Hard-Mask Missing Feature Theory for Robust Speaker Recognition

Shin-Cheol Lim, Sei-Jin Jang, Soek-Pil Lee, and Moo Young Kim, *Senior Member*, IEEE

**Abstract — ***Compared with conventional full-band speaker recognition systems, Advanced Missing Feature Theory (AMFT) produces a much lower error rate, but requires increased computational complexity. We propose a weighting function for the score calculation algorithm in AMFT. The weighting function is estimated by calculating the number of reliable spectral components. A modified mask is also proposed to reduce the number of reliable components based on the estimated weighting function. In the proposed Hard-mask MFT-8 (HMFT-8), only 8 elements are selected out of 10 spectral components in a feature vector. Compared with the full-band system and the AMFT, the proposed HMFT-8 gives a lower identification error rate by 16.95% and 2.67%, respectively. In terms of computational complexity, AMFT and HMFT-8 require 307 and 41 arithmetic and conditional operations for each frame, respectively*[1].

Index Terms — Speaker recognition, missing feature theory, MFT, AMFT.

I. Introduction

Biometrics such as speaker, iris, fingerprint, and face recognition have gained a great deal of attention in modern consumer devices. Among them, speaker recognition can be implemented with a simpler interface between human and computer [1], [2].

Fig. 1 illustrates the basic building blocks of the speaker recognition system and some examples for consumer devices. From the input speech signal, feature vectors are extracted for the likelihood score calculation with multiple speaker models. The best matched model provides the speaker identity of the given speech signal. This system can be applied to access control to consumer devices, secure e-banking through mobile phones, and Music Information Retrieval (MIR) based on the user information. We also applied the designed speaker recognition system to a Karaoke machine to generate a user-specific favorite-song list.

However, the performance of speaker identification is adversely affected by background noise. In handheld devices, noise characteristics are highly time-varying because of the mobile nature of the system [2]. The identification rate is dramatically decreased even with a small amount of background noise. Thus, recent studies have investigated robust speaker identification against background noise by reducing the influence of background noise. For this purpose, speech enhancement and Missing Feature Theory (MFT) have been proposed as major feature-domain approaches [3]-[15].

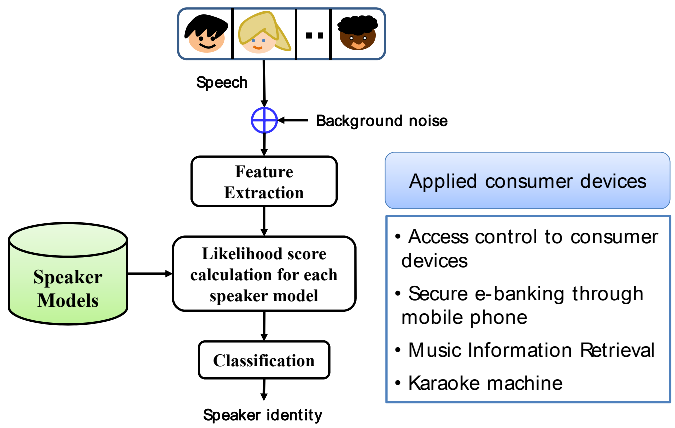


Fig. 1. Block diagram of the speaker recognition system and its application to the consumer devices.

In speech enhancement, the estimated noise spectrum is subtracted from the input noisy speech spectrum based on spectral subtraction [3] and Wiener filter [4]. Noise estimation techniques include Voice Activity Detection (VAD), Minimum Statistics (MS) [5], Weighted Spectral Averaging (WSA) [6], and Improved Minima Controlled Recursive Averaging (IMCRA) [7]. Using speech enhancement as a preprocessor of the speaker identification system, we can obtain an increased identification rate. However, the increase is not obvious especially for non-stationary noise. Precise estimation of particular noise characteristics at a given time instant remains an issue for non-stationary noise. Since most noises, such as babble noise and mobile phone ringtone, have highly time-varying characteristics, speech enhancement may not be a practical tool for robust speaker identification.

MFT has been proposed as a feature-domain approach. In MFT, a prior knowledge of the noise statistics is assumed to be given. Based on the Signal-to-Noise Rate (SNR) estimate [8], Bayesian estimation [9], and combined approach [10], MFT determines a mask that defines a time-frequency component as reliable or unreliable one. A set of reliable time-frequency components that are relatively less corrupted by the background noise can be selected as a feature vector for speaker identification.

[1]This work was supported by the Ministry of Knowledge Economy grant funded by the Korea government (No. 10037244).

S.-C. Lim and M. Y. Kim* are with the Human Computer Interaction Laboratory, Department of Information and Communication Engineering, Sejong University, Seoul, Korea (e-mail: en.shincheol@gmail.com and mooyoung@sejong.ac.kr*).

S.-J. Jang and S.-P. Lee are with the Digital Media Research Center, Korea Electronics Technology Institute, Seoul, Korea (e-mail: sjjang@keti.re.kr and lspbio@keti.re.kr).

Contributed Paper
Manuscript received 07/02/11
Current version published 09/19/11
Electronic version published 09/19/11.

Extended MFT (EMFT) calculates the likelihood scores for all possible combinations of features and selects the most reliable one without using prior knowledge of the noise statistics [11-14]. EMFT was reported to produce a better recognition rate than MFT at a cost of huge computational complexity. A bottom-up score-calculation algorithm was introduced in Advanced Missing Feature Theory (AMFT) to reduce the computational complexity of EMFT [15]. AMFT also considers the cross-term in Gaussian mixture calculation such that its recognition rate is superior to MFT and EMFT even under the non-stationary background noise.

In Section II, we introduce the conventional EMFT and AMFT including the bottom-up score-calculation algorithm. Since AMFT still requires high computational complexity, to further reduce it, we propose a top-down score-calculation algorithm in Section III. Using the proposed method, we can also increase the recognition rate. Section IV and V presents the experimental results and the conclusion, respectively.

II. EMFT AND AMFT

In EMFT and AMFT, K-dimensional Decorrelated Filter Bank (DFB) is used as a feature vector $X = \{x_1, x_2, \cdots, x_K\}$ [11-15]. The likelihood score for EMFT can be calculated as [11-14]

$$p(X \mid \lambda_S) = \prod_{k=1}^{K} \sum_{m=1}^{M} w_{S,m} N(x_k; \mu_{S,m,k}, \sigma^2_{S,m,k}) \tag{1}$$

where λ_S, M, $w_{S,m}$, and $N(x_k; \mu_{S,m,k}, \sigma^2_{S,m,k})$ are the model of the speaker S, the number of mixtures, an m-th mixture weighting factor, and a Gaussian function for the input feature x_k, mean $\mu_{S,m,k}$, and variance $\sigma^2_{S,m,k}$, respectively. We note that $\lambda_S = \{w_{S,m}, \mu_{S,m,k}, \sigma^2_{S,m,k}\}$.

However, considering the dependency between vector components, we can calculate the likelihood score as given by

$$p(X \mid \lambda_S) = \sum_{m=1}^{M} w_{S,m} \prod_{k=1}^{K} N(x_k; \mu_{S,m,k}, \sigma^2_{S,m,k}). \tag{2}$$

In AMFT, the marginal likelihood score for a subset, $X_{subset} \subset X$, is calculated by

$$p(X_{subset} \mid \lambda_S) = \sum_{m=1}^{M} w_{S,m} \prod_{x_k \in X_{subset}} N(x_k; \mu_{S,m,k}, \sigma^2_{S,m,k}). \tag{3}$$

Because the length of each subset can be different, the likelihood score in (3) should be normalized before comparison as follows:

$$p(\lambda_S \mid X_{subset}) = \frac{p(X_{subset} \mid \lambda_S)\, p(\lambda_S)}{\sum_{S'} p(X_{subset} \mid \lambda_{S'})\, p(\lambda_{S'})} \tag{4}$$

where $p(\lambda_{S'})$ is a prior probability for a speaker S'. From the equal prior assumption of $p(\lambda_{S'})$ (4) can be interpreted as a normalized version of (3). Thus, using (4), we can calculate the maximum likelihood score of X for a given speaker model $\lambda_{S'}$ as given by

$$p(X \mid \lambda_S) \propto \max_{X_{subset} \subset X} p(\lambda_S \mid X_{subset}). \tag{5}$$

