Noise-Robust Speaker Recognition Using Subband Likelihoods and Reliable-Feature Selection

Sungtak Kim, Mikyong Ji, and Hoirin Kim

We consider the feature recombination technique in a multiband approach to speaker identification and verification. To overcome the ineffectiveness of conventional feature recombination in broadband noisy environments, we propose a new subband feature recombination which uses subband likelihoods and a subband reliable-feature selection technique with an adaptive noise model. In the decision step of speaker recognition, a few very low unreliable feature likelihood scores can cause a speaker recognition system to make an incorrect decision. To overcome this problem, reliable-feature selection adjusts the likelihood scores of an unreliable feature by comparison with those of an adaptive noise model, which is estimated by the maximum a posteriori adaptation technique using noise features directly obtained from noisy test speech. To evaluate the effectiveness of the proposed methods in noisy environments, we use the TIMIT database and the NTIMIT database, which is the corresponding telephone version of TIMIT database. The proposed subband feature recombination with subband reliable-feature selection achieves better performance than the conventional feature recombination system with reliable-feature selection.

Keywords: Speaker recognition, Gaussian mixture model, universal background model, feature recombination, mel-frequency cepstral coefficient, subband likelihood, reliable feature selection, adaptive noise model.

Manuscript received Apr. 11, 2007; revised Nov. 23, 2007.

This work is a result of the URC project sponsored by the MIC, Rep. of Korea.

Sungtak Kim (phone: +82 42 866 6221, email: stkim@icu.ac.kr), Mikyong Ji (email: lindaji@icu.ac.kr), and Hoirin Kim (email: hrkim@icu.ac.kr) are with the School of Engineering, Information and Communications University, Daejeon, Rep. of Korea.

I. Introduction

Speaker recognition, which can be classified into identification and verification, is a process of automatically recognizing who is speaking on the basis of individual information included in speech signal. Speaker identification finds the correct speaker of a given test utterance among registered speakers, and speaker verification determines whether the claimed speaker is accepted based on the score of the test utterance. In recent years, methods based on Gaussian mixture models (GMMs) [1] and the GMM universal background model (UBM) [2] have been dominant for text-independent speaker identification and verification. Most speaker recognition systems based on these methods provide very good performance under laboratory conditions. However, in real situations, such as hand-free car applications, the presence of interfering noises can dramatically lower the accuracy of speaker recognition systems. This performance degradation is mainly caused by mismatch between enrollment and recognition conditions. To overcome this mismatch problem, a number of techniques have been proposed. These techniques can be categorized into three classes: the feature domain approach [3], [4], in which the noisy test speech is modified or enhanced to move toward clean speech as closely as possible; the model domain approach [5], [6], in which the speaker models are modified or adapted to match the statistical properties of noisy test speech; and the score domain approach [7], [8], in which the scores of speaker models are adjusted to minimize the effect introduced by environment variability. In this paper, we focus on the multiband approach and the reliable-feature selection included in the feature domain approach and the score domain approach, respectively.

Widely used feature parameters, namely, the mel-frequency cepstral coefficients (MFCCs), are obtained by using the filter-bank approach, in which filters have equal bandwidths in the mel-scale frequency domain. The commonly used feature extraction is computed over the full band of the spectral representation of speech. A major drawback of the full-band-based computation is that even partial band-limited noise corruption affects all the feature vector components. The multiband approach deals with this problem by performing acoustic feature analysis independently on a set of frequency subbands. Since the resulting coefficients are computed independently, a band-limited noise does not spread over all of the feature components. In previous works on the multiband approach, likelihood recombination and feature recombination techniques were employed. Feature recombination yields better performance than likelihood recombination [9]-[12] because it enables the modeling of the correlation between subband feature vectors and better class discrimination.

Feature recombination tends to be more noise-robust than the full-band approach in the band-limited noise condition. However, in the case of the broadband noise condition, the improvement in the performance of feature recombination is not notable compared with that of the full-band approach. Even when the speech is corrupted by broadband noise, the individual subbands may be corrupted to differing degrees. Therefore, it is still effective to process each subband independently. However, in the conventional feature recombination technique, the likelihood scores are computed by using all subband features as shown in Fig. 1(c). This likelihood computation is not effective, even if the subband features are extracted separately. To cope with this drawback, we introduce a re-formulation of the subband likelihood computation and propose a new feature recombination using the subband likelihood computation.

In speaker identification under noise conditions, some input speech frames may have very low likelihood scores for the correct speaker, and that can cause the correct speaker not to be selected as the correct speaker. In speaker verification, this problem occurs frequently. If the effects of these low likelihood scores can be removed or reduced, the performance degradation could be decreased. Therefore, we propose a reliable-feature selection method based on an adaptive noise model corresponding to a score domain approach. Additionally, we apply the reliable-feature selection method to the subband level. To determine whether a feature is reliable, we compare the likelihoods of features for a speaker model or UBM with those for an adaptive noise model. If the features are determined to be unreliable, the likelihoods of the unreliable features are substituted by those for an adaptive noise model to minimize the effect of low likelihood scores of unreliable features.

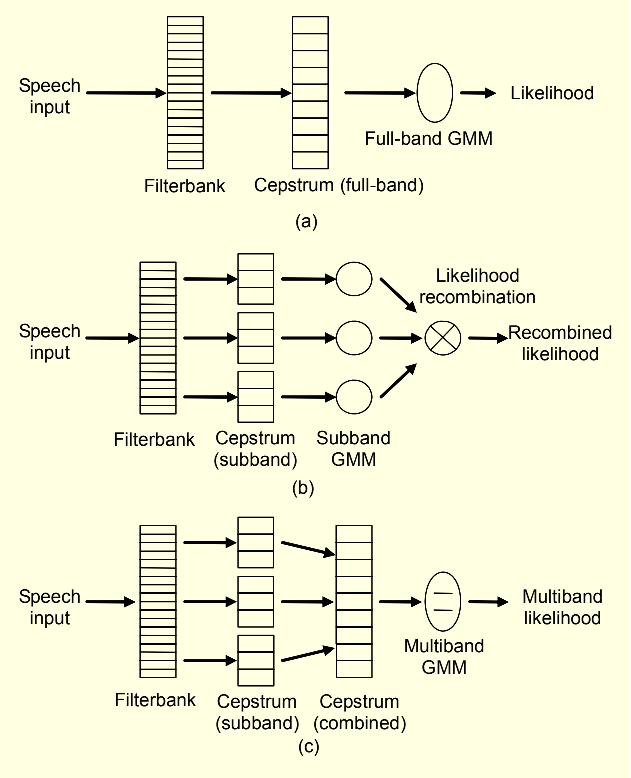


Fig. 1. Block diagrams of (a) full-band speaker recognition, (b) multiband speaker recognition by likelihood recombination, and (c) multiband speaker recognition by feature recombination.

In section II, a method to compute multiband MFCCs [13] is briefly reviewed and the proposed subband likelihood computation is presented. Conventional speaker identification and verification methods are briefly described in section III. In section IV, the proposed subband feature recombination with reliable-feature selection for speaker recognition is explained. Finally in sections V and VI, the experimental results of the proposed methods and conclusions are given.

