Text-Independent Speaker Identification using Soft Channel Selection in Home Robot Environments

Mikyong Ji, Sungtak Kim, Hoirin Kim, *Member*, IEEE, and Ho-Sub Yoon

Abstract — *With the aim of achieving the best possible speaker identification rate in a distant-talking environment, we developed a multiple microphone-based text-independent speaker identification system using soft channel selection. The system selects and combines the identification results based on the reliability of an individual channel result using a single perceptron. Thus, it allows for user-customized service with high identification accuracy in home robot environments. From the experimental results, it is shown that the proposed system is effective in a distant-talking environment, thereby providing a speech interface for a wide range of potential hands-free applications in a ubiquitous environment.* [1]

Index Terms — Human robot interaction, hypothesis combination, speaker identification, speaker recognition, multiple microphones.

I. Introduction

Speaker identification is the task of determining which enrolled speaker has provided a given utterance among a set of known users. This technique makes it possible to use a speaker's voice to control access to services such as voice dialing, data access services, and information retrieval services. This capability is effective for robot applications where multiple users share the same access privileges to some application, but where the individual speaker must be uniquely identified from a group in order to provide the user with a customized service depending on his/her preference. Therefore, speaker identification technology is expected to create various useful services that will make our daily lives more convenient.

The ultimate goal of speaker identification systems is to achieve the best possible identification performance at hand. Current state-of-the-art speaker identification systems have achieved high identification accuracy. And they are known to perform reasonably well when the speech signals are captured in noise-free environments using close-talking microphones worn near the speaker's mouth. However, even if one of the current technologies yield a best performance, its identification rate could be abruptly degraded due to a variety of causes such as the distance between the speaker and the microphone, the location of the microphone and/or noise, and the direction of the speaker in adverse distant-talking environments where the speaker is at a distance from the microphones.

To deal with such a problem, microphone array-based speaker identification technologies have been successfully applied to improve the identification rate through speech enhancement by combining the speech signals from the microphones in order to increase their SNR [1], [2]. However, an accurate estimation of the time delays between different speech signals is still not an easy task due to room reverberation, background noise, the non-stationary characteristics of the speech signal, etc. Among those causes, generally, room reverberation is considered to be the main problem for time delay estimation, but adverse noisy environments can also considerably decrease the performance of a time delay. Even though array processing technologies effectively improve the SNR of the resulting signal, these improvements are not directly translated into substantial gains in classification problems. It is assumed that the best array processing will result in the best performance. However, the speaker identification system does not interpret the waveform itself; rather the feature is extracted from the speech waveform. Whereas this is certainly appropriate if the speech signal is to be interpreted by a human listener, it may not necessarily be the right criteria if the signal is to be interpreted by the speaker identification system. A microphone array is any number of microphones operating in tandem. Typically, arrays are formed using numbers of closely spaced microphones. And there exists a restriction with a fixed physical relationship in space between the different individual microphone array elements. In addition, there has been another approach based on feature compensation for robust speaker identification in a multi-microphone environment [3].

The variety of causes that exist in a distant-talking environment can have different effects on an individual channel. Thus, speaker identification errors misclassified by different speech inputs are not always the same. This suggests that the composite output could potentially have a lower error rate than any of the individual outputs. Therefore, we propose a multiple microphone-based speaker identification method that merges the identification results obtained from multiple microphones using soft channel selection. No restrictions with respect to space between separate microphones are imposed. To recognize a user's identity exactly is very important in providing a user with a customized service in a robot environment, thereby enhancing the quality of human and robot interaction.

The remainder of this paper is organized as follows. In Section II, we review the conventional multiple microphone-based speaker identification methods based on the classical Gaussian mixture models [4], [5], and we describe a new multiple microphone-based speaker identification method in

[1] This work is a result of the URC project sponsored by the MIC of the Korean government.

Mikyong Ji, Sungtak Kim, and Hoirin Kim are with Information and Communications University (ICU), Daejeon, Korea (e-mail: lindaji@icu.ac.kr, stkim@icu.ac.kr, and hrkim@icu.ac.kr).