As the dimensionality of an input feature vector, K, increases, the complexity in calculating (3) and (4) increases since the possible number of subset candidates is $\sum_{n=1}^{K} {}_K C_n$. Thus, instead of $p(X_{subset} \mid \lambda_S)$ in (3),

$$p(X_N \mid \lambda_S) = \sum_{X_{subset} \subset X_N} p(X_{subset} \mid \lambda_S) \tag{6}$$

is calculated where X_N is a collection of all possible X_{subset} whose dimensionality is N ($1 \le N \le K$). Then, the maximum likelihood score in (5) can be replaced by

$$p(X \mid \lambda_S) \propto \max_{1 \le N \le K} p(\lambda_S \mid X_N) \tag{7}$$

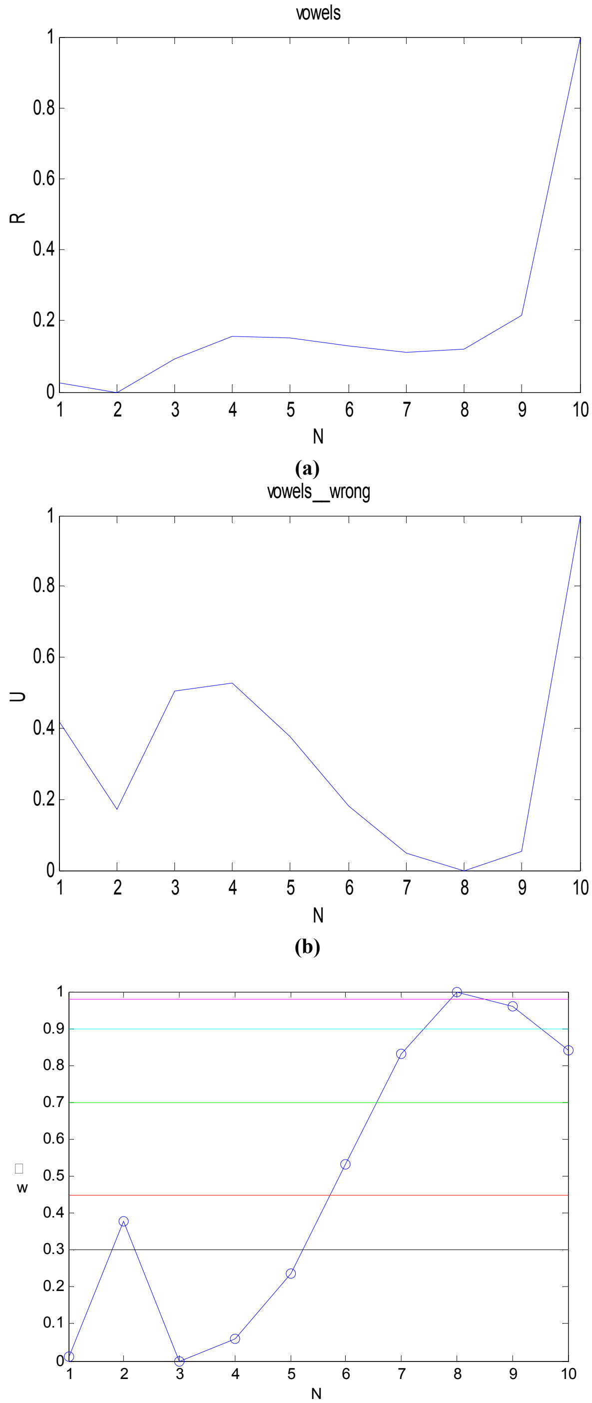
where

$$p(\lambda_S \mid X_N) = \frac{p(X_N \mid \lambda_S)\, p(\lambda_S)}{\sum_{S'} p(X_N \mid \lambda_{S'})\, p(\lambda_{S'})}. \tag{8}$$

In AMFT, we also proposed a fast score-calculation algorithm [15]. As shown in Table I, $p(X_N \mid \lambda_S)$ in (6) is calculated by $p(X_{K,N} \mid \lambda_S)$ in a recursive way. Compared with the computational complexity of $O(MK2^K)$ in EMFT, AMFT requires $O(MK + \frac{K(K+1)}{2})$, which gives a significant reduction in complexity for a higher value of K.

III. PROPOSED METHOD

AMFT provides superior performance in the recognition rate to EMFT, which yields much better performance than the conventional GMM-based system. AMFT also has functionality to reduce computational complexity by introducing a bottom-up score-calculation algorithm as shown in Table I [15]. In this section, we propose a weighting function considering the number of reliable components. By adjusting the threshold in the weighting function, we find the proper N not only to increase the recognition rate but to decrease the computational complexity. To further reduce the complexity, we also propose a top-down score-calculation algorithm.



(c)

Fig. 2. R_N, U_N, **and the corresponding weighting** $\overline{w}_N$.

First, we apply the weighting function, $\overline{w}_N$, depending on the number of reliable components, N, as given by

$$p(X \mid \lambda_S) \propto \max_{1 \le N \le K} \overline{w}_N \, p(\lambda_S \mid X_N) \tag{9}$$

where $\overline{w}_N$ is selected based on the following off-line learning method. First, the GMM parameters for each speaker are estimated based on the expectation-maximization (EM) algorithm. Second, the maximum likelihood score for each frame of the input data is calculated to discover the estimated number of reliable components for each frame as given by

$$N^* = \arg\max_{\substack{1 \le N \le K \\ 1 \le \lambda_S \le \Omega}} p(\lambda_S \mid X_N) \tag{10}$$

where Ω is the number of all speaker models. The corresponding N^* is collected to count the number of successes and failures according to the recognition result, which is decided not on a frame-by-frame basis, but by observing a whole frame of test vectors. For each N, the number of successes and failures are converted to the corresponding histograms, R_N and U_N, using min-max normalization. The final weighting factor, $\overline{w}_N$, is estimated using $w_N = R_N + 1/U_N$ after min-max normalization of it. Because the recognition rate is adversely affected by the increase in U_N, we use the reciprocal form of U_N to calculate w_N.

To calculate R_N and U_N, TIMIT database recorded from 630 talkers was used. For each talker, 8 and the remaining 2 sentences were used for training and test, respectively. From the training data set, 5 and the remaining 3 sentences of vowel sections were used to build a GMM model and to count N^*, respectively.

Fig. 2 (a), (b), and (c) illustrate R_N, U_N, and the corresponding $\overline{w}_N$, respectively. Since most of test vectors are classified successfully, U_N is measured with insufficient number of frames compared with R_N Thus, U_N and the corresponding $\overline{w}_N$ have abrupt artifact for N=2.

Not only the recognition rate, but the computational complexity is important in designing the speaker recognition system. To reduce the complexity, we build a modified mask as given by

$$\hat{w}_N = \begin{cases} \overline{w}_N, & \overline{w}_N > \delta_{TH} \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

where δ_{TH} is the pre-determined threshold. Using (11), we select a subset of X_N that gives more influence to the recognition rate. As shown in Fig. 2 (c), the number of reliable components, N, is selected as {2, 6, 7, 8, 9, 10}, {6, 7, 8, 9, 10}, {7, 8, 9, 10}, {8, 9}, and {8}. We add the case of {9} for the complexity issue.

For the case of 8 or 9 of N, it is not required to calculate the max operation in (7) and the denominator in (8). Thus, (7) and (8) can be rewritten as a simpler form of

$$p(X \mid \lambda_S) = p(X_N \mid \lambda_S). \tag{12}$$

TABLE I
FAST SCORE-CALCULATION ALGORITHMS: CONVENTIONAL BOTTOM-UP AND PROPOSED TOP-DOWN APPROACHES.

