BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **8**, No 4

Sofia • 2008

A Speaker Recognition Method Based on Personal Identification Voice and Trapezoidal Fuzzy Similarity

Nguyen T. H. Lien[1], *Fangyan Dong*[1], *Yoshinori Arai*[2], *Kaoru Hirota*[1], *Hiroyuki Sato*[3], *Teruhiko Hayashi*[3]

[1]*Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology, Japan*
E-mails: {nhlien, tou, hirota}@hrt.dis.titech.ac.jp

[2]*Department of Applied Computer Science, Tokyo Polytechnic University, Japan*
E-mail: arai@cs.t-kougei.ac.jp

[3]*Advanced Technology Systems Division, Soliton Systems K. K, Japan*
E-mails: {hiroyuki.sato teruhiko.hayashi}@soliton.co.jp

Abstract*: A text-dependent speaker recognition method is proposed using trapezoidal fuzzy similarity function to measure the similarity of voice features between a test user and the registered speaker who has nearest distance. The trapezoidal fuzzy similarity function is constructed based on three-time data recorded during enrolment process as personal identification voice (PIV) and statistical data of an individual recorded many times in a long time period to cover the intra-variation. A set of acoustic voice features is also introduced to present some general speaker and text dependent characteristics that are effective for modeling PIV, thus allowing to capture the inter- variation from one speaker to another. The experimental results on 24 speakers recorded in four different sessions show that, without false acceptation, the proposed system can decrease 30.05% of false rejection cases, compared to the traditional nearest neighbor approach. The focus of this work is on applications which require fast processing and few burdens for users.*

Keywords*: Speaker recognition, acoustic feature, fuzzy membership, nearest neighbor.*

1. Introduction

Speaker recognition, i. e., a technique to automatically recognize speakers from their voices, has various applications to access control to restricted services such as access to banking, database services, shopping or voice mail, and access to secure equipments or areas where mostly required a real-time processing with high security level and as fewer burdens for users as possible.

Focusing on the application of access control to restricted areas, this work limits the number of times that an user has to utter during the enrolment (registration) process to three times. During the enrolment process, speakers are asked to utter the same word or sentence three times in the same way as possible which is considered to be their personal identification voice (PIV). The PIV here includes characteristics of both word or sentence being spoken and the speaker's voice. The system only accepts registered users with their registered PIV.

Two fundamental issues regarding a text-dependent speaker recognition system are feature extraction and matching. The former involves finding features which can distinguish the Personal Identification Voice (PIV) of one person to another. The later is the process of recognizing users automatically using those features. For access control system, this involves identifying users and authenticating their identity. Furthermore, for the system that implies few burdens for users or few data for enrolment, the capture of feature variations of a person in different times is also a critical issue.

There are two popular approaches for feature extraction. The first one is to use traditional acoustic features such as formant frequencies, pitch, energy of the registered voice, which can present physical characteristics of the speaker and the utterance [14]. This approach is successfully used in several difficult tasks [5, 6]. The second approach is to use spectral representation of speech signal such as linear prediction coding [13], mel-frequency cepstrum coefficients [12], being this approach is suitable for matching based on statistical model. The feature extraction process in the second approach, however, has a large computation cost due to the repetition of the process on a large number of small segments of the speech, and thus it is still not appropriate for real-time applications. In this work, we follow the first approach and propose two new features besides the traditional ones. Those are Energy Increasing Frequency (EIF) and Voice Valley Number (VVN), as well as the introduction of Bounded Variation Quantity (BVQ) as an additional voice feature.

The techniques for matching speakers based on the extracted features can be categorized into two main approaches. The first approach is to model the speaker-dependent acoustic features and then compare these acoustic features in a test utterance using models such as Gaussian Mixture Model (GMM) [4], Vector Quantization (VQ) [15-17, 19]. Finally, the speaker recognition is performed using either maximum likelihood [4] or nearest neighbor distance [18, 15, 16]. The second approach is the use of discriminative neural networks (NN) such as multi-layer perceptrons [7], time-delay NN [8], and radial basis functions [9]. This approach has good speaker recognition performance, compared to others but its limitation is that the complete network has to be retrained when a new speaker is added to the system. The proposed system falls into the first approach that builds reference models for recognition. Different from GMM or VQ approaches which build a probabilistic model by repeatedly extracting features from many small segments of the speech, and thus are capacity and time-consuming, the proposed system uses PIV features of the whole speech signal and stores a vector of mean values of each feature as the reference model. This allows the system to easily adapt to new data and hence, it can be implemented on real-time applications.

The main issue of this approach is that the use of only the nearest neighbor distance is not enough to decide whether to accept or reject a test user. This issue could be solved by setting distance thresholds for the acceptance and rejection regions. It, however, actually does not improve performance considerably because if one changes the threshold to a lower false acceptance rate, then the false rejection rate increases. In this work, a trapezoidal fuzzy similarity (TFS) function is proposed to evaluate the similarity between the test user and the speaker to whom the test user has the nearest distance. The TFS is derived from the data that the user uttered during the enrolment and the statistical data of a user recorded in a long period to generalize and absorb the variation of feature value of one person, thus giving more accurate decisions.

Speakers are asked to utter the same word or sentence, which is considered to be their Personal Identification Voice (PIV), three times for the enrolment, and another utterance of the same PIV is used to test the proposed system. From the enrolment data, speakers' features are extracted and then vectors containing the mean values of each feature are stored as the speaker reference model. At the same time, the system stores the TFS function for each speaker to be used in the decision process. In addition, for the test utterance, the system extracts features of this uttered PIV, finds the speaker with the nearest distance and calculates the fuzzy similarity of the test user to that speaker to decide either to accept or to reject.

A database of 24 speakers' PIV recorded in four different sessions is used to evaluate the performance of the proