. The normalization in frequency domain also provides the compensation to additive noise when the testing speech is corrupted by colored noise. This is similar to the application of CMN method to compensate the additive noise corruption [25].

The other finding in Table III is that the assumption of corrupting by stationary noise during the derivation of RAS-MFCC and CHNRAS-MFCC does not hurt too much to the performance in our experiments. The babble noise is usually the case of nonstationary noise. Table III(d) shows that RAS-MFCC and CHNRAS-MFCC can still obtain relatively high accuracy rates for babble noise corruption even in low SNR.

5) Noisy Speech Testing—Corrupted by Additive Noise and Channel Distortion: The testing speech is polluted by additive noise and channel distortion simultaneously. 500 14-digit testing utterances are used for this evaluation. These testing speech data are artificially corrupted by the noises, and then dis-

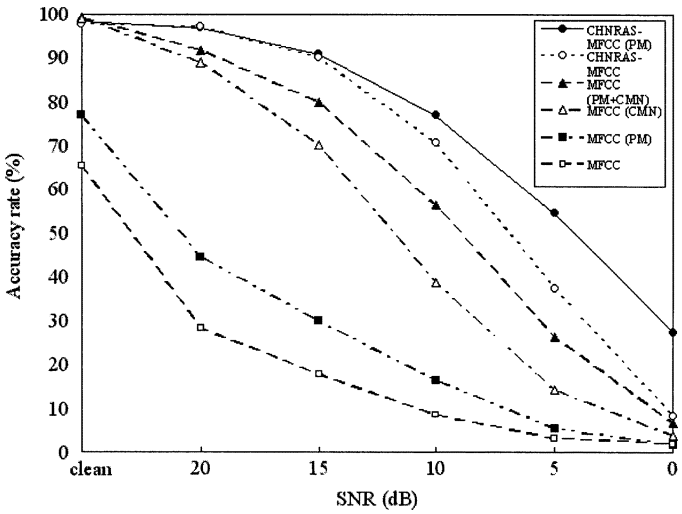

Fig. 4. Accuracy rate (%) for testing speech corrupted by additive noise and channel distortion (clean enrollment).

torted by the channel effect. For comparison, the cases of traditional MFCC compensated with CMN and projection measure are included. Table IV(a)–(d) list all the results. Fig. 4 plots the average accuracy rates shown in Table IV. For clean testing utterances, the performance of MFCC with mean normalization and projection measure is the best. But when the noise level increases, CHNRAS-MFCC becomes better. The use of projection measure with CHNRAS-MFCC gives further improvement especially in cases of low SNR.

B. Speaker Models Based on Noisy Enrollment

The previous experiments have shown the effectiveness of CHNRAS-MFCC and RAS-MFCC based on clean enrollment. In the following experiments we want to see whether these features are still good in case of noisy enrollment.

The 120 training utterances of each speaker in the first three sessions are split into six subsets. The six subsets represent a clean condition and five noise scenarios. The noisy environment is expressed in different SNR levels, i.e., 20 dB, 15 dB, 10 dB, 5 dB, and 0 dB. Therefore, a total of 120 multi-condition utterances are used for generating a speaker model which is a Gaussian mixture of 64 component densities. Both of static and dynamic features are used. With this method, we produce two sets of speaker models involving respectively the babble noise and the white noise. For each enrollment, four testing conditions are specified, i.e., corrupted by inside noise, corrupted by outside noise, corrupted by inside noise and channel distortion, and corrupted by outside noise and channel distortion. The corrupting noises are white noise, factory noise, F16 noise, and babble noise. A corrupting noise is called the inside noise if it is the same type as the noise in the enrollment speech. The corrupting noise of the type different from the noise in the enrollment speech is called the outside noise. For example, the babble noise is considered the inside noise when the speaker model is trained by speech corrupted by babble noise. On the other hand, F16 noise, factory noise, and white noise are considered the outside noises for the speaker model trained by speech corrupted by babble noise. The

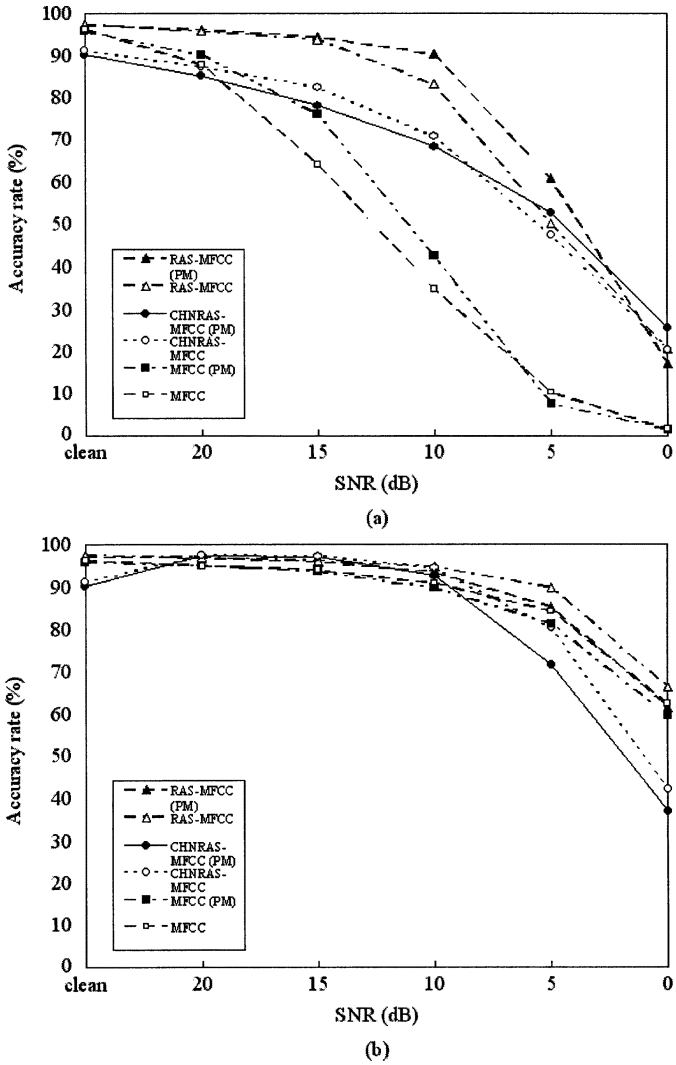

Fig. 5. Noisy speech testing: Testing speech corrupted by inside noise: (a) enrollment speech corrupted by babble noise and (b) enrollment speech corrupted by white noise.

channel distortion is simulated by convoluting the speech with a randomly selected channel filter. The "clean" means no additive noise. The experiments are conducted for cases of enrollment speech corrupted by babble noise and by white noise.

1) Noisy Speech Testing—Corrupted by Inside Noise: Fig. 5 shows the accuracy rates with the inside noise at various SNR levels. For the enrollment speech corrupted by babble noise [Fig. 5(a)], the accuracy rates deteriorate in decreasing SNR. With the clean testing speech, RAS-MFCC and MFCC can achieve the comparable performance. As the SNR decreases, the MFCC degrades rapidly, and the RAS-MFCC can maintain relatively good performance. The CHNRAS-MFCC is better than the MFCC, but worse than the RAS-MFCC. For the enrollment speech corrupted by white noise [Fig. 5(b)], the performances of three features, i.e., MFCC, RAS-MFCC, and CHNRAS-MFCC, are comparable. When the SNR is below 5 dB, the RAS-MFCC gives the best accuracy. In general, the RAS-MFCC performs well in the matched noise environment. The projection method offers only a little improvement for these three features.

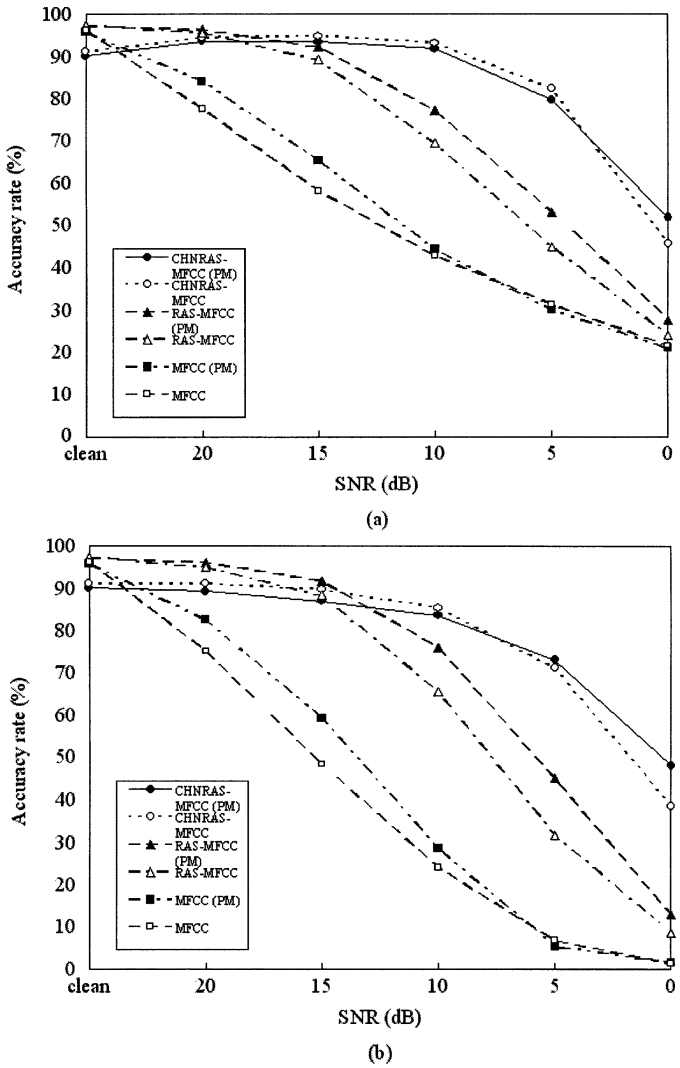

Fig. 6. Noisy speech testing: Testing speech corrupted by outside noises: (a) enrollment speech corrupted by babble noise and (b) enrollment speech corrupted by white noise.

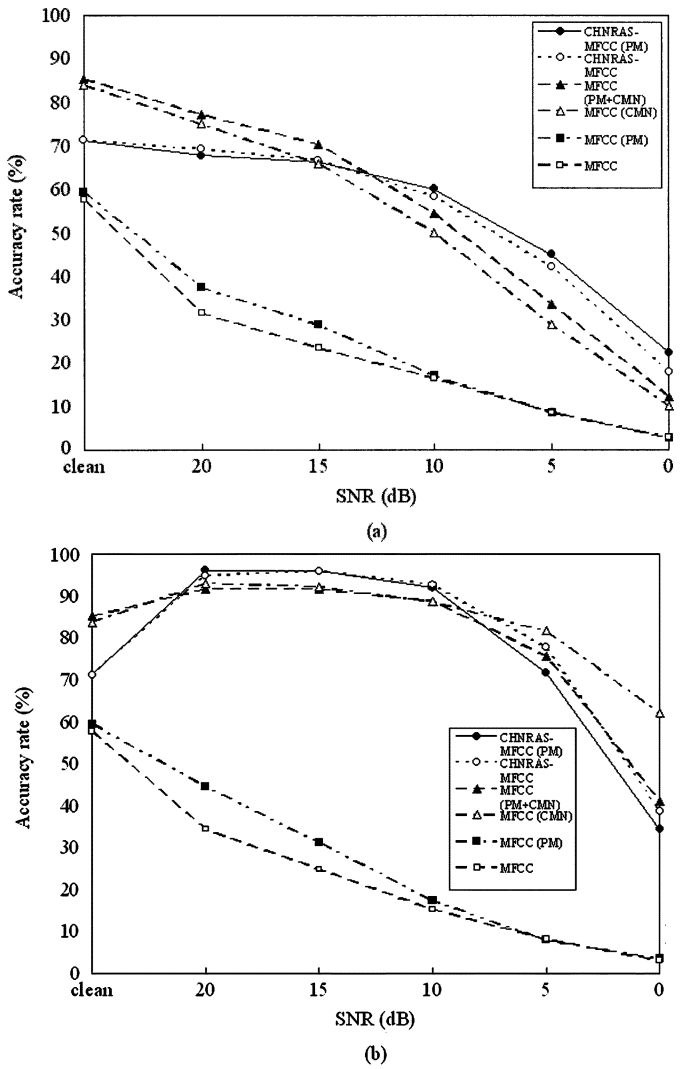

Fig. 7. Noisy speech testing: Testing speech corrupted by inside noise and channel distortion: (a) enrollment speech corrupted by babble noise and (b) enrollment speech corrupted by white noise.

2) Noisy Speech Testing—Corrupted by Outside Noises: Fig. 6 shows the accuracy rates with outside noises at various SNR levels. Both figures show a consistent performance. The CHNRAS-MFCC again is worse than RAS-MFCC and MFCC on the clean testing speech. However, the CHNRAS-MFCC performs better than the RAS-MFCC when the SNR becomes lower. The projection method offers only a little improvement for RAS-MFCC and MFCC, and almost no improvement for CHNRAS-MFCC. The reason is that the training data are corrupted by noise also so that the norm shrinkage is not obvious. The projection measure does not show any advantage in this environmental mismatch condition. This experiment also shows that the procedure of mean normalization in frequency-domain during the derivation of CHNRAS-MFCC provides the compensation to mismatched noises.

3) Noisy Speech Testing—Corrupted by Inside Noise and Channel Distortion: Fig. 7 shows the accuracy rates with the inside noise and channel distortion. The RAS-MFCC is not considered in this experiment since it is not expected to be able to compensate the channel effect. We compare the performance of CHNRAS-MFCC and MFCC with CMN. The projection measure provides only a little bit of improvement on CHNRAS-MFCC and MFCC (CMN). For the speech corrupted only by channel distortion, i.e., the cases denoted by "clean", the CHNRAS-MFCC is worse than the MFCC with CMN. However, for the enrollment speech corrupted by babble noise (Fig. 7(a)), the CHNRAS-MFCC is better than the MFCC with CMN when the SNR is below 10 dB. For the enrollment speech corrupted by white noise [Fig. 7(b)], the CHNRAS-MFCC is better than the MFCC with CMN at the higher SNR.

4) Noisy Speech Testing—Corrupted by Outside Noises and Channel Distortion: Fig. 8 shows the accuracy rates with outside noises and channel distortion. Both figures show a consistent performance. With only channel distortion, the CHNRAS-MFCC is worse than the MFCC with CMN. However, the CHNRAS-MFCC becomes better when the SNR decreases. The projection method again provides very few of improvement on CHNRAS-MFCC. The results tell that the CHNRAS-MFCC is more robust in lower SNR cases.

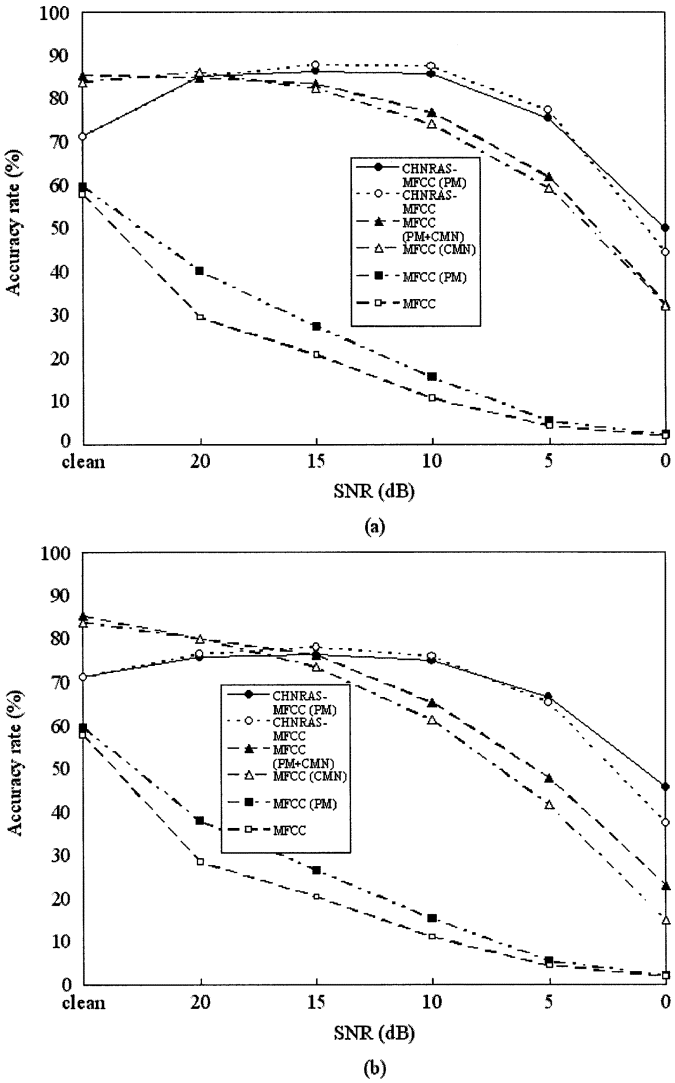

Fig. 8. Noisy speech testing: Testing speech corrupted by outside noises and channel distortion: (a) enrollment speech corrupted by babble noise and (b) enrollment speech corrupted by white noise.

VII. Conclusion

In this paper, two robust features, RAS-MFCC and CHNRAS-MFCC, are introduced for speaker identification. The RAS-MFCC is robust to additive noise. The CHNRAS-MFCC is robust to both additive noise and channel distortion. For the case of clean enrollment, the CHNRAS-MFCC in combination with the projection measure technique gives the best performance. For the case of noisy enrollment, the projection measure technique contributes only small improvement. Several types of noise corruption to the testing utterances are evaluated. Experimental results show that the proposed robust feature, CHNRAS-MFCC, is effective for overcoming the corruption of noise and channel distortion in cases of lower SNR environment. The method is effective for colored noise corruption also. The weakness of this proposed method is that the derivation of RAS and CHNRAS is based on the assumption of corrupting by stationary noise. This may limit the application of RAS and CHNRAS to a more diverse environment. However, this assumption does not hurt too much to the performance in our experiments.

References

[1] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Commun.*, vol. 16, pp. 261–291, Apr. 1995.

[2] M. J. F. Gales and S. J. Young, "Robust speech recognition in additive and convolutional noise using parallel model combination," *Comput. Speech Lang.*, vol. 9, pp. 289–307, 1995.

[3] A. Sankar and C.-H. Lee, "A maximum-likelihood approach to stochastic matching for robust speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 4, pp. 190–202, May 1996.

[4] O. Siohan and C. H. Lee, "Iterative noise and channel estimation under the stochastic matching algorithm framework," *IEEE Signal Process. Lett.*, vol. 4, pp. 304–306, Nov. 1997.

[5] Y. Zhao, "Channel identification and signal spectrum estimation for robust automatic speech recognition," *IEEE Signal Process. Lett.*, vol. 5, no. 12, pp. 305–308, Dec. 1998.

[6] M. Afify, Y. Gong, and J. P. Haton, "A general joint additive and convolutive bias compensation approach applied to noisy Lombard speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 6, pp. 524–538, Nov. 1998.

[7] H. A. Murthy, F. Beaufays, L. P. Heck, and M. Weintraub, "Robust text-independent speaker identification over telephone channels," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 5, pp. 554–568, Sep. 1999.

[8] L. P. Heck, Y. Konig, M. K. Sönmez, and M. Weintraub, "Robustness to telephone handset distortion in speaker recognition by discriminative feature design," *Speech Commun.*, vol. 31, pp. 181–192, 2000.

[9] K. Sönmeza, L. Heck, and M. Weintraub, "Multiple speaker tracking and detection: Handset normalization and duration scoring," *Digital Signal Process.*, vol. 10, no. 1–3, pp. 133–142, Jan. 2000.

[10] D. A. Reynolds, "The effects of handset variability on speaker recognition performance experiments on the switchboard corpus," in *Proc. ICASSP-96*, vol. 1, 1996, pp. 113–116.

[11] T. F. Quatieri, D. A. Reynolds, and G. C. O'Leary, "Estimation of handset nonlinearity with application to speaker recognition," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 5, pp. 567–584, Sep. 2000.

[12] H. G. Hirsch, P. Meyer, and H. Ruchl, "Improved speech recognition using high-pass filtering of subband envelopes," in *Proc. EUROSPEECH'91*, Genova, Italy, 1991, pp. 413–416.

[13] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[14] C. Avendano and H. Hermansky, "Study on the dereverberation of speech based on temporal envelope filtering," in *Proc. ICSLP'96*, vol. 2, Philadelphia, PA, 1996, pp. 889–892.

[15] K. H. Yuo and H. C. Wang, "Robust features for noisy speech recognition based on temporal trajectory filtering of short-time autocorrelation sequences," *Speech Commun.*, vol. 28, pp. 13–24, 1999.

[16] ——, "Robust features derived from temporal trajectory filtering for speech recognition under corruption of additive and convolutional noises," in *Proc. ICASSP'98*, Seattle, WA, 1998, pp. 577–580.

[17] D. Mansour and B. H. Juang, "The short-time modified coherence representation and noisy speech recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 6, pp. 795–804, 1989.

[18] J. Hernando and C. Nadeu, "Linear prediction of the one-sided autocorrelation sequence for noisy speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 1, pp. 80–84, 1997.

[19] D. Mansour and B. H. Juang, "A family of distortion measures based upon projection operation for robust speech recognition," *IEEE Trans. Acoust., Speech., Signal Process.*, vol. 37, no. 11, pp. 1659–1671, Nov. 1989.

[20] A. Carlson and M. A. Clements, "A projection-based likelihood measure for speech recognition in noise," *IEEE Trans. Speech Audio Process.*, pt. 1, vol. 2, no. 1, pp. 97–102, Jan. 1994.

[21] K. H. Yuo, T. H. Hwang, and H. C. Wang, "Combination of temporal trajectory filtering and projection measure for robust speaker identification," in *Proc. ICSLP2000*, Beijing, China, 2000.

[22] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-29, no. 2, pp. 254–272, Apr. 1981.

[23] J. T. Chien, L. M. Lee, and H. C. Wang, "Channel-effect-cancellation method for speech recognition over telephone system," in *Proc. Inst. Elect. Eng.—Vis., Image, Signal Process.*, vol. 142, 1995, pp. 395–399.

[24] A. Varga and H. J. M. Steeneken, "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 Commun.*, vol. 12, pp. 247–251, 1993.

[25] K. Parssinen, P. Salmela, M. Harju, and I. Kiss, "Comparing jacobian adaptation with cepstral mean normalization and parallel model combination for noise robust speech recognition," in *Proc. ICASSP 2002*, pp. 193–196.

Kuo-Hwei Yuo received the B.S. degree in electronic engineering from Chung Yuan Christian University (CYCU), Chung-Li, Taiwan, R.O.C., in 1983, and the M.S. and Ph.D. degrees in electrical engineering from National Tsing Hua University (NTHU), Hsinchu, Taiwan, in 1988 and 2000, respectively.

He joined the Chung-Shan Institute of Science and Technology (CSIST) in 1983, and has involved in the projects of switching power supply, automatic test system, system on chip, and rapid thermal control system. He is an adjunct Assistant Professor with the CYCU for courses of speech signal processing and digital signal processing since 2000. His current interests include robust speech/speaker recognition, digital signal processing, statistical signal processing, and communication engineering.

Tai-Hwei Hwang received the B.S. degree in electronic engineering from Chung Yuan Christian University, Chung-Li, Taiwan, R.O.C., in 1987 and the M.S. and Ph.D. degrees in electrical engineering from National Tsing Hua University, Hsinchu, Taiwan, in 1992 and 2000, respectively.

He has been with the Computer and Communication Laboratories, Industrial Technology Research Institute (ITRI), Taiwan, since 1999, and is involved in the projects of speech/speaker recognition. His current interests include robust speech/speaker recognition, and digital signal processing.

Hsiao-Chuan Wang (M'76–SM'88) received the B.S. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, R.O.C., in 1969, and the M.S. and Ph.D. degrees in electrical engineering from the University of Kansas, Lawrence, in 1973 and 1977, respectively.

He joined the Department of Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan, in 1977. He was the Chair of Department of Electrical Engineering (August 1986–July 1992). He has served on the editorial board of the *Journal of CIEE* since 1993 and was the Editor-in-Chief (1993–1995) and then the Chair of editorial board (1996–1999). His current research interests include speech recognition, speech coding, speech enhancement, and digital signal processing.

Dr. Wang has served as Chair of the IEEE Taipei Section (1997–1999) and Associate Editor of IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (March 1999–February 2002). He is a member of Association of Computational Linguistics and Chinese Language Processing, and has been the President of the Association (December 1999–December 2001). He is a member of Chinese Institute of Electrical Engineering (CIEE).