Bottom-Up Approach	Proposed Top-Down Approach	
AMFT [15]	***N*=9**	***N*=8**
for N= 1…K for n = 1…N if (N = 1 and n = 1) $p(X_{N,n} \mid \lambda_S) = p(x_N \mid \lambda_S)$ else if (n = 1) $p(X_{N,n} \mid \lambda_S) = p(X_{N-1,n} \mid \lambda_S) + p(x_N \mid \lambda_S)$ else if (n = N) $p(X_{N,n} \mid \lambda_S) = p(X_{N-1,n} \mid \lambda_S) * p(x_N \mid \lambda_S)$ else if (1 < n < N) $p(X_{N,n} \mid \lambda_S) = p(X_{N-1,n} \mid \lambda_S)$ $+ p(X_{N-1,n-1} \mid \lambda_S) * p(x_N \mid \lambda_S)$ end end end	temp= $p(x_1 \mid \lambda_S)$ for i=2…K temp*= $p(x_i \mid \lambda_S)$ end $p(X_9 \mid \lambda_S)$ = temp/ $p(x_1 \mid \lambda_S)$ for i=2…k $p(X_9 \mid \lambda_S)$ += temp/ $p(x_1 \mid \lambda_S)$ End	temp= $p(x_k \mid \lambda_S)$ for *i*=(K-1)…4 temp*= $p(x_i \mid \lambda_S)$ end $p(\lambda_S \mid X_8)$ =temp* $p(x_3 \mid \lambda_S)$ A=temp/ $p(x_k \mid \lambda_S)$ B=A/ $p(x_{k-1} \mid \lambda_S)$ for i=(K-1)…5 A += temp/ $p(x_i \mid \lambda_S)$ B += A/ $p(x_{i-1} \mid \lambda_S)$ end A+=temp/ $p(x_4 \mid \lambda_S)$ C=temp+A* $p(x_3 \mid \lambda_S)$ $p(X_8 \mid \lambda_S)$ += $p(x_2 \mid \lambda_S)$ *C+ $p(x_1 \mid \lambda_S)$ *[C+ $p(x_2 \mid \lambda_S)$ *{A+ $p(x_3 \mid \lambda_S)$ *B}]

TABLE II
SPEAKER-IDENTIFICATION ERROR RATE (%) FOR FULL-BAND SYSTEM [16], AMFT [15], AND AMFT USING A WEIGHTING FUNCTION.

Noise type	SNR (dB)	Full band	AMFT	AMFT $+\overline{w}_N$
Clean		1.51	1.03	1.03
Volvo	20	5.79	1.51	1.35
	15	19.60	2.38	2.22
	10	46.51	3.73	3.17
Babble	20	1.75	2.14	1.75
	15	4.44	5.48	4.92
	10	16.83	21.35	20.32
Monophonic ringtone	20	6.11	1.35	1.11
	15	20.00	1.98	1.59
	10	45.40	3.49	3.17
Polyphonic ringtone	20	4.60	3.25	3.02
	15	16.67	6.27	5.95
	10	50.48	24.44	23.57
Machinegun	20	32.22	5.48	5.08
	15	31.67	5.95	5.16
	10	32.14	6.59	5.79
F16	20	41.51	24.44	23.02
	15	57.78	46.98	47.06
	10	83.57	79.29	79.84
Average		27.29	13.01	12.59

TABLE III
SPEAKER-IDENTIFICATION ERROR RATE WITH A MODIFIED MASK IN (11).

Noise type	SNR (dB)	the number of reliable components, *N*						
		1~10	2,6~10	6~10	7~10	8~9	8	9
Clean		1.03	1.03	0.79	0.63	0.63	0.56	0.48
Volvo	20	1.35	1.27	1.19	1.03	0.56	0.63	0.71
	15	2.22	2.22	1.67	1.35	0.79	0.87	1.03
	10	3.17	3.10	2.46	2.38	1.11	1.11	1.27
Babble	20	1.75	1.75	1.35	1.27	1.27	1.35	0.95
	15	4.92	4.84	3.41	3.49	3.33	3.41	2.94
	10	20.32	20.16	17.62	15.95	14.05	15.24	12.62
Monophonic ringtone	20	1.11	1.11	1.03	1.03	0.95	0.87	1.35
	15	1.59	1.51	1.43	1.35	1.19	1.03	2.46
	10	3.17	3.17	2.70	2.61	2.46	2.38	8.10
Polyphonic ringtone	20	3.02	2.78	2.38	2.22	1.75	1.83	1.67
	15	5.95	5.79	5.32	4.60	4.92	5.00	5.40
	10	23.57	23.41	22.22	21.43	20.71	20.73	25.56
Machinegun	20	5.08	5.00	3.81	3.57	2.38	2.46	3.10
	15	5.16	5.08	3.97	3.89	2.46	2.46	3.10
	10	5.79	5.79	4.76	4.37	2.78	2.70	3.17
F16	20	23.02	22.94	20.00	18.41	16.11	16.11	16.11
	15	47.06	47.14	46.75	45.40	40.79	41.51	41.35
	10	79.84	79.84	79.13	78.25	76.43	76.19	76.75
Average		12.59	12.52	11.68	11.22	10.25	10.34	10.95