II. Subband Likelihood Scoring in Multiband MFCCs

If there is an M-subband system with a total of N channels and L MFCCs per subband, the j-th multiband MFCC of the i-th subband of a frame is

$$x_j^{(i)} = \sqrt{\frac{2}{N/M}} \sum_{n=1}^{N/M} LFB_n^{(i)} \cos\left[(n-0.5)\frac{j\pi}{N/M}\right], \quad 1 \le j \le L \le \frac{N}{M}, \tag{1}$$

where $LFB_n^{(i)}$ is the logarithm of the n-th channel energy of the i-th subband. The example of the feature extraction process for a two-subband system is depicted in Fig 2. By using these multiband MFCCs, which are combined feature vectors,

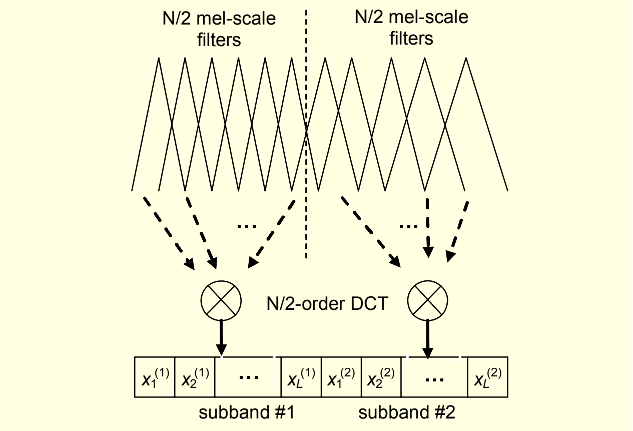


Fig. 2. Example of the feature extraction for 2 subbands.

speaker models and UBM are estimated. To compute subband likelihoods, we need an independent assumption and marginalization process. In the M-subband feature recombination system, the combined feature vector $\mathbf{X}$ is partitioned into subband parts, that is, $\mathbf{X}=(\mathbf{x}^{(1)},\mathbf{x}^{(2)},\cdots,\mathbf{x}^{(M)})$. On the assumption that the feature of each subband is statistically independent, we can obtain

$$\begin{aligned} p(\mathbf{X}\,|\,\lambda) &= \sum_{w=1}^{W} p(w\,|\,\lambda)\, p(\mathbf{x}^{(1)},\mathbf{x}^{(2)},\cdots,\mathbf{x}^{(M)}\,|\,w,\lambda) \\ &= \sum_{w=1}^{W} p(w\,|\,\lambda) \prod_{i=1}^{M} p(\mathbf{x}^{(i)}\,|\,w,\lambda), \end{aligned} \tag{2}$$

where λ is a GMM, $p(w\,|\,\lambda)$ is the mixture weight, and W is the number of mixtures in the GMM.

The feature vector $\mathbf{x}^{(i)}$ of a specific subband is incomplete data compared with combined feature vector $\mathbf{X}$. To compute the likelihood score of the incomplete data, marginalization is necessary. By marginalizing the likelihood scores of all of the other subbands, the likelihood score of a specific subband can be obtained as

$$\begin{aligned} & p(\mathbf{x}^{(i)}\,|\,\lambda) \\ &= \iint\cdots\int P(\mathbf{X}|\lambda)\, d\mathbf{x}^{(1)} d\mathbf{x}^{(2)} \cdots d\mathbf{x}^{(i-1)} d\mathbf{x}^{(i+1)} \cdots d\mathbf{x}^{(M)} \\ &= \iint\cdots\int \sum_{w=1}^{W} p(w|\lambda) \prod_{i=1}^{M} p(\mathbf{x}^{(i)}|w,\lambda) \quad d\mathbf{x}^{(1)} d\mathbf{x}^{(2)} \\ &\qquad \cdots d\mathbf{x}^{(i-1)} d\mathbf{x}^{(i+1)} \cdots d\mathbf{x}^{(M)} \\ &= \sum_{w=1}^{W} p(w\,|\,\lambda)\, p(\mathbf{x}^{(i)}|\,w,\lambda) \underbrace{\prod_{m=1,m\neq i} \int p(\mathbf{x}^{(m)}|\,w,\lambda)\, d\mathbf{x}^{(m)}}_{1} \\ &= \sum_{w=1}^{W} p(w\,|\,\lambda)\, p(\mathbf{x}^{(i)}|\,w,\lambda) \quad . \end{aligned} \tag{3}$$

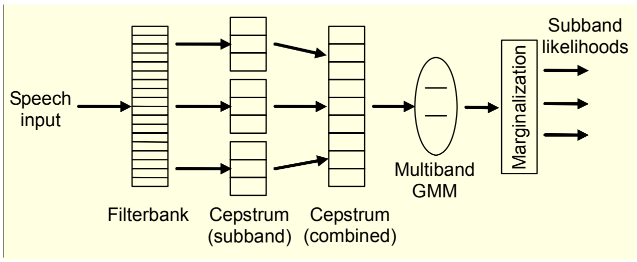


Fig. 3. Subband likelihood scoring in the feature recombination.

Using (3), the subband features are processed separately, and this process can complement the drawback of conventional feature recombination, which uses the total subband features to compute a single likelihood score. The output of multiband GMM, the subband likelihoods shown in Fig. 3, is different from the single output in Fig. 1(c).

III. Conventional Speaker Recognition System

The GMM-based speaker identification and GMM-UBM-based speaker verification are simple and provide good performance in text-independent speaker recognition. In speaker identification, given a group of speakers $S=\{1,2,\cdots,K\}$ and test speech feature vector sequence $X=\{x_1,x_2,\cdots,x_T\}$, the goal is to find the speaker model that has maximum accumulated log-likelihood:

$$\hat{S} = \arg\max_{k\in S} P(X\,|\,\lambda_k) = \arg\max_{k\in S} \sum_{t=1}^{T} \log p(x_t\,|\,\lambda_k)\,. \tag{4}$$

Speaker verification determines whether X was spoken by the claimed speaker c. In GMM-UBM-based speaker verification, the likelihood ratio is compared with a predefined threshold to decide if the claimed speaker is accepted or rejected as follows:

$$\begin{aligned} &\text{Accept if } P(X\,|\,\lambda_c)/P(X\,|\,\lambda_{UBM}) \geq \theta, \\ &\text{Reject if } P(X\,|\,\lambda_c)/P(X\,|\,\lambda_{UBM}) < \theta, \end{aligned} \tag{5}$$

$$P(X\,|\,\lambda_c) = \frac{1}{T}\sum_{t=1}^{T} \log p(x_t\,|\,\lambda_c)\,, \tag{6}$$

$$P(X\,|\,\lambda_{UBM}) = \frac{1}{T}\sum_{t=1}^{T} \log p(x_t\,|\,\lambda_{UBM})\,, \tag{7}$$

where λ_c is a claimed speaker model and λ_{UBM} is a UBM. If some noise exists in the recognition environment, a few likelihood scores of the speaker model and UBM may be very low. These very low likelihoods can cause the speaker recognition system to make an incorrect decision. To deal with this problem, we propose a reliable-feature selection method as explained in the following section.