Ho-Sub Yoon is with Electronics and Telecommunications Research Institute, Daejeon, Korea (e-mail: yoonhs@etri.re.kr).

Manuscript received January 14, 2008

Section III. Section IV illustrates the experimental results. Finally, we draw our conclusions in Section V.

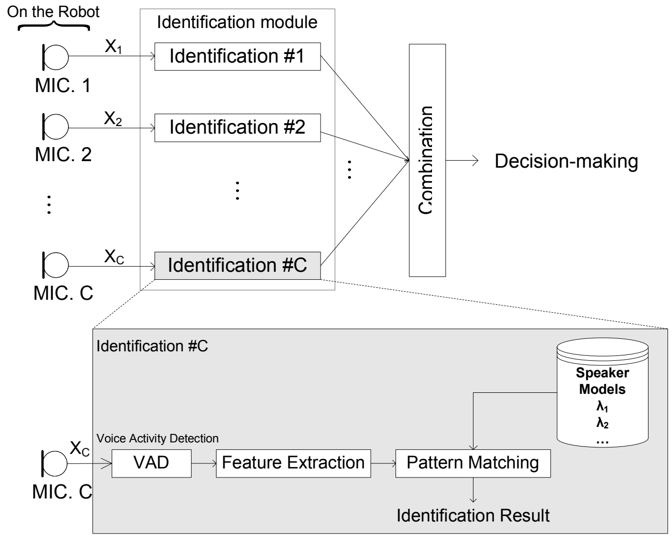


Fig. 1. Diagram of integrating the identification results in multiple microphone-based speaker identification.

II. Combining Speaker Identification Results

In multiple microphone-based speaker identification, the process of integrating the identification results obtained by multiple microphones is illustrated into a block diagram, as shown in Fig. 1. Given different speech inputs X_1, X_2, …, X_C simultaneously recorded through C multiple microphones, the speaker who provides given test utterances X_1, X_2, …, X_C among a set of known speakers $\mathcal{S} = \{1, 2, \ldots, S\}$ is generally identified by (1). Each enrolled speaker is modeled individually by Gaussian mixture model (GMM) $\lambda_1, \lambda_2, \ldots, \lambda_S$.

$$\hat{S} = \arg\max_{1\le k\le S} P(\lambda_k \mid X_1, X_2, \ldots, X_C). \tag{1}$$

Depending on the assumptions made, (1) can be rewritten as one of the following combination rules, CS (combination by sum), CM (combination by max), or CV (combination by majority vote) as in (2), (3) and (4), respectively [6].

1) Combination by sum (*CS*)

$$\hat{S} \cong \arg\max_{1\le k\le S} \sum_{c=1}^{C} P(\lambda_k \mid X_c). \tag{2}$$

2) Combination by max (*CM*)

$$\hat{S} \cong \arg\max_{1\le k\le S} \max_{1\le c\le C} P(\lambda_k \mid X_c). \tag{3}$$

3) Combination by majority vote (*CV*)

$$\hat{S} \cong \arg\max_{1\le k\le S} \sum_{c=1}^{C} \Delta_{kc}, \tag{4}$$

where Δ_{ks} is further defined by

$$\Delta_{kc} = \begin{cases} 1 & if\ P(\lambda_k \mid X_c) = \max_{1\le k'\le S} P(\lambda_{k'} \mid X_c) \\ 0 & otherwise \end{cases}. \tag{5}$$

III. Proposed Speaker Identification with Multiple Microphones

A. Frame's Entropy by Posterior Probabilities

In probability or information theory, the entropy $H(Y)$ of a discrete random variable $Y = \{y_1, y_2, \ldots, y_N\}$ introduced by Shannon is defined as:

$$H(Y) = H(P(y_1), \ldots, P(y_N)) = -\sum_{i=1}^{N} P(y_i) \log_b P(y_i), \tag{6}$$

where $P(y_i) = Pr(Y = y_i)$, $P(y_i) \ge 0$, and $\sum_i P(y_i) = 1$. If all the outcomes are equally likely to be ($P(y_i) = 1/N$), then the entropy should be maximal. For all N, therefore, it follows that

$$H(P(y_1), P(y_2), \ldots, P(y_N)) \le H(\frac{1}{N}, \frac{1}{N}, \ldots, \frac{1}{N}) = \log_b^N. \tag{7}$$