In this case, the relative weighting as a function of *N* in (11) is not needed to be applied to (12) because only the single value of *N* is chosen. We call this special case the Hard-mask MFT (HMFT). Thus, AMFT and HMFT produce $\sum_{n=1}^{K} {}_K C_n$ and ${}_K C_N$ possible combinations, respectively. All possible combinations are used to calculate the maximum likelihood score for each frame.

In AMFT, the fast score-calculation algorithm in (6) is a bottom-up approach since it calculates $p(X_N \mid \lambda_S)$ from *N*=1 to *N*=10. However, in HMFT with *N*=8 and 9, $p(X_N \mid \lambda_S)$ is not required to be calculated for other *N*'s. For example, if *N*=8, it is simpler to calculate $p(X_8 \mid \lambda_S)$ not from *N*=1, but from *N*=10 as given by

$$p(X_8 \mid \lambda_S) = \frac{p_{1,2,\cdots,10}}{p_{1,2}} + \frac{p_{1,2,\cdots,10}}{p_{1,3}} + \cdots + \frac{p_{1,2,\cdots,10}}{p_{9,10}} \quad (13)$$

where

$$p_{\alpha,\alpha+1,\cdots,\beta} = \prod_{\alpha}^{\beta} p(x_i \mid \lambda_S). \quad (14)$$

We call this fast method the top-down approach of HMFT.

TABLE IV
COMPUTATIONAL COMPLEXITY OF THE SCORE-CALCULATION ALGORITHMS IN AMFT AND THE PROPOSED HMFT WITH N=8 AND 9.

		AMFT	HMFT-8 (N=8)		HMFT-9 (N=9)	
Fast score-calculation algorithm		**Bottom-up**	**Bottom-up**	**Top-down**	**Bottom-up**	**Top-down**
Fast score-calculation algorithm Complexity	Addition	45	44	16	45	9
	Multiplication	45	43	12	44	9
	Division	0	0	13	0	10
	Conditioning	217	217	0	217	0
	Total	307	304	41	306	28
Identification Error Rate (%)		13.01	10.34		10.95	

To further reduce the computational complexity of (13), we also propose a score-calculation algorithm for $p(X_8 \mid \lambda_S)$ where the common parts in calculation are more efficiently utilized as given by

$$p(X_8 \mid \lambda_S) = p_{3,4,\cdots,10} + p_2 C + p_1\{C + p_2(A + p_3 B)\} \qquad (15)$$

where $A = \sum_{i=4}^{10} T_i$, $B = \sum_{i=4}^{9} T_{i+1} / p_i$, $C = p_{4,5,\cdots,10} + p_3 A$, and $T_i = p_{4,5,\cdots,10} / p_i$. Table I summarizes the fast score-calculation algorithms such as the conventional bottom-up approach and the proposed top-down approach.

IV. EXPERIMENTAL RESULTS

Speaker recognition systems were evaluated with TIMIT database. We used 8 and 2 sentences of each speaker for training and evaluation, respectively. The 32-mixture GMM for each speaker was estimated using an expectation maximization algorithm. Test sentences were additively corrupted by Volvo, babble, mono ringtone, polyphonic ringtone, machinegun, F16 noise in the NOISEX database with a 10, 15, and 20dB signal-to-noise ratio (SNR).