IV. Feature Recombination-Based Speaker Recognition Using Subband Likelihood Scores and Reliable-Feature Selection

In this paper, we classify input features into reliable or unreliable features depending on the likelihood scores. When we use this approach, we must consider two points: how to determine whether a feature is reliable and how to reduce the effect of unreliable features. To find a reliable feature more effectively, we use an adaptive noise model. This noise model is estimated by the maximum *a posteriori* (MAP) adaptation technique [2] using noise features obtained from test speech after a simple voice activity detector (VAD). In the VAD, the way to decide if a frame is a speech or non-speech frame is to compare the current frame energy with the average energy of the first 10 frames of test speech. The VAD determines that a given frame is a speech or non-speech frame, and the feature extracted from the non-speech frame is roughly assumed as a noise feature. The advantages of using this adaptive noise model are that we do not need prior information about noise. Moreover, it accurately reflects the characteristics of current noise because the noise features are directly extracted from a noisy test speech. To determine whether a feature is reliable or unreliable, the likelihood of this feature for a speaker model is compared with that for an adaptive noise model. If the likelihood for a speaker model is lower than that for an adaptive noise model, this feature is determined to be an unreliable feature. To reduce the effects of lower likelihoods of unreliable features, the likelihoods of unreliable features are substituted by those of an adaptive noise model. The likelihoods of features for an adaptive noise model are assumed to be the lower bound of the likelihoods of features. Figure 4 shows the proposed speaker recognition system using reliable-feature selection. The proposed reliable-feature selection can be applied to both full-band and subband levels.

Speaker identification using full-band reliable-feature selection based on an adaptive noise model is formulated as

$$\hat{S} = \arg\max_k \tilde{P}(X \mid \lambda_k) = \arg\max_k \sum_{t=1}^{T} \log \hat{p}(x_t \mid \lambda_k)\,, \tag{8}$$

$$\hat{p}(x_t \mid \lambda_k) = \begin{cases} p(x_t \mid \lambda_k) & \text{if } p(x_t \mid \lambda_k) \ge p(x_t \mid \lambda_{Noise}), \\ p(x_t \mid \lambda_{Noise}) & \text{if } p(x_t \mid \lambda_k) < p(x_t \mid \lambda_{Noise}). \end{cases} \tag{9}$$

From (8) and (9), comparing $p(x_x \mid \lambda_k)$ and $p(x_t \mid \lambda_{Noise})$ is the process to determine whether feature x_t is reliable or unreliable. Substituting $p(x_x \mid \lambda_k)$ by $p(x_t \mid \lambda_{Noise})$ is the process of reducing the effect of an unreliable feature.

In speaker verification using full-band reliable-feature selection, the way to determine whether a feature is reliable or unreliable is the same as in speaker identification. In speaker verification using reliable-feature selection based on an adaptive noise model, the likelihood ratio test is computed using the following equations:

$$\begin{aligned} &\text{Accept if } \tilde{P}(X \mid \lambda_c)/\tilde{P}(X \mid \lambda_{UBM}) \ge \theta, \\ &\text{Reject if } \tilde{P}(X \mid \lambda_c)/\tilde{P}(X \mid \lambda_{UBM}) < \theta, \end{aligned} \tag{10}$$

$$\tilde{P}(X \mid \lambda_c) = \frac{1}{T}\sum_{t=1}^{T} \log \hat{p}(x_t \mid \lambda_c), \tag{11}$$

$$\tilde{P}(X \mid \lambda_{UBM}) = \frac{1}{T}\sum_{t=1}^{T} \log \hat{p}(x_t \mid \lambda_{UBM}), \tag{12}$$

$$\hat{p}(x_t \mid \lambda_c) = \begin{cases} p(x_t \mid \lambda_c) & \text{if } p(x_t \mid \lambda_c) \ge p(x_t \mid \lambda_{Noise}), \\ p(x_t \mid \lambda_{Noise}) & \text{if } p(x_t \mid \lambda_c) < p(x_t \mid \lambda_{Noise}), \end{cases} \tag{13}$$

$$\hat{p}(x_t \mid \lambda_{UBM}) = \begin{cases} p(x_t \mid \lambda_{UBM}) & \text{if } p(x_t \mid \lambda_{UBM}) \ge p(x_t \mid \lambda_{Noise}), \\ p(x_t \mid \lambda_{Noise}) & \text{if } p(x_t \mid \lambda_{UBM}) < p(x_t \mid \lambda_{Noise}), \end{cases} \tag{14}$$

where λ_c and λ_{UBM} are the claimed speaker model and the UBM, respectively.

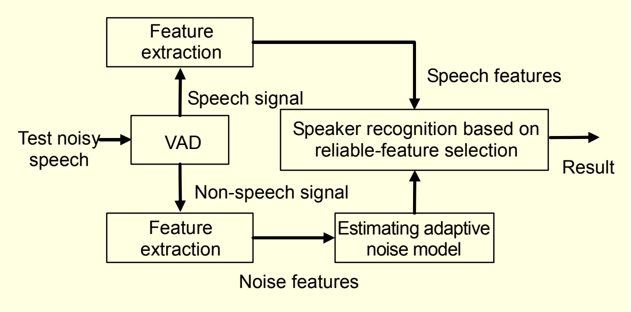


Fig. 4. Speaker recognition system based on reliable-feature selection.

Combining the subband likelihood scoring and subband reliable-feature selection, the speaker identification based on feature recombination is formulated as

$$\hat{S} = \arg\max_k \sum_{t=1}^{T}\left(\sum_{i=1}^{M} \log\left(\hat{p}(x_t^{(i)} \mid \lambda_k)\right)\right), \tag{15}$$

$$\hat{p}(x_t^{(i)} \mid \lambda_k) = \begin{cases} p(x_t^{(i)} \mid \lambda_k) & \text{if } p(x_t^{(i)} \mid \lambda_k) \ge p(x_t^{(i)} \mid \lambda_{Noise}), \\ p(x_t^{(i)} \mid \lambda_{Noise}) & \text{if } p(x_t^{(i)} \mid \lambda_k) < p(x_t^{(i)} \mid \lambda_{Noise}), \end{cases} \tag{16}$$

where M is the number of subbands. In the speaker verification based on feature recombination using subband likelihood scoring and subband reliable-feature selection, the likelihood ratio test is modified as follows:

$$\begin{aligned}&\text{Accept if } \tilde{P}(X\,|\,\lambda_c)/\tilde{P}(X\,|\,\lambda_{UBM}) \ge \theta,\\ &\text{Reject if } \tilde{P}(X\,|\,\lambda_c)/\tilde{P}(X\,|\,\lambda_{UBM}) < \theta,\end{aligned} \tag{17}$$

$$\tilde{P}(X\,|\,\lambda_c) = \frac{1}{T}\sum_{t=1}^{T}\left(\sum_{i=1}^{M}\log \hat{p}(x_t^{(i)}\,|\,\lambda_c)\right), \tag{18}$$