The entropy should be unchanged even if the outcomes y_i are re-ordered as

$$H(P(y_1), P(y_2), \ldots, P(y_N)) = H(P(y_2), P(y_1), \ldots, P(y_N)). \tag{8}$$

In this work, entropy is used to measure the degree of a frame's contribution to speaker identification. Let us assume that there exist S enrolled speakers, and the prior probabilities for all the speakers are equal ($P(\lambda_k) = 1/S$). Then, the posterior probability of speaker k at frame t is given as follows:

$$p_{tk} = P(\lambda_k \mid x_t) = \frac{p(x_t \mid \lambda_k)}{\sum_{k'=1}^{S} p(x_t \mid \lambda_{k'})}. \tag{9}$$

The posterior probability of speaker k in (9) represents the accuracy of speaker model λ_k producing an observation x_t. Thus, if the values of the posterior probabilities are similar, it is reasonably concluded that frame t does not really affect the identification. Reversely, if the posterior probability for one speaker is relatively higher than the others, it is reasonable to infer that frame t does affect the identification result. Then, the frame's entropy is defined by

$$H(P_t) = H(p_{t1}, p_{t2}, \ldots, p_{tS}) = -\sum_{k=1}^{S} p_{tk} \log_b p_{tk}, \tag{10}$$

where $p_{tk} = P(\lambda_k \mid x_t) \ge 0$, $\sum_{k=1}^{S} p_{tk} = 1$, and $P_t = \{p_{t1}, p_{t2}, \ldots, p_{tS}\}$ is a set of posterior probabilities for all enrolled speakers. The frame's entropy in (10) should be maximal if the posterior probabilities are equal for all speakers k ($p_{tk} = 1/S$) in the same manner as (7).

In order to examine if the frame's entropy is actually related to identification success, the entropy is computed frame by frame. The distribution of the frame's entropy is shown in Figs. 2 and 3, which are divided into two cases, respectively: when the true speaker is correctly identified, and when a false speaker is incorrectly identified, each under the entire observation sequence $X=\{x_1, x_2, \ldots, x_T\}$. In both cases, we categorize each frame into one of two groups depending on whether the best hypothesis is recognized on that frame or not. As shown in the figures, the frames on which the best hypothesis is recognized are considerably more distributed over a low entropy value in the case of identification success than in the case of identification failure. Accordingly, this implies that the frame's entropy affects identification success or failure.

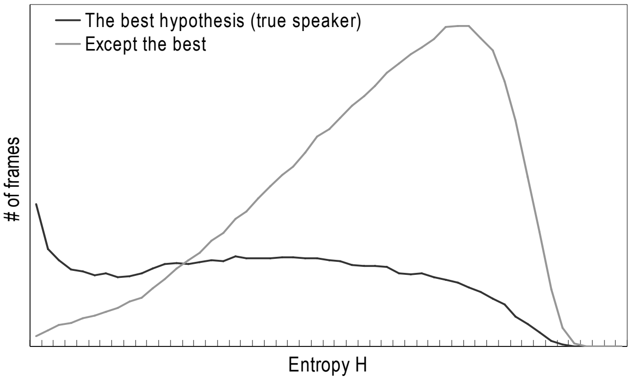


Fig. 2. Distribution of the frame's entropy where the true speaker is correctly identified (identification success).

