Speaker Recognition for Wire/Wireless Communication Systems

Mohamed Abdel Fattah, Fuji Ren, and Shingo Kuroiwa
Faculty of Engineering, The University of Tokushima, Japan

Abstract: *Recently data communication spread to the mobile wireless world. The complexity of medium and large speech & speaker recognition systems are beyond the memory and computational resources of the small portable devices. Moreover, the most common approach to speaker recognition today is the use of global Gaussian Mixture Models (GMM) which ignores knowledge of the underlying phonetic content of the speech, so it does not take advantage of all available information. In this paper we address the solution of these two problems by investigating the phoneme effect on speaker recognition system. We used YOHO database for speaker identification task. We found that some phonemes have strong effect on speaker identification. Segmenting the most effective phoneme for speaker recognition task from a speaker utterance and send this phoneme only through the wireless communication system will decrease the complexity of medium and speed up the authentication process though mobile communication system. We have applied different approaches on YOHO corpus, several of these approaches were able outperform previously published results on the speaker ID task. One of our approaches could achieve 0.7% error rate by using only an average segment of 4.45% of the testing utterance for recognition.*

Keywords: *Speaker recognition, speaker identification, speech recognition, wireless communications.*

Received August 24, 2004; accepted September 172004

1. Introduction

There are two ways to perform vocal authentication on telecommunication area: The processing can be performed either locally or remotely. Remote authentication may be appropriate for high security transactions over a telecommunications link. Implementation of security services in general for Universal Mobile Telecommunication System (UMTS), and mutual authentication of user and network, in particular, includes the authentication protocols of the network and also the authentication of a user to the network based on face and/or speech of the user [15]. But for embedded environments, such as cell phones, it is desirable to have small speaker models and low computational complexity to decrease use of memory and increase battery life. For server environments, scalability is highly desirable. More transactions per unit time and small speaker models help maximize throughput and decrease server cost [2].

Speaker recognition is the process of automatically recognizing who is speaking by using speaker-specific information included in speech signal. Speaker recognition can be classified into *identification* and *verification*. Speaker identification is the process of determining which registered speaker provides a given utterance. Speaker verification, on the other hand, is the process of accepting or rejecting the identity claim of a speaker. Speaker recognition methods can also be divided into *text-independent*, *text-dependent* and *text-prompted* methods. In a text-independent system, speaker models capture characteristics of somebody's speech which show up irrespective of what one is saying. In a text-dependent system, the recognition of the speaker's identity is based on his or her speaking one or more specific phrases, like passwords, card numbers, PIN codes, etc. In text-prompted systems the user is asked to repeat a phrase. In this paper we addressed text independent and text-prompted speaker identification task since it is very beneficial when dealing with robots. It is convenient that the robot identifies the person who talks to with any utterance.

The use of Gaussian Mixture Models (GMM's) for speaker identification was shown to provide superior performance compared with several existing techniques [3, 5, 7, 12]. For example, Reynolds D. could achieve error rates as low as 0.7% using YOHO corpus [13], while Pellom B. L. reported the same error rate with reduction of the time to identify a speaker by a factor of 140 [11]. All previous mentioned researches could not achieve 0.0% Error Rate (ER). Moreover these researches used all speaker utterance for speaker recognition task which increased the recognition time and increased the transmitted data in the case of wire/wireless communication systems.

Some other researchers used different techniques for speaker recognition. Genoudy, used neural-network acoustic models of a hybrid connectionist-HMM speech recognizer to adapt a speaker-independent network by performing a small amount of additional training using data from the target speaker, giving an acoustic model specifically tuned to that speaker [6].

Thyes *et al.* [14], used "eigenvoice" approach, in which client and test speaker models are confined to a low-dimensional linear subspace obtained previously from a different set of training data. He reported 5% ER for Eigenvoice dimension of 70 using YOHO database. Wan *et al.* [16], reported identification error rate of 4.5% using polynomial order of 10 for Support Vector Machines approach when applied on YOHO corpus [16]. Campbell, used Polynomial Classifiers for Text-Prompted Speaker Recognition. His best identification error rate was 0.38% using second order Polynomial Classifiers for YOHO database [15]. However the different techniques used for all previous mentioned researches, it is strongly required to achieve as low ER as possible and also decrease the required speaker utterance part for recognition.