$$\tilde{P}(X\,|\,\lambda_{UBM}) = \frac{1}{T}\sum_{t=1}^{T}\left(\sum_{i=1}^{M}\log \hat{p}(x_t^{(i)}\,|\,\lambda_{UBM})\right), \tag{19}$$

$$\begin{aligned}&\hat{p}(x_t^{(i)}\,|\,\lambda_c)\\ &=\begin{cases} p(x_t^{(i)}\,|\,\lambda_c) & \text{if } p(x_t^{(i)}\,|\,\lambda_c) \ge p(x_t^{(i)}\,|\,\lambda_{Noise}),\\ p(x_t^{(i)}\,|\,\lambda_{Noise}) & \text{if } p(x_t^{(i)}\,|\,\lambda_c) < p(x_t^{(i)}\,|\,\lambda_{Noise}),\end{cases}\end{aligned} \tag{20}$$

$$\begin{aligned}&\hat{p}(x_t^{(i)}\,|\,\lambda_{UBM})\\ &=\begin{cases} p(x_t^{(i)}\,|\,\lambda_{UBM}) & \text{if } p(x_t^{(i)}\,|\,\lambda_{UBM}) \ge p(x_t^{(i)}\,|\,\lambda_{Noise}),\\ p(x_t^{(i)}\,|\,\lambda_{Noise}) & \text{if } p(x_t^{(i)}\,|\,\lambda_{UBM}) < p(x_t^{(i)}\,|\,\lambda_{Noise}).\end{cases}\end{aligned} \tag{21}$$

In the proposed reliable-feature selection, the likelihoods of unreliable features are substituted by those of the adaptive noise model. However, exclusion of unreliable features has to be considered. From the experimental results for exclusion of unreliable features in the next section, the substitution of unreliable feature yielded slightly better performance in both speaker identification and verification. We analyze the effects of substitution and exclusion on speaker identification and verification separately. First, to analyze this effect on speaker identification, we compute the mean and variance of the differences of normalized log-likelihoods of the best and second best speakers after decoding for each test utterance. Table 1 shows the means and variances of the differences.

Table 1. Means and variances of the differences between log-likelihoods of best and second best speakers after speaker identification over various SNRs of airport noise from the TIMIT database.

Method / SNR	Exclusion		Substitution	
	Mean	Variance	Mean	Variance
20 dB	0.146	0.0115	0.388	0.0755
15 dB	0.128	0.0105	0.333	0.0646
10 dB	0.099	0.0079	0.238	0.0462
5 dB	0.064	0.0044	0.138	0.0197
0 dB	0.043	0.0020	0.084	0.0078

Table 2. Means and variances of the log-likelihood ratios of test utterances over various SNRs of airport noise from the TIMIT database.

Method / SNR	Exclusion				Substitution			
	Claimed		Impostor		Claimed		Impostor	
	Mean	Var.	Mean	Var.	Mean	Var.	Mean	Var.
20 dB	0.29	0.021	-0.22	0.024	0.22	0.021	-0.21	0.020
15 dB	0.27	0.021	-0.21	0.024	0.19	0.019	-0.20	0.017
10 dB	0.23	0.020	-0.19	0.022	0.15	0.012	-0.18	0.013
5 dB	0.18	0.018	-0.16	0.020	0.09	0.010	-0.15	0.011
0 dB	0.13	0.017	-0.12	0.018	0.04	0.006	-0.11	0.007

Table 3. Ratios of reliable features in test utterances over various SNRs in airport noise condition of TIMIT database. (%)

SNR / System	20 dB	15 dB	10 dB	5 dB	0 dB
Speaker identification	80.59	77.03	72.52	63.78	54.61
Speaker veritication	81.33	78.10	73.33	66.83	58.37

From Table 1, the mean of substitution is larger than that of exclusion. As the difference between likelihood scores of the best and second best speakers become smaller, the confusion between the best and second best speakers increases more than the larger one, and this higher confusion causes the speaker identification system to make more incorrect decisions. Therefore, substitution shows better performance than exclusion in speaker identification. Next, to analyze the effect of exclusion and substitution on speaker verification, we compute the means and variances of log-likelihood ratios of test utterances. These test utterances are separated according to whether claimed speakers or impostors have spoken. The means and variances are shown in Table 2.

As shown in Table 2, the variances of log-likelihood ratios of substitution are smaller than those of exclusion for both claimed speakers and impostors. As the variances of log-likelihood ratios are reduced, the rates of false alarms and false rejects are reduced in comparison with those in the case of large variance. Table 3 represents the ratios of reliable features in test utterances over various signal-to-noise ratios (SNRs) under airport noise conditions from the TIMIT database during speaker recognition.

The ratio of reliable features increases as the SNR of test utterances becomes higher. The reason that the ratios of speaker identification and verification are different is that in speaker verification, there are impostor utterances.

V. Experiments

1. Database

We evaluate speaker recognition systems in noisy environments using the TIMIT and NTIMIT databases. The NTIMIT database is corresponding telephone version of the TIMIT database [14]. For our experiments, 100 male and 100 female speakers are selected as enrolled speakers, and 158 male and 42 female speakers are used as impostors in both the TIMIT and NTIMIT database, respectively. A UBM is trained from an additional 50 male and 50 female speakers. Of ten sentences uttered by each enrolled speaker, five sentences are used to estimate the speaker GMMs based on MAP adaptation, and the other five sentences are assigned to evaluate the speaker identification system. Additionally, five sentences of each impostor are used to evaluate speaker verification. For the noise condition, we artificially added acoustic noise from the Aurora 2 noise database [15] to clean test speech down-sampled to 8 kHz for various SNRs. The speech analysis frame rate is set to 20 ms with 10 ms intervals. The UBM, speaker models, and the adaptive noise model contain 160 Gaussian components. In the case of the full-band system, the eighteen-dimensional (18D) MFCCs are extracted from the outputs of 33 channels. The details of the front-end in the multiband system are presented in Table 4. The final dimensions of features are 18 for the two-subband and three-subband systems and 16 or 20 for the four-subband system.

Table 4. Channel numbers and dimensions of MFCCs in the multiband system.

System / Parm.	Multiband system			
	2 subbands	3 subbands	4 subbands (16D)	4 subbands (20D)
Channel number	32	33	32	32
MFCC (dimension)	9 (18)	6 (18)	4 (16)	5 (20)

2. Experimental Results of Speaker Identification

A. Evaluation of the TIMIT Database

Table 5 shows the error rates and error reduction rates (ERR) of speaker identification systems under clean conditions. Conventional feature recombination degrades the performance under clean conditions, but the proposed feature recombination using subband likelihoods, or subband feature recombination, improves performance under clean conditions.

Table 6 shows the error rates of the full-band system for eight kinds of noise with various SNRs.

Table 5. Error rates of the full-band, feature recombination (FR), and subband feature recombination (SFR), and error reduction rates over full band under clean conditions (TIMIT database).