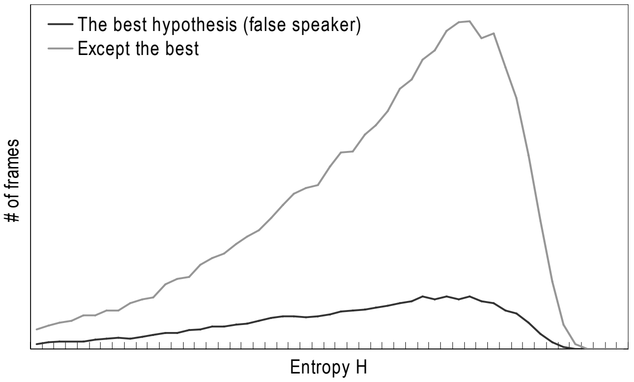


Fig. 3. Distribution of the frame's entropy where a false speaker is incorrectly identified (identification failure).

B. Measuring the Degree of Confidence in an Identification Result based on a Frame's Entropy

A procedure to measure the degree of confidence in an identification result is proposed based on a frame's entropy as previously introduced. First of all, the best hypothesis s^{h1} is determined by identifying the entire observation sequence $X = \{x_1, x_2, \ldots, x_T\}$. Second, the feature components of X are re-sorted with their entropies in ascending order. Then, the accumulated log-likelihoods of all enrolled speakers are computed in the order of the re-arranged observations frame by frame. At every frame, the speaker with the maximum accumulated log-likelihood $s(t)$ is compared with the best hypothesis s^{h1}. As a result, the confidence in identification result s^{h1} is determined by how long the accumulated log-likelihood of speaker s^{h1} is successively the maximum among those enrolled speakers right before speaker s^{h1} is finally confirmed as the output. Thus, we propose a feature called identified speaker's continuity (ISC), which represents the confidence in the identified speaker. The pseudo code for this is described in detail in Fig. 5.

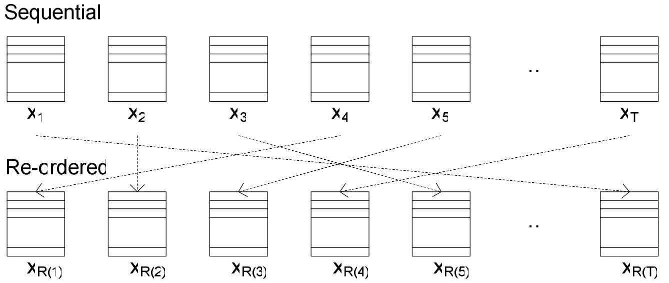
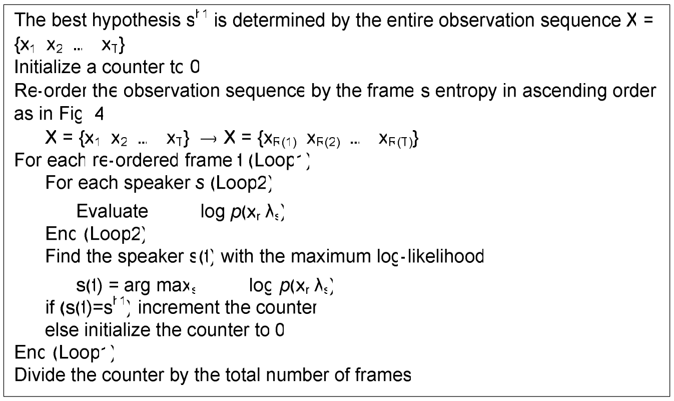


Fig. 4. Re-arranged observations with their entropies in ascending order.



```
The best hypothesis $s^{h1}$ is determined by the entire observation sequence $X = \{x_1, x_2, \ldots, x_T\}$
Initialize a counter to 0
Re-order the observation sequence by the frame's entropy in ascending order as in Fig. 4
    $X = \{x_1, x_2, \ldots, x_T\} \rightarrow X = \{x_{R(1)}, x_{R(2)}, \ldots, x_{R(T)}\}$
For each re-ordered frame t (Loop1)
    For each speaker s (Loop2)
        Evaluate    $\log p(x_t|\lambda_s)$
    End (Loop2)
    Find the speaker s(t) with the maximum log-likelihood
        $s(t) = \arg\max_s$    $\log p(x_t|\lambda_s)$
    if $(s(t)=s^{h1})$ increment the counter
    else initialize the counter to 0
End (Loop1)
Divide the counter by the total number of frames
```

Fig. 5. Pseudo code for identified speaker's continuity (ISC).

C. Soft Channel Selection by Perceptron

By selecting only reliable channels among multiple channels, and by then combining the identification results obtained from them, the accuracy of speaker identification can be improved upon even further. As shown in Fig. 6, a single perceptron learned by a gradient descent algorithm is used for soft channel selection, which gives weight to the individual identification result. The reliability of each channel is selected by the output of the two-input perceptron, and this reliability is applied as weight before integrating the identification results by all channels. The ISC and voting rate [7], both of which represent the degree of confidence in the identified result and range from 0 to 1, are used as the two inputs to the perceptron.

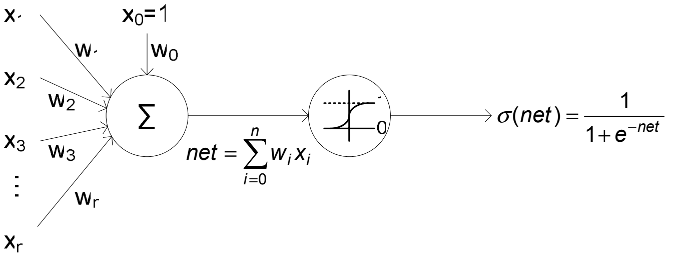

x_1
$x_0=1$
w_1
w_0
x_2
w_2
x_3
w_3
⋮
x_r
w_r
Σ
$net = \sum_{i=0}^{n} w_i x_i$
0
$\sigma(net) = \frac{1}{1+e^{-net}}$

Fig. 6. A single perceptron used for soft channel selection.

The reliability selected by the perceptron is used as weight w_c to the identified result obtained by identifying each channel input X_c. In the end, the conventional combination rules after the application of soft channel selection are modified as

1) CS after the application of soft channel selection

$$\hat{S} = \arg\max_{1 \le k \le S} \sum_{c=1}^{C} w_c P(\lambda_k \mid X_c). \tag{11}$$

2) CM after the application of soft channel selection

$$\hat{S} = \arg\max_{1 \le k \le S} \max_{1 \le c \le C} w_c P(\lambda_k \mid X_c). \tag{12}$$

3) CS after the application of soft channel selection

$$\hat{S} = \arg\max_{1 \le k \le S} \sum_{c=1}^{C} w_c \Delta_{kc}, \tag{13}$$

where Δ_{kc} is further defined by

$$\Delta_{kc} = \begin{cases} 1 & if\ P(\lambda_k \mid X_c) = \max\limits_{1 \le k \le S} P(\lambda_k \mid X_c) \\ 0 & otherwise \end{cases}. \tag{14}$$

IV. Experiment

A. Experimental Setup

In order to achieve the best possible speaker identification performance in distant-talking environments, a proposed multiple microphone-based speaker identification is evaluated using a database recorded by 30 speakers (23 males and 7 females) in the same environment as shown in Fig. 7. Nearly 60 conversational sentences per speaker, with short lengths of about one to two seconds each, were recorded in a quiet environment. Then, each sentence was re-recorded again into multiple microphones on a robot simultaneously by playing the original recording back on a loudspeaker placed at center (0°) or diagonal (45°) with distances of 1m, 2m, 3m, or 5m facing a robot (mock-up) in a home environment, as shown in Fig. 7. Among the recordings, 30 different sentences per speaker, each of which was recorded only at center and diagonal with a 1m distance by eight microphones on the robot simultaneously, were used to train an enrolled speaker model λ_k, whereas the rest of them were used for performance evaluation. Eight low-cost omni-directional microphones distributed on the robot were employed to collect the database, with the speech signals sampled at 16 kHz, and 12 Mel-frequency cepstral coefficients and their corresponding delta coefficients were used as the features. Each enrolled speaker was modeled by an 80 component GMM using the expectation-maximization algorithm [8]. Every utterance was pre-emphasized with a factor of 0.97, and a 20 ms Hamming window was applied with 10 ms overlapping.

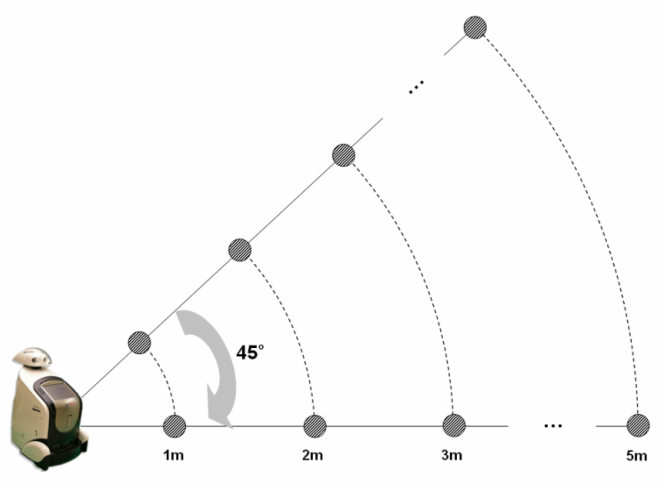


Fig. 7. Distant-talking multi-microphone environment.

B. Experimental Results

The accuracy of speaker identification is generally measured by the identification rate. For performance comparison, the conventional combination rules illustrated in Section II are employed to integrate the identification results obtained by separate microphones distributed on the robot in the same environment, as shown in Fig. 7. First of all, to examine if ISC actually presents confidence in the identification result, its distribution in both identification success and identification failure is shown in Fig. 8. The ISC values of the utterances are widely distributed throughout a range from 0 to1 in the case of identification failure. In the case of identification success, however, the distribution of those values is localized close to 1. This figure indicates that the value of ISC reflects the degree of confidence in the identification result.

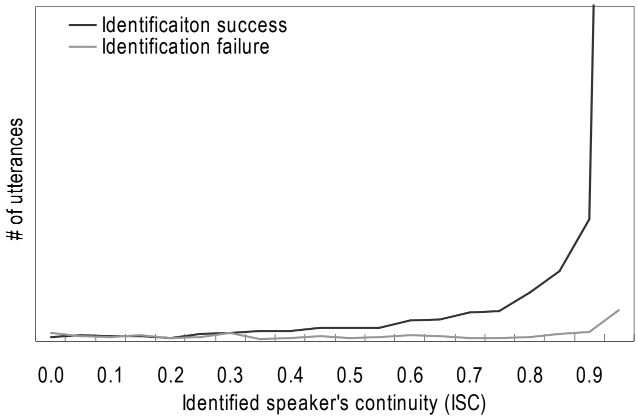


Fig. 8. Distribution of confidence in the identification results of both identification success and identification failure.

For the purpose of evaluating a performance improvement by soft channel selection, we compared the identification rates before and after the application of soft channel selection to those conventional combination rules as shown in Fig. 9. The experimental results show the identification rate per speaker's location. When the speakers are located near the microphones, the performances before and after the application of soft channel selection are comparable as shown in the figure. However, as the speakers distance themselves from the microphones, the new multiple microphone-based speaker identification method improves gradually. And, the relative improvement on CM after the application of soft channel selection is greater than the other rules. From this result, it is inferred that the weight selected using a single perceptron reflects the reliability in each identification result well.

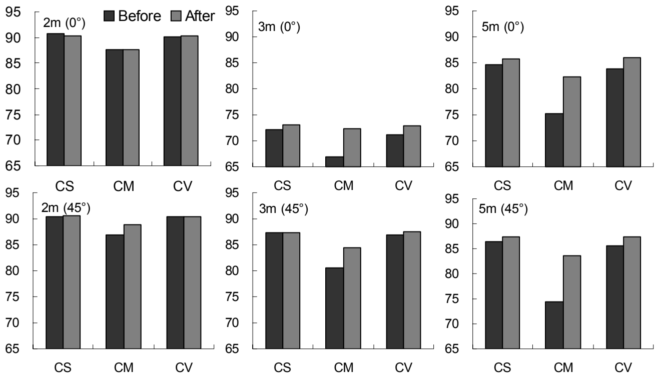


Fig. 9. Identification results before and after the application of soft channel selection to the conventional combination rules.

V. Conclusion

In this paper, we proposed a multiple microphone-based speaker identification method, which can effectively improve the identification rate by integrating the identification results using soft channel selection with a single perceptron in a distant-talking environment. The experimental results confirm that the proposed speaker identification method improves the identification performance even more as the speaker is at a distant from the microphones. We suggest that the proposed method can be used not only to accomplish the performance improvement of distant-talking speaker identification but also to provide a speech interface for a wide range of potential hands-free applications in robot environments, thereby significantly enhancing the quality of human and robot interaction. Also, the proposed speaker identification using soft channel selection is expected to be useful for performance improvement in various ubiquitous environments using multiple spatially-distributed microphones.

References

[1] Q. Lin, E Jan, and J. Flanagan,, “Microphone arrays and speaker identification,” *IEEE Trans. Speech, Audio Processing*, vol. 2, no. 4, pp. 622-629, Oct. 1994.

[2] I. A. McCowan, J. Pelecanos, and S. Sridharan, “Robust speaker recognition using microphone arrays,” in *Proc. Speaker Odyssey: The Speaker Recognition Workshop* (*Odyssey*2001), Crete, Greece, 2001, pp. 101-106.

[3] Q. Jin, Y. Pan, and T. Schultz, “Far-field speaker recognition,” in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing* (*ICASSP*2006), Toulouse, France, 2006, pp. 937-940.

[4] D. A. Reynolds, “Speaker identification and verification using Gaussian mixture speaker models,” *Speech Commu.*, vol. 17, pp. 91-108. 1995.

[5] D. A. Reynolds, and R. C. Rose, “Robust speaker identification using Gaussian mixture speaker models,” *IEEE Trans. Speech, Audio Processing*, vol. 3, no. 1, pp. 72-83. 1995.

[6] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, “On combining classifiers,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226-239. 1998.

[7] B. Narayanaswamy and R. Gangadharaiah, “Extracting additional information from Gaussian mixture model probabilities for improved text-independent speaker identification,” in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing* (ICASSP2005), Philadelphia, USA, 2005, vol. 1, pp. 621-624.

[8] A. Dempster, N. Laird, and D. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society*, vol. 39, pp. 1-38. 1977.

Mikyong Ji received her B.S. degree from the Dept. of Information Engineering, Hansung University, Seoul, Korea, and M.S. degree from the School of Engineering, Information and Communications University (ICU), Daejeon, Korea in 2000 and 2002, respectively. She is currently a Ph. D. candidate student in ICU. Her research interests include speech recognition, speaker recognition and Bayesian belief network.

Sungtak Kim received his B.S. degree in electronics engineering from the Ulsan University and M.S. degree from the School of Engineering, Information and Communications University (ICU), Korea in 2000 and 2003, respectively. He is currently pursuing the Ph. D degree at ICU. His research interests include robust speech recognition and speaker recognition.

Hoirin Kim was born in Seoul, Korea in 1961. He received his M.S. and Ph.D. degrees from the Dept. of Electrical and Electronics Engineering, KAIST, Korea in 1987 and 1992, respectively. From Oct. 1987 to Dec. 1999, he was a senior researcher in the Spoken Language Processing Lab. at the Electronics and Telecommunications Research Institute (ETRI). From June 1994 to May 1995, he was on leave at 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 Jan. 2000, he has been an Associative Professor at Information and Communications University, Korea. His research interests include signal processing for speech & speaker recognition, audio indexing & retrieval, and spoken language processing.

Ho-Sub Yoon received his B.S. and M.S. degrees in computer science from Soongsil University, Seoul, Korea in 1989 and 1991 respectively. He received his Ph. D. degree in image processing from KAIST, Daejeon, Korea in 2003. He joined KIST/SERI in 1991 and transferred to ETRI in 1999. He is currently the leader of the Human Robot Interaction (HRI) team in the intelligence robotics group. His major interests include HRI, image processing, audio processing, and pattern recognition.