DFB was selected as a feature vector instead of mel-frequency cepstral coefficient (MFCC) because of its better performance in speaker recognition [14]. Using 21 channel mel-scale filter banks, we calculated 21 spectral amplitudes in a log domain for each filter bank as $\{c_1, c_2, \ldots, c_{21}\}$. By applying a high-pass filter, $H(z) = 1 - z^{-1}$, 20-dimensional intra-frame difference features were calculated by $\{d_1, d_2, \ldots, d_{20}\} = \{c_2 - c_1, c_3 - c_2, \ldots, c_{21} - c_{20}\}$. Then, the DFB sub-band feature x_k was composed of $\{x_1, \ldots, x_{10}\} = \{(d_1, d_2), \ldots, (d_{19}, d_{20})\}$. DFB showed better performance than MFCC as in [14]. For the training and evaluation, we used the phoneme information of TIMIT instead of designing voice activity detection.

Table II represents speaker-identification error rate for the full-band system [16], AMFT [15], and AMFT using a weighting function $\overline{w}_N$. AMFT produces a lower error rate than the full-band system. AMFT with $\overline{w}_N$ gives a lower error rate than AMFT without using $\overline{w}_N$ except for the F16 noise, but the difference is not distinct. In average, the full-band system, AMFT, and AMFT with $\overline{w}_N$ gives error rates of 27.29%, 13.01%, and 12.59%, respectively.

Table III represents the identification error rate with the modified mask in (11). Based on the pre-determined threshold δ_{TH}, the number of reliable components is selected as shown in Table III. As we increase δ_{TH}, the average error rate can be decreased and, at the same time, the computational complexity is also decreased. In terms of the error rate, {8}, {9}, and {8, 9} produce 10.34%, 10.95%, and 10.25%, respectively, while AMFT gives 13.01% on average. Although {8, 9} shows the best error rate, it requires additional complexity for the normalization step in (8). On the other hand, {8} gives almost the same performance as {8,9} without a normalization step as in (12). We call the cases of {8} and {9} HMFT-8 and HMFT-9, respectively.

Table IV shows the computational complexity of AMFT and the proposed HMFT. The number of arithmetic operations including addition, multiplication, and division, and conditional statements for each fast score-calculation algorithm were calculated. The proposed top-down approach gives much lower addition, multiplication, and conditional operations than the conventional bottom-up approach. The conditional operation is not required because N is fixed during the process. However, 13 and 10 division operations are required for HMFT-8 and HMFT-9, respectively. If we assume all the operations have the same weight, bottom-up AMFT, top-down HMFT-8, and top-down HMFT-9 require 307, 41, and 28 operations for each frame, respectively. Compared with AMFT, HMFT-8 and HMFT-9 produce much lower computational complexity as well as a significantly better identification error rate.

V. CONCLUSION

For the robustness of the speaker recognition system under background noise conditions, we proposed a hard-mask based missing feature theory (HMFT). It determines the optimal number of reliable components for the score calculation algorithm. Compared with the conventional AMFT where all the possible combinations are used to calculate the maximum likelihood score, HMFT-8 selects only 8 elements out of 10 spectral elements in a DFB feature vector. We note that the number of combinations are reduced from 1023 in AMFT to 45 in HMFT. To reduce the computational complexity, we proposed the top-down approach in score calculation. Compared with the AMFT, the proposed HMFT-8 gives a lower identification error rate by 2.67% with 7.49 times lower complexity. Thus, the proposed HMFT-8 is a practical alternative to the conventional AMFT-based speaker recognition system. Future work will aim to increase the performance by combining a speech enhancement algorithm.

REFERENCES

[1] M. Ji, S. Kim, H. Kim, and H. Yoon, "Text-independent speaker identification using soft channel selection in home robot environments," *IEEE Trans. Consumer Electron*., vol. 54, no. 1, pp. 140-144, 2008.

[2] H. Lee, S. Chang, D. Yook, and Y. Kim, "A voice trigger system using keyword and speaker recognition for mobile devices," *IEEE Trans. Consumer Electron*., vol. 55, no. 4, pp. 2377-2384, 2009.

[3] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustics, Speech, Signal Process.*, vol. 27, no. 2, pp. 113-120, 1979.

[4] J. Lim and A. V. Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. Acoustics, Speech, Signal Process.*, vol. ASSP-26, no. 3, pp. 197-210, 1978.

[5] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Acoustics, Speech, Signal Process.*, vol. 9, no. 5, pp. 504-512, 2001.

[6] H. Hirsch and C. Ehrlicher, "Noise estimation techniques for robust speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, pp. 153-156, 1995.

[7] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466-475, 2003.

[8] A. Vizinho, P. Green, M. Cooke, and L. Josifovski, "Missing data theory, spectral subtraction and signal-to-nose estimation for robust ASR: An integrated study," in *Proc. Eurospeech*, pp. 2407–2410, 1999.

[9] M. L. Seltzer, B. Raj, and R. M. Stern, "A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition," *Speech Commun.*, vol. 43, no. 4, pp. 379–393, 2004.

[10] D. Pullella, M. Kuhne, and R. Togneri, "Robust speaker identification using combined feature selection and missing data recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, pp. 4833–4836, 2008.

[11] J. Ming, P. Jancovic, and F. J. Smith, "Robust speech recognition using probabilistic union models," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 6, pp. 403-414, 2002.

[12] J. Ming and F. J. Smith, "A posterior union model for improved robust speech recognition in nonstationary noise," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, pp. 420–423, 2003.

[13] J. Ming, "Universal compensation - an approach to noisy speech recognition assuming no knowledge of noise," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 961–964, 2004.

[14] J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds, "Robust speaker recognition in noisy conditions," *IEEE Trans. Audio, Speech, Language Process*., vol. 15, no. 5, pp. 1711-1723, 2007.

[15] J. Jung, K. Kim, and M. Y. Kim, "Noise robust speaker identification based on the advanced missing feature theory," *Electronics Letters*, vol. 46, no. 14, pp. 1027-1029, 2010.

[16] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Tran. Speech Audio Process*., vol. 3, no. 1, pp. 72–83, 1995.

[17] H. C. Nadeu, J. Hernando, and M. Gorricho, "On the decorrelation of the filter-bank energies in speech recognition," in *Proc. Eurospeech*, pp. 1381–1384, 1995.

[18] H.-W. Yang, Y.-L. Liu, and D.-Z. Huang, "Speaker recognition based on weighted mel-cepstrum," in *Int. Conf. Commun. Information Technology*, pp. 200-203, 2009.

[19] X.-T. Luo, L.-X. Ji, and S.-M. Li, "Weighted distortion measure on standard deviation for VQ-based speaker identification," in *Int. Conf. e-Business Information System Security*, pp. 1-4, 2010.

BIOGRAPHIES

Shin-Cheol Lim is a Master Student at the Dept. of Information and Communication Engineering, Sejong University, Seoul, Korea. He received an B.Sc. degree in Information and Communication Engineering from Sejong University, Seoul, Korea, in 2009. His research interests include speech and speaker recognition, speech enhancement, and music information retrieval.

Sei-Jin Jang received BS and MS degrees in Electronics Engineering from Kyungpook National University, Daegu, South Korea, in 1995 and 1997, respectively. From 1997 to 2002, he worked as a Senior Research Staff at Daewoo Electronics, Seoul, Korea. He is currently a head of Next-Generation Sound Supporting Center of Korea Electronics Technology Institute. His research interests include A/V signal processing and music information retrieval.

Seok-Pil Lee received BS and MS degrees in Electrical Engineering from Yonsei University, Seoul, South Korea, in 1990 and 1992, respectively. In 1997, he earned a PhD degree in Electrical and Electronics Engineering also at Yonsei University. He is currently a head of Digital Media Research Center of Korea Electronics Technology Institute. His research interests include A/V signal processing and the convergence of digital broadcast and telecommunication.

Moo Young Kim (M'96, SM'10) is an Associate Professor at the Dept. of Information and Communication Engineering, Sejong University, Seoul, Korea. He received an M.Sc. degree in electrical engineering from Yonsei University, Seoul, Korea, in 1995, and a Ph.D. degree in electrical engineering from KTH (the Royal Institute of Technology), Stockholm, Sweden, in 2004, respectively. From 1995 to 2000, he worked as a Member of Research Staff of the Human Computer Interaction Laboratory at Samsung Advanced Institute of Technology, Kiheung, Korea. From 2005 to 2006, he was a Senior Research Engineer of the Dept. Multimedia Technologies at Ericsson Research, Stockholm, Sweden. His research interests include speech and audio coding based on information theory, biometrics including speaker recognition, speech enhancement, joint source and channel coding, and music information retrieval.