(%)

		Error rate	ERR
Full-band		6.1	-
FR	2 subbands	9.9	-62.29
	3 subbands	9.5	-55.74
	4 subbands (16D)	9.8	-60.66
	4 subbands (20D)	11.6	-90.16
SFR	2 subbands	5.0	18.03
	3 subbands	4.3	29.50
	4 subbands (16D)	5.9	3.28
	4 subband (20D)	6.2	-1.64

Table 6. Error rates of full-band system for various noise types with various SNRs (TIMIT database).

(%)

SNR / Noise	20 dB	15 dB	10 dB	5 dB	0 dB
Airport	11.7	18.1	36.4	61.9	83.8
Babble	13.3	26.8	51.0	72.2	85.1
Car	19.9	39.0	61.4	82.4	92.6
Exhibition	30.7	56.9	78.1	89.2	94.5
Restaurant	13.7	25.1	45.1	68.7	85.1
Street	9.3	14.2	24.8	41.4	69.9
Subway	33.2	56.3	76.2	87.6	94.2
Train	15.1	26.8	50.2	71.6	89.9
Average	18.36	32.90	52.90	71.88	86.89

A comparison of feature recombination and subband feature recombination is shown in Table 7. Subband feature recombination achieves better performance than the conventional method, especially at high SNRs. The feature recombination system and the subband feature recombination system with 3 subbands achieve the best average ERRs of 3.05% and 8.33%, respectively, over the full-band system. The performance improvement of the conventional feature recombination is not notable compared with the full-band system under realistic noise conditions.

To verify the effectiveness of the reliable-feature selection, we employ a widely used SNR-based frame-dropping technique as a comparative method. To drop a frame, if the SNR of a given frame is lower than a threshold, this frame is excluded in test speech. In these experiments, the SNR threshold is set to 30 dB. The SNR estimation at the *t*-th frame is

Table 7. ERRs of the conventional feature recombination (FR) and the subband feature recombination (SFR) over full-band system (TIMIT database). (%)

Method \ SNR		Error reduction rate over full-band					
		20 dB	15 dB	10 dB	5 dB	0 dB	Ave.
2 sub.	FR	-2.25	7.18	5.60	0.77	-1.34	1.30
	SFR	16.32	14.24	8.92	2.03	-0.45	8.21
3 sub.	FR	2.18	8.40	5.27	0.82	-1.40	3.05
	SFR	17.29	16.11	8.13	1.06	-0.92	8.33
4 sub. (16D)	FR	-0.41	5.47	3.57	-0.45	-1.65	1.99
	SFR	1.91	4.41	4.99	-0.50	-1.70	1.82
4 sub. (20D)	FR	-14.77	3.31	5.95	4.31	2.88	0.34
	SFR	12.39	16.75	11.03	4.30	2.88	9.47

Table 8. Error rates of conventional feature recombination (FR), and ERRs of the conventional feature recombination with frame dropping (FR+FD) or full-band reliable-feature selection (FR+FRFS), and subband feature recombination combined with frame dropping (SFR+FD) or subband reliable-feature selection (SFR+SRFS) (3 subband system, TIMIT database). (%)

Method \ SNR	FR error rate (%)	Error reduction rate over FR			
		FR		SFR	
		+ FD	+ FRFS	+ FD	+ SRFS
20 dB	17.96	2.85	8.28	25.26	30.34
15 dB	30.14	7.63	11.90	23.14	33.51
10 dB	50.11	10.48	13.79	17.98	30.41
5 dB	71.29	6.87	11.12	9.84	21.34
0 dB	88.10	3.72	6.85	1.84	12.57
Average ERR		6.31	10.39	15.61	25.63

computed as

$$SNR_t = 10\log_{10}\left[\frac{\sum_{k=1}^{K}\left|S_t(k)\right|^2}{\sum_{k=1}^{K}\left|\overline{N}(k)\right|^2}\right], \tag{22}$$

$$\left|S_t(k)\right| = \max\left\{\left|X_t(k)\right| - 1.1\left|\overline{N}(k)\right|,\ 0.001\left|\overline{N}(k)\right|\right\}, \tag{23}$$

where k is the frequency index, $\left|X_t(k)\right|$ and $\left|S_t(k)\right|$ are the magnitude spectra of noisy speech and estimated speech, respectively, and $\left|\overline{N}_t(k)\right|$ is the averaged magnitude spectrum of noise. Noise power is estimated by averaging the non-speech frames in each utterance. Whether a frame is speech or non-speech is determined by simply comparing the current frame energy with the average energy of the first 10 frames in input test speech.

Table 9. Speaker identification accuracy of subband feature recombination using the reliable-feature selection in the exclusion or substitution process (3 subband system, TIMIT database). (%)

SNR \ Method	Subband feature recombination	
	+ SRFS (exclusion)	+ SRFS (substitution)
20 dB	84.82	87.49
15 dB	77.40	79.96
10 dB	61.96	65.13
5 dB	40.24	43.93
0 dB	19.25	22.98

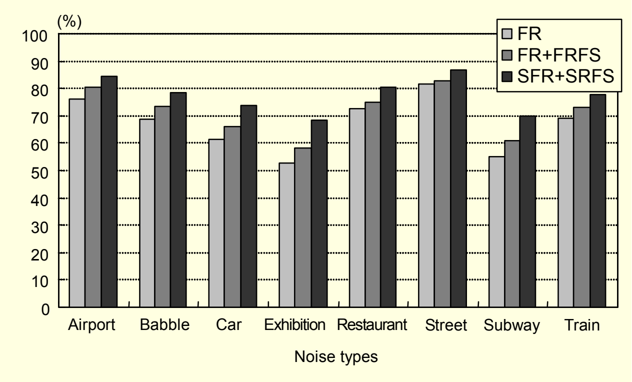


Fig. 5. Average speaker-identification accuracy of conventional feature recombination (FR), conventional feature recombination combined with full-band reliable-feature selection (FR+FRFS), and subband feature recombination together with sub-band reliable-feature selection (SFR+SRFS) (3 sub-band system, TIMIT database).

The performance improvements of the proposed methods are shown in Table 8. Combining the subband feature recombination with subband reliable-feature selection techniques yields much better performance improvement than combining the conventional feature recombination with full-band reliable-feature selection.

In Table 9, the likelihood of the unreliable feature is excluded from the accumulated likelihood computation instead of the substitution process of the proposed method. The results demonstrate that the substitution process achieves better performance than the exclusion process.