André *et al.* [1], segmented speech to 5 classes. Unvoiced segment class in addition of 4 different classes based on rising and falling of energy and fundamental frequency (f0). Park *et al.* [10], segmented speech to eight phonetic classes and used several approaches for speaker identification task based on YOHO corpus. His best identification error rate was 0.25% when he used multiple classifiers (phonetically structured GMM + speaker adaptive) [10]. However André and Park segmented the speech, they did not take the advantage of the whole effect of all phonemes in it.

Although the goal of text independent speaker recognition has led to an increased focus on global speaker modeling, it is well known that some phones have better speaker distinguishing capabilities than others [4, 9]. For instance, in [4] vowels and nasals were found to be most discriminating phoneme groups. Global speaker modeling techniques like the GMM approach are not able to take optimal advantage of the acoustic differences of diverse phonetic events. No doubt that taking the advantage of speech segmentation is enhancing the identification error rate as well as decreasing the required speech segments for speaker identification task. This advantage was not taken into account for the traditional speaker recognition models.

In this paper we investigate the phoneme effect on speaker recognition task. Our targets are:

1. Decreasing the identification error rate.
2. Decreasing the required speech segment for speaker identification task to decrease the system complexity and speed up the speaker identification process.

In order to achieve the above targets, we conducted several experiments using different approaches. First we used the traditional GMM for speaker identification then we investigated the speaker phonemes effect on the speaker identification task using different techniques. Our results outperform previously published results on the speaker ID from the precision point of view as well as minimum speech segment required for identification process. Section 2 describes YOHO database, section 3 illustrates the proposed system, section 4 illustrates the proposed system implementation in detail, and section 5 presents conclusions and future work.

2. YOHO Database

The data consists of 138 speakers - 106 males and 32 females - recorded in a span of 3 months. To record the data, a high quality telephone handset was used. For each speaker, both training, also referred to as enrollment, and testing, or verification, sessions have been created. The enrollment sessions consist of four sessions each containing 24 utterances while the verification data has 10 sessions of 4 utterances each. Each speaker has the same training data set where testing data are different for each speaker. Each utterance consists of "combination lock" phrases which are each a set of three doublets of digits, for example "23-42-91" pronounced as (twenty three, forty two, ninety one). The sampling rate for the speech files is 8 kHz, and the sample coding is 12-bit linear (stored as 16-bit words). The total number of pronounced phoneme types in YOHO database is 18 phoneme types. Figure 1 illustrates the training data phoneme frequencies for each speaker. Figure 2 shows testing data phoneme average frequencies for each speaker utterance.

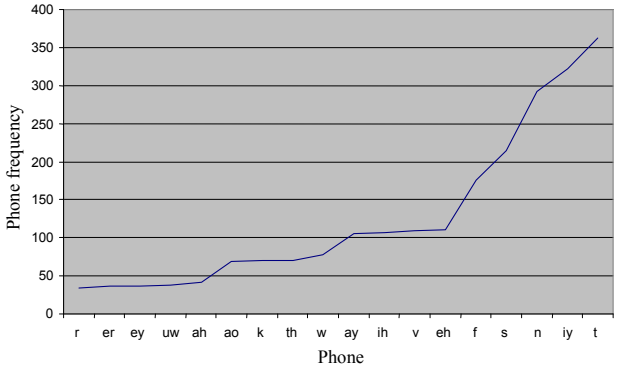

Figure 1. Training data phoneme frequencies for each speaker.

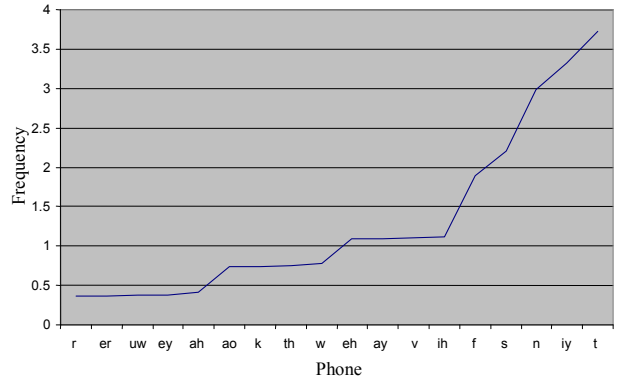

Figure 2. Testing data phoneme average frequencies for each speaker utterance.

3. The Proposed Approach

Our target is to decrease the speaker identification error rate and decrease the speech segment periods required for speaker identification process. To achieve these targets, we proposed the following approach:

- Feature extraction process for the training and testing data occurred as a pre-processing step.
- Use all speakers training data to construct Speaker Independent (SI HMM) phoneme model.
- Use the constructed SI HMM phoneme model to segment all training data to phoneme segments.
- Construct phoneme based speaker dependent model for each speaker.
- Segment testing data, and use each phoneme segment for speaker identification task.
- Select the most effective phonemes on speaker identification task, and use them for wire/wireless communication systems.

4. Implementation

The following sections describe the above proposed approach in detail. First we introduce the traditional Gaussian Mixture Model (GMM) approach which is the most common approach for speaker recognition task, and then we describe the proposed approach.

4.1. Traditional Gaussian Mixture Model Background

The most widespread paradigm for statistical acoustic modeling in speaker recognition involves the use of Gaussian mixture model. With this approach, the probability density function for a feature vector $\vec{z}$ is a weighted sum, or *mixture*, of K class-conditional Gaussian distributions. For a given speaker, s, the probability of observing $\vec{z}$ is given by:

$$p(\vec{z} \mid s) = \sum_{k=1}^{K} w_{s,k} N(\vec{z}; \vec{\mu}_{s,k}, \sum_{s,k}) \quad (1)$$

since $w_{s,k}$, $\vec{\mu}_{s,k}$, $\sum_{s,k}$, are the mixture weight, mean, and covariance matrix, respectively, for the i-th component, which has a Gaussian distribution given by:

$$N(\vec{z}; \vec{\mu}, \sum) = \frac{1}{\sqrt{(2\pi)^n \mid \sum \mid}} e^{-\frac{1}{2}(\vec{z}-\vec{\mu})' \sum^{-1} (\vec{z}-\vec{\mu})} \quad (2)$$

Where n is the size of $\vec{z}$. We used $\sum$ as diagonal covariance matrices to reduce computation. Given a set of training vectors of a certain speaker, an initial set of means is estimated using the *k*-means clustering. The mixture weights, means, and covariances are then iteratively trained using the Expectation Maximization (EM) algorithm.

4.2. Implementation Using Gaussian Mixture Model

Before making GMM speaker model, we have to choose a certain feature parameter to be extracted from the speaker speech. Linear Predictive Cepstral Coefficient (LPCC) was the first choice as a speaker feature parameter since LPC parameters in general have the following characteristics:

- For the quasi steady state voiced regions of speech, the all-pole model of LPC provides a good approximation to the vocal tract spectral envelop. During unvoiced and transient regions of speech, the LPC model is less effective than for voiced regions, but it still provides an acceptably useful model for speaker recognition purposes.
- The way in which LPC is applied to the analysis of speech signals leads to a reasonable source-vocal tract separation. As a result, a parsimonious representation of the vocal tract characteristics becomes possible.

A low (4-8) order LPC analysis captures the gross features of the envelope of speech spectrum. Speaker Information (SI) may be lost in such a representation [8]. So, 14 order LPCC + power were used as features to characterize the identity of speakers. The speech is first pre-emphasized (0.97); then, a sliding Hamming window with a length of 25ms and a shift of 10ms was positioned on the signal. Cepstral mean normalization also performed. Delta LPCC was used. So the feature vector size was 30.

After extracting the feature vectors from training and testing data, we have constructed speaker independent HMM phoneme model of 3 states, 16 mixtures for the 18 phones plus "sil" (silence) and "sp" (short pause) using all YOHO training data. The constructed HMM phoneme model was used to segment all training data. Then we represented each utterance as (sil + X + sp + X + sp + X + sil), since "X" is the voice segment, "sil" is the silence segment, and "sp" is the short pause segment. After that we used YOHO training data to construct a model for each speaker as follows:

- HMM of 3 states, 16 mixtures for "sil".
- HMM of 1 state (GMM), 16 mixtures for "sp".
- GMM of 64 mixtures for "X".

Then we combined "sil", "sp", and "X" in one model to represent each speaker. We could achieve identification error rate of 1.68% when we used maximum likelihood of segment "X" for testing data. Segmentation of each utterance as voice (X) and silence (sil, sp) enabled the system to use only "X" which is useful for speaker identification and discard silences which do not have any speaker information. However this GMM approach is not able to take optimal advantage of the acoustic differences of

diverse phonetic events. One of the disadvantages of the GMM's global model is that the acoustic variability of phonetic events in the test utterance is not taken into account when comparing different speakers. Although it has been shown that some phonetic classes have higher speaker distinguishing capabilities than others [4], much of this information is lost when all enrollment data is mapped to a single acoustic model. To overcome this problem, Park A., in [10] segmented speech to eight phonetic classes to take the contribution of each phonetic class for speaker recognition task. No doubt that segmentation of speech to phonemes instead of phonetic classes will decrease the identification error rate. But phone level speaker modeling techniques may exhibit poor performance due to insufficient training data at the phone level. However, speech corpus like YOHO database contains suitable amount of data for most of its phonemes (except "r" and "er" phonemes, see Figure 1) that are suitable to construct phoneme based GMM model.

4.3. Phonemes Effect on Speaker Identification

The probability density function for a feature vector $\vec{z}$ is a weighted sum, or *mixture*, of *K* class-conditional Gaussian distributions. For a given phone of a certain speaker, s_p, the probability of observing $\vec{z}$ is given by:

$$p(\vec{z} \mid s_p) = \sum_{k=1}^{K} w_{s_p,k} N(\vec{z}; \vec{\mu}_{s_p,k}, \sum\nolimits_{s_p,k}) \qquad (3)$$

Where $w_{s_p,k}$, $\vec{\mu}_{s_p,k}, \sum_{s_p,k}$ are the mixture weight, mean, and covariance matrix, respectively, for the i-th component, which has a Gaussian distribution given by equation (2).

Using equation (3), we have constructed the same speaker model as in section (4.2) but instead of combining all phonemes as a speech segment (X), we constructed phoneme model for each speaker. So each speaker utterance was represented as: (sil < ($phonemes) > (sp) < ($phonemes) > (sp) < ($phonemes) > sil). Since ($phonemes) is some phoneme combination of the 18 phonemes of YOHO database represented as:

$$\$phonemes = ah \mid ao \mid ay \mid eh \mid er \mid ey \mid f \mid ih \mid iy \mid k \mid n \mid r \mid s \mid t \mid th \mid uw \mid v \mid w;$$

Using this approach, we constructed speaker dependent model for each phoneme except 2 phonemes which are ("r" and "er") since the system failed to construct them for some speakers because the frequencies of these 2 phonemes are low as shown in Figure 1 and their training data duration times are low too. After that we used each phoneme model for each speaker to test each separate speaker phoneme for speaker identification task. Table 1 and Figure 3 illustrate the identification error rate for each speaker phoneme.

From Table 1 and Figure 3, ER depends on phoneme type and the frequency of the phoneme in the training data. ER is inversely proportional to the phoneme frequency of the training data since as the phoneme frequency increases the GMM phone based model accuracy increases too. ER is low in the case of vowels. A diphthong phone like "ay" has good identification results where as "ey" does not have good results since "ey" frequency is low. Mid vowel phoneme like "ao" has good identification result however it does not have high frequency, whereas "ah" does not have good result. A front vowel "eh" has strong effect on speaker recognition where "ih" and "iy" do not have that effect. The effect of voiced fricative consonant "v" is stronger than "th" because of the frequencies difference between them. Unvoiced fricative consonants ("f" and "s") and unvoiced stop consonants ("k" and "t") have weak effect on the identification task in general however the phoneme model of "t" has better results because it has the highest frequency over all phonemes.

The achieved results in Table 1 and Figure 3 are calculated for each separate phoneme of the utterance. It is common for speaker identification task to calculate the ER using the whole utterance. In the next section, we take the contribution of all phonemes of the same type for a certain utterance to calculate the identification error rate.

Table 1. Speaker identification results using separate phones.

Phoneme	**ay**	**ao**	**eh**	**ah**
ER	1%	1.6%	1.6%	5.9%
Phoneme	**n**	**t**	**ey**	**th**
ER	11.9%	14.7%	20.1%	20.6%
Phoneme	**v**	**ih**	**w**	**iy**
ER	6.1%	8.7%	10%	10.1%
Phoneme	**f**	**s**	**uw**	**k**
ER	21.7%	24.1%	25.9%	37.5%

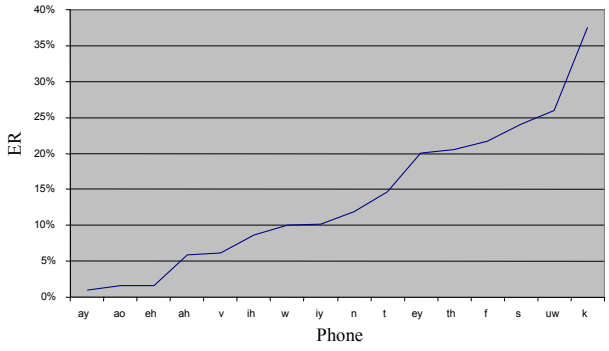

Figure 3. Speaker identification results using separate phones.

4.4. Identification Error Rate Using the Contribution of Utterance Phonemes of the Same Type

We conducted the above experiment, but we took the contribution of all phonemes of the same type in each utterance into account to calculate the ER. Table 2 and Figure 4 illustrates the identification error rate for each speaker phoneme when taking all utterance phonemes of the same type into account.

Table 2. Speaker identification results using contribution of utterance phones of the same type.

Phoneme	**ay**	**eh**	**ao**	**iy**
ER	0.7%	1.1%	1.4%	1.7%
Phoneme	**ih**	**w**	**s**	**f**
ER	7%	8.2%	14.4%	15%
Phoneme	**n**	**v**	**t**	**ah**
ER	4.2%	4.3%	4.8%	5.8%
Phoneme	**th**	**ey**	**uw**	**k**
ER	19.4%	20.1%	25.6%	36.7%

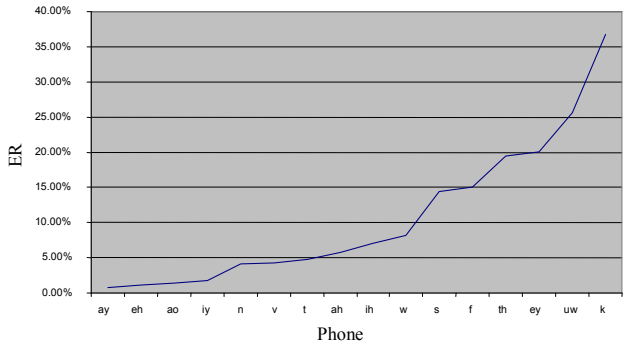

Figure 4. Speaker identification results using contribution of utterance phones of the same type.

It is clear that the ER improved in general. The effect of testing data phoneme frequencies on speaker identification task is very strong as shown in Table 2, and Figure 2. Phonemes "iy" and "t" have the highest frequencies over all phonemes of training and testing data so the ER values are lower than that of Table1. Phoneme "t" is an unvoiced stop consonant which does not have a strong effect on speaker identification task. However the ER associated with phoneme "t" is not bad because of its high frequency. It leads us to say that phoneme frequencies in testing data have the strongest effect on speaker recognition. The ER of phoneme "ay" is less than that of all phonemes because phoneme "ay" frequency is higher than most of the rest vowel frequencies in testing data.

For wire/wireless communications purposes, it is convenient to extract all "ay" phones from the speaker utterance at client terminal then send only these phones for speaker recognition purposes. Figure 5 shows an example of one speaker utterance containing 1 segment of phoneme "ay". This segment may be sent through wire/wireless system media. We have estimated the average duration time of all "ay" segments in the whole YOHO testing database and we found that: ("ay" segments duration time in all testing phrases)/ (total testing data time) = 4.45%. So we can achieve speaker identification error rate = 0.7% by sending only 4.45% of the speaker utterance through wire/wireless communication system.

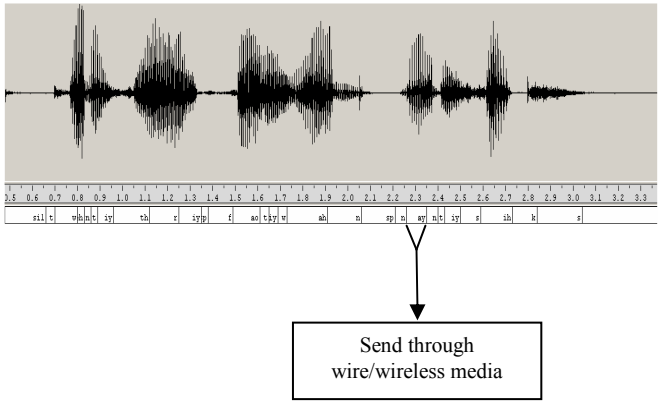

Figure 5. An example of one speaker utterance containing one segment of phoneme "ay".

Taking the contribution of all phonemes of the same type of each utterance into account improved ER.

We have conducted the same experiment taking all utterance phonemes contribution into account and we achieved 0.29% identification error rate.

All the above results were obtained using text-prompted approach in order to be able to segment speaker testing speech. It is strongly required to conduct text-independent experiment.

4.5. Text-Independent Identification Error Rate Using the Contribution of All Utterance Phonemes

Applying the approach of section (4.3) on text independent YOHO testing data, we could achieve ER = 0.94%. Table 3 shows a summary of the whole utterance speaker identification results.

Table 3. Summary of the whole utterance speaker identification results.

Approach	**Traditional GMM**	**Text-Prompted ER Using the Contribution of All Utterance Phonemes**	**Text-Independent ER Using the Contribution of All Utterance Phonemes**
ER	1.68%	0.29%	0.94%

5. Conclusions

In this paper we have investigated the phonemes effect on the speaker recognition task. We found that ER is inversely proportional to the frequency of the phoneme in the training and testing data and it depends on the phoneme type too. In general vowel phonemes contain a lot of speaker dependent features so they are very beneficial for speaker recognition task. Phoneme

segmentation of a speaker utterance before recognition is strongly enhancing speaker recognition results. For YOHO database we found that the phoneme "ay" gives 0.7% identification error rate. We can pick up the "ay" phone segments from speaker utterance and send it through the wire/wireless communication system for speaker identification process to decrease the system computations and increase the system speed. Taking the contributions of "ay" segments into account when calculating ER, we could achieve 0.7% identification error rate using 4.45% of the speaker utterance only.

In the future work we will investigate the rest phonemes effect on speaker recognition using other speech corpora.

Acknowledgment

This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (B), 14350204 and 14380166, in 2004, Hoso-Bunka Foundation (HBF) and International Communications Foundation (ICF).

References

[1] André G. and Hynek H., "Segmentation of Speech for Speaker and Language Recognition," *in Proceedings of Eurospeech Conference*, Geneva, 2003.

[2] Campbell M. and Broun C., "Text-Prompted Speaker Recognition with Polynomial Classifiers," *in Proceedings of the Speaker Recognition Workshop (ODYSSEY)*, pp. 183-188, 2001.

[3] Douglas A., "Automatic Speaker Recognition Using Gaussian Mixture Speaker Models," *The Lincoln Laboratory Journal*, vol. 8, no. 2, pp. 173-192, 1995.

[4] Eatock J. P. and Mason J. S., "A Quantitative Assessment of the Relative Speaker Discriminant Properties of Phonemes," *in Proceedings of ICASSP*, Adelaide, vol. 1, pp. 133-136, 1994.

[5] Fine S., Navratil J., and Gopinath R., "Enhancing GMM Scores Using SVM Hints," *in Proceedings of Eurospeech Conference*, Aalborg, pp. 1760-1767, 2001.

[6] Genoud D., Ellis D., and Morgan N., "Combined Speech and Speaker Recognition with Speaker-Adapted Connectionist Models," *in Proceedings of ASRU-99 Workshop*, Keystone CO, 1999.

[7] Kharroubi J., Petrovska D., and Chollet G., "Combining GMM's with Support Vector Machines for Text-Independent Speaker Verification," *in Proceedings of Eurospeech Conference*, Aalborg, pp. 1761-1764, 2001.

[8] Misra H., Ikbal S., and Yegnanarayana B., "Speaker-Specific Mapping for Text-Independent Speaker Recognition," *Speech Communication*, vol. 39, no. 3-4, pp. 301-310, 2003.

[9] Nolan F., *The Phonetic Bases of Speaker Recognition*, Cambridge University Press, Cambridge, 1983.

[10] Park A. and Timothy J., "ASR Dependent Techniques for Speaker Identification," *in Proceedings of the International Conference on Spoken Language Processing*, Denver, Colorado, 2002.

[11] Pellom B. L. and Hansen J. H. L., "An Efficient Scoring Algorithm for Gaussian Mixture Model Based Speaker Identification," *IEEE Signal Processing Letters*, vol. 5, no. 11, pp. 281-284, 1998.

[12] Reynolds D. and Rose R., "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," *IEEE Transactions on Speech Audio Processing*, vol. 3, no. 1, pp. 72-83, 1995.

[13] Reynolds D., "Speaker Identification and Verification Using Gaussian Mixture Speaker Models," *Speech Communication*, vol. 17, pp. 91-108, 1995.

[14] Thyes O., Kuhn R., Nguyen P., and Junqua J. C., "Speaker Identification and Verification Using Eigenvoices," *in Proceedings of the International Conference on Spoken Language Processing (ICSLP'2000)*, 2000.

[15] Timothy J., Weinstein E., Kabir R., and Park A., "Multi-Modal Face and Speaker Identification on a Handheld Device," *in Proceedings of Workshop on Multimodal User Authentication*, Santa Barbara, California, pp. 113-120, 2003.

[16] Wan V. and Campbell W. M., "Support Vector Machines for Speaker Verification and Identification," *Neural Networks for Signal Processing X*, pp. 775-784, 2000.

Mohamed Abdel Fattah received his BE degree in 1994 and ME degree in 2003 from Cairo University, Egypt. Currently he is a PhD student in Tokushima University, Japan. His research interests include natural language processing, information retrieval, speech recognition, and speaker recognition.

Fuji Ren received the BE degree in 1982 and ME degree in 1985 from Beijing University of Posts and Telecommunications, China and his PhD degree from Hokkaido University, Japan in 1991. His research interests include natural language processing, machine translation, fault simulation of digital systems, artificial intelligence,

language understanding and communication, multi-lingual multi-function multi-media intelligent system, automatic abstracting and information retrieval, super-function methodology, automatic derivation of programs from natural language descriptions, robust method for dialogue understanding, and sensitive information processing.

Shingo Kuroiwa received his BE, ME and PhD degrees in electrocommunications from The University of Electro Communications, Tokyo, Japan, in 1986, 1988, and 2000, respectively. From 1988 to 2001 he had been a researcher at the KDD R&D Laboratories. Since 2001, he has been with the Faculty of Engineering, Tokushima University, Tokushima, Japan. Currently, he is an associate professor. His research interests include speech recognition, speaker recognition, natural language processing, and information retrieval. He is a member of the Information Processing Society, the Acoustical Society of Japan, and the Institute of Electronics, Information and Communication Engineers.