Figure 5 shows the accuracy averages of speaker identification for various noise types at SNRs of 20 dB, 15 dB, and 10 dB. These SNRs are the minimum SNRs at which the speaker identification system has to guarantee satisfactory performance in real applications. These results demonstrate that the subband feature recombination system together with subband reliable-feature selection shows the best performance

Table 10. Error rates of full-band system for various noise types at various SNRs (NTIMIT database). (%)

Noise \ SNR	20 dB	15 dB	10 dB	5 dB	0 dB
Airport	35.4	42.2	51.7	64.5	81.8
Babble	34.8	41.2	57.8	76.9	89.7
Car	37.3	48.8	65.0	83.4	93.3
Exhibition	42.6	56.6	78.0	90.6	96.7
Restaurant	34.2	39.8	53.6	68.6	86.5
Street	34.6	41.1	51.3	68.8	84.2
Subway	52.6	65.3	83.8	95.4	97.4
Train	42.1	52.1	61.3	74.9	87.8
Average	39.19	48.39	62.81	77.89	89.68

Table 11. ERRs of conventional feature recombination (FR) and subband feature recombination (SFR) over full-band system (NTIMIT database). (%)

Method \ SNR		Error reduction rate over full-band					
		20 dB	15 dB	10 dB	5 dB	0 dB	Ave.
2 sub.	FR	-0.32	1.83	-0.16	-1.20	-1.02	-0.17
	SFR	2.62	1.78	-1.93	-1.65	-0.71	0.02
3 sub.	FR	4.15	6.97	4.43	3.18	1.21	3.99
	SFR	3.00	5.71	3.44	3.51	2.70	3.67
4 sub. (16D)	FR	-2.39	0.03	-0.40	-0.05	0.53	-0.46
	SFR	-6.35	-2.61	-0.68	1.46	1.80	-1.27
4 sub. (20D)	FR	10.11	13.77	13.51	11.15	6.87	11.08
	SFR	11.64	15.27	15.36	13.71	9.65	13.12

for all kinds of noise.

B. Evaluation of NTIMIT Database

Table 10 shows the identification error rates of the full-band system under various noise conditions. Table 11 presents a performance comparison of feature recombination and subband feature recombination under noise conditions according to the number of subbands. As Table 11 shows, in the case using 4 subbands with 20 dimensional feature vectors, feature recombination and subband feature recombination achieve the best performance, and subband feature recombination yields better performance than feature recombination.

Table 12 shows the performance of various speaker identification systems. The proposed system, which uses subband feature recombination and subband reliable-feature selection, obtains the highest performance improvement compared with the other systems in terms of average ERR; however, at high SNRs of 20 dB and 15 dB, feature recombination combined with full-band reliable-feature selection shows slightly better performance than subband feature recombination together with subband reliable-feature selection. The results of Tables 8 and 12 demonstrate that, compared with the conventional method, the proposed reliable-feature selection method yields better performance improvement in cases in which there is no convolution noise than in cases in which there is convolution noise. This is because the proposed method is not a method to compensate the effects of convolution noise; rather, it is a method to compensate the effects of additive noise. Table 13 shows speaker identification accuracy for cases in which the likelihoods of unreliable features are substituted or excluded. As in the previous section, the substitution process shows slightly better performance than exclusion. Figure 6 shows the average speaker identification performance at SNRs of 20 dB, 15 dB, and 10 dB. With some noise types, namely, exhibition noise and street noise, subband feature recombination with

Table 12. Error rates of conventional feature recombination (FR), ERRs of the conventional feature recombination combined with frame dropping (FR+FD) or full-band reliable-feature selection (FR+FRFS), and subband feature recombination together with frame dropping (SFR+FD) or subband reliable-feature selection (SFR+SRFS) (4 subband system (20D), NTIMIT database). (%)

SNR \ Method	FR error rate	Error reduction rate over FR			
		FR		SFR	
		+ FD	+ FRFS	+ FD	+ SRFS
20 dB	35.23	-1.85	4.05	-0.57	2.16
15 dB	41.73	0.87	6.95	2.31	6.74
10 dB	54.33	4.53	6.53	8.12	12.98
5 dB	69.20	4.71	7.19	7.37	14.11
0 dB	83.51	3.25	5.57	1.26	13.65
Average ERR over FR		2.30	6.06	3.70	9.93

Table 13. Speaker identification accuracy of subband feature recombination using the subband reliable-feature selection in exclusion or substitution process (4 subband system (20D), NTIMIT database). (%)

SNR \ Method	Subband feature recombination	
	+ SFRS (exclusion)	+ SFRS (substitution)
20 dB	60.4	65.54
15 dB	54.65	61.09
10 dB	44.35	52.73
5 dB	30.46	40.56
0 dB	17.50	27.88

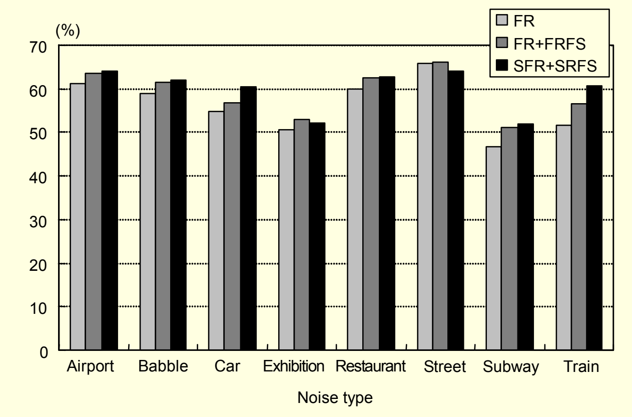


Fig. 6. Average speaker identification accuracy of conventional feature recombination (FR), conventional feature recombination together with full-band reliable-feature selection (FR+FRFS), and subband feature recombination together with subband reliable-feature selection (SFR+SRFS) (4 subband system, NTIMIT database).

subband reliable-feature selection yields slightly worse performance than feature recombination with full-band reliable-feature selection; however, this results demonstrates that the proposed reliable-feature selection technique is effective in noisy environments.

Table 14. EERs of the full-band, feature recombination (FR), and subband feature recombination (SFR), and ERRs over full band under clean condition (TIMIT database). (%)

		EER	ERR
Full-band		2.7	-
FR	2 subbands	3.5	-29.63
	3 subbands	3.1	-14.81
	4 subbands (16D)	2.8	-3.7
	4 subbands (20D)	4.4	-62.96
SFR	2 subbands	2.0	25.93
	3 subbands	2.1	22.22
	4 subbands (16D)	2.3	14.81
	4 subband (20D)	2.3	14.81

Table 15. EERs of full-band system for various noise types with various SNRs (TIMIT database). (%)

SNR / Noise	20 dB	15 dB	10 dB	5 dB	0 dB
Airport	3.6	5.1	7.6	13.9	21.1
Babble	4.2	6.5	11.6	19.0	27.0
Car	5.1	8.3	15.4	24.2	32.3
Exhibition	7.3	14.6	23.4	33.2	42.0
Restaurant	3.8	5.9	10.8	18.0	27.6
Street	3.3	3.8	5.8	8.7	16.1
Subway	8.1	14.3	24.1	32.6	40.0
Train	4.0	6.5	10.6	17.0	23.6
Average	4.93	8.13	13.66	20.83	28.71

Table 16. ERRs of the conventional feature recombination (FR) and the subband feature recombination (SFR) over full-band system (TIMIT database). (%)

SNR / Method		Error reduction rate over full-band					
		20 dB	15 dB	10 dB	5 dB	0 dB	Ave.
2 sub.	FR	-14.21	-2.31	2.84	2.88	0.09	-2.14
	SFR	16.50	25.23	24.79	18.49	8.88	18.78
3 sub.	FR	-10.91	2.92	10.61	9.66	4.92	3.44
	SFR	5.84	18.00	22.60	21.07	14.37	16.37
4 sub. (16D)	FR	-10.21	-4.30	10.80	12.30	6.53	3.03
	SFR	3.30	15.08	18.21	16.69	12.28	13.11
4 sub. (20D)	FR	-48.73	-24.62	-9.42	-0.84	-0.78	-16.88
	SFR	4.65	18.46	22.05	19.75	12.58	15.68

3. Experimental Results of Speaker Verification

A. Evaluation of the TIMIT Database

To investigate the operational characteristics of the conventional and the proposed speaker verification systems, two error terms are used: false alarm rate and false rejection rate. As the threshold changed, we found equal error rates (EERs). The EER is the error rate when the false alarm rate is equal to the false rejection rate. Table 14 shows a comparison with the full-band using feature recombination and subband feature recombination under clean conditions.

Tables 15 and 16 present the performance of the full-band system, feature recombination, and subband feature recombination under noisy conditions. Conventional feature recombination obtains slightly better performance than the full-band system as in the experiment of speaker identification, but subband feature recombination is more effective than conventional feature recombination under noisy conditions.

The performance of feature recombination and subband feature recombination together with frame dropping, full-band reliable-feature selection, or subband reliable-feature selection are shown in Table 17. When frame dropping or reliable-feature selection is combined with subband feature recombination, the performance is dramatically improved compared with that of conventional feature recombination. Therefore, reliable-feature selection is more effective than the frame dropping. A comparison of the exclusion and substitution of the likelihoods of unreliable features is shown in Table 18. Substitution achieves better performance than that of

Table 17. EERs of conventional feature recombination (FR) and ERRs of the conventional feature recombination combined with frame dropping (FR+FD) or full-band reliable-feature selection (FR+FRFS), and subband feature recombination combined with frame dropping (SFR+FD) or subband reliable-feature selection (SFR+SRFS) (3 subband system, TIMIT database). (%)

Method SNR	FR EER	Error reduction rate over FR			
		FR		SFR	
		+ FD	+ FRFS	+ FD	+ SRFS
20 dB	5.46	-6.18	6.18	18.99	21.97
15 dB	7.89	0.63	7.77	25.83	29.00
10 dB	12.21	4.30	8.50	29.58	33.47
5 dB	18.81	2.79	8.64	23.06	35.81
0 dB	27.28	2.34	10.68	14.12	35.61
Average ERR		0.77	8.35	22.31	31.17

Table 18. EERs of subband feature recombination using reliable-feature selection with exclusion or substitution (3 subband system, TIMIT database). (%)

Method SNR	Subband feature recombination	
	+ SRFS (exclusion)	+ SRFS (substitution)
20 dB	4.38	4.26
15 dB	5.96	5.60
10 dB	8.84	8.13
5 dB	13.36	12.08
0 dB	18.16	17.56

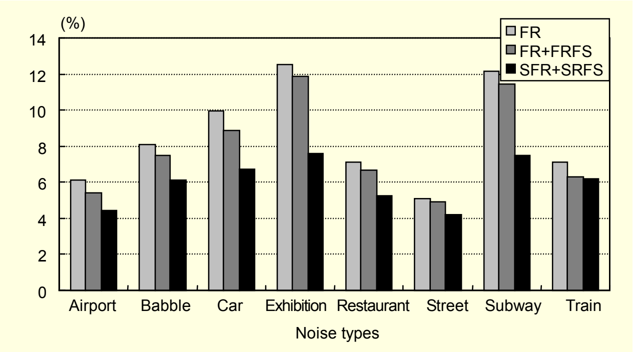


Fig. 7. Average EERs of speaker verification using conventional feature recombination (FR), conventional feature recombination together with full-band reliable-feature selection (FR+FRFS), and subband feature recombination together with subband reliable-feature selection (SFR+SRFS) (3 subband system, TIMIT database).

the exclusion as in speaker identification.

The average EERs of speaker verification for various noise types at SNRs of 20 dB, 15 dB, and 10 dB are given in Fig 7.

Table 19. EERs of full-band system for various noise types with various SNRs (NTIMIT database). (%)

SNR Noise	20 dB	15 dB	10 dB	5 dB	0 dB
Airport	9.0	10.9	11.8	15.2	21.8
Babble	9.3	10.5	14.2	20.7	31.1
Car	9.7	10.6	15.2	23.6	31.0
Exhibition	11.0	14.7	21.8	29.6	37.1
Restaurant	9.4	9.7	12.3	18.1	27.6
Street	9.4	9.9	12.9	16.4	22.1
Subway	12.0	16.0	23.2	31.7	36.5
Train	11.1	12.2	14.9	17.8	24.9
Average	10.11	11.81	15.79	21.64	29.01

Table 20. ERRs of conventional feature recombination (FR) and subband feature recombination (SFR) over full-band system (NTIMIT database). (%)

SNR Method		Error reduction rate over full-band					
		20 dB	15 dB	10 dB	5 dB	0 dB	Ave.
2 sub.	FR	0.49	-2.12	-5.94	-6.12	-7.28	-4.19
	SFR	7.58	6.11	6.80	6.86	1.45	5.76
3 sub.	FR	7.91	-1.90	3.64	3.87	-1.29	2.45
	SFR	10.26	7.94	10.61	12.71	10.43	10.39
4 sub. (16D)	FR	0.37	-1.74	1.74	-1.10	-2.97	-0.75
	SFR	-0.25	1.38	7.10	9.01	7.71	5.01
4 sub. (20D)	FR	2.10	0.95	2.85	1.62	-0.17	1.47
	SFR	17.92	15.87	19.56	22.13	18.74	18.84

From this result, the subband feature recombination system together with subband reliable-feature selection achieves the best performance for all kinds of noise as in speaker identification.

B. Evaluation of NTIMIT Database

Tables 19 and 20 present EERs of the speaker verification system for full-band, and a performance comparison of feature recombination and subband feature recombination in terms of ERR over the full-band system. Subband feature recombination with 4 subbands (20D) achieves the best performance.

Table 21 shows ERRs of the various speaker verification systems over conventional feature recombination. Subband feature recombination together with subband reliable-feature selection make speaker verification more noise-robust than conventional feature recombination together with full-band reliable-feature selection. Figure 8 shows the average EERs of

Table 21. EERs of conventional feature recombination (FR), conventional feature recombination together with frame dropping (FR+FD) or full-band reliable-feature selection (FR+FRFS), and subband feature recombination together with frame dropping (SFR+FD) or subband reliable-feature selection (SFR+SRFS) (4 subband system (20D), NTIMIT database). (%)

Method / SNR	FR EER (%)	Error reduction rate over FR			
		FR		SFR	
		+ FD	+ FRFS	+ FD	+ SRFS
20 dB	9.90	-6.44	4.29	17.42	18.31
15 dB	11.70	-3.74	5.98	19.23	21.26
10 dB	15.34	0.08	6.11	21.52	26.89
5 dB	21.29	1.94	9.10	20.78	33.71
0 dB	29.06	4.30	8.34	22.19	35.74
Average ERR		-0.77	6.77	20.23	27.18

Table 22. EERs of subband feature recombination using subband reliable-feature selection with exclusion or substitution (4 subband system (20D), NTIMIT database). (%)

Method / SNR	Subband feature recombination	
	+ SRFS (exclusion)	+ SRFS (substitution)
20 dB	8.66	8.09
15 dB	10.06	9.21
10 dB	13.25	11.21
5 dB	17.75	14.11
0 dB	25.18	18.68

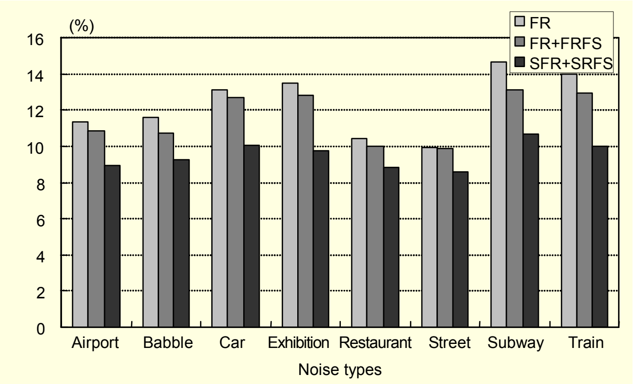


Fig. 8. Average EERs of speaker verification using conventional feature recombination (FR), conventional feature recombination together with full-band reliable-feature selection (FR+FRFS), and subband feature recombination together with subband reliable-feature selection (SFR+SRFS) (4 subband system (20D), NTIMIT database).

the conventional and proposed methods for various noise types at SNRs of 20 dB, 15 dB, and 10 dB. The proposed method shows better performance with all kinds of noise.

VI. Conclusions

In this paper, we introduced a method to compute subband likelihoods for feature recombination in a multiband approach to speaker recognition and verification. We proposed subband feature recombination using subband likelihood. The proposed subband feature recombination method is more effective than conventional feature recombination for both speaker identification and verification under broadband noisy conditions. In addition, reliable-feature selection for noise robust speaker recognition is proposed. The experimental results demonstrate that the proposed reliable-feature selection achieves better performance than the conventional SNR-based frame-dropping technique for both speaker identification and verification. When the proposed method is used together with subband feature recombination and subband reliable-feature selection, the performance improvements are remarkable in speaker identification and verification. Finally, we analyzed and tested the cases of likelihood exclusion and substitution of unreliable features. Experimental results demonstrated that the proposed likelihood substitution is more effective than likelihood exclusion in speaker recognition.

References

[1] D. Reynold and R.C. Rose, "Robust Text Independent Speaker Identification Using Gaussian Mixture Speaker Models," *Proc. IEEE Tran. Speech and Audio Processing*, vol. 3, Jan. 1995, pp. 72-83.

[2] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, 2000, pp. 19-41.

[3] A. Drygajlo and M. El-Maliki, "Speaker Verification in Noisy Environments with Combined Spectral Subtraction and Missing Feature Theory," *Proc. ICASSP*, vol. 2, 1998, pp. 121-124.

[4] C. Barras and J. Gauvain, "Feature and Score Normalization for Speaker Verification of Cellular Data," *Proc. ICASSP*, 2003, pp. 49-52.

[5] K. Yiu, M. Mak, and S. Kung, "Environment Adaptation for Robust Speaker Verification," *Proc. EUROSPEECH*, 2003, pp. 2973-2976.

[6] H.J. Qing, Z. Lei, and W. Chengfa, "An Environment Adaptation Method for Robust Speech Recognition," *Proc. ICSP*, 2000, pp. 726-729.

[7] D. Ramos-Castro, J. Fierrez-Aquilar, J. Gonzalez-Rodriquez, and J. Ortega-Garcia, "Speaker Verification Using Speaker- and Test-Dependent Fast Score Normalization," *Pattern Recognition Letters*, vol. 28, 2007, pp. 90-98.

[8] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score Normalization for Text-Independent Speaker Verification

Systems," *Digital Signal Processing*, vol. 10, 2000, pp. 42-54.

[9] S. Okawa, E. Bocchieri, and A. Potamianos, "Multiband Speech Recognition in Noise Environments," *Proc. ICASSP*, 1998, pp. 641-644.

[10] H. Hermansky, S. Tibrewala, and M. Pavel, "Toward ASR on Partially Corrupted Speech," *Proc. ICSLP*, 1996.

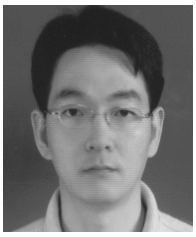
[11] S. Tibrewala and H. Hermansky, "Subband Based Recognition of Noisy Speech," *Proc. ICASSP*, 1997, pp. 1255-1258.

[12] W. Chen, C. Hsieh, and E. Lai, "Multiband Approach to Robust Text-Independent Speaker Identification," *Computational Linguistics and Chinese Language Processing*, vol. 9, no. 2, 2004, pp. 63-76.

[13] B. Mak, "A Mathematical Relationship Between Full-Band and Multiband Mel-Frequency Cepstral Coefficients," *IEEE Signal Processing Letters*, vol. 9, no. 8, 2002, pp. 241-244.

[14] D. Reynold, "Large Population Speaker Identification Using Clean and Telephone Speech," *IEEE Signal Processing Letters*, vol. 2, no. 3, 1995, pp. 46-48.

[15] D. Pearce and H. Hirsch, "The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noise Conditions," *Proc. ICSLP*, vol. 4, 2000, pp. 29-32.



Sungtak Kim received the BS degree in electronics engineering from the Ulsan University, and the MS degree in multimedia communications and processing from the Information and Communications University (ICU), Rep. of Korea, in 2000 and 2003, respectively. He is currently pursuing a PhD degree in multimedia communications and processing at the Information and Communications University. His research interests are robust speech recognition and speaker recognition.

Mikyong Ji received the BS degree in information engineering from the Hansung University, and the MS degree in multimedia communications and processing from the Information and Communications University (ICU), Rep. of Korea, in 2000 and 2002, respectively. She is currently pursuing a PhD degree in multimedia communications and processing at the Information and Communications University. Her research interests are distant-talking speech recognition, multi-microphone speaker recognition, environment compensation, and BBN.

Hoirin Kim received the MS and PhD degrees from the Dept. of Electrical and Electronics Engineering, KAIST, Rep. of Korea, in 1987 and 1992, respectively. From October 1987 to December 1999, he was a senior researcher with the Spoken Language Processing Lab. at the Electronics and Telecommunications Research Institute (ETRI). From June 1994 to May 1995, he was on leave to the ATR-ITL, Kyoto, Japan. From July 2006 to July 2007, he was with the Institute of Neural Computation, UCSD, USA as a visiting researcher. Since January 2000, he has been an associate professor with Information and Communications University (ICU), Rep. of Korea. His research interests are signal processing for speech and speaker recognition, audio indexing and retrieval, and spoken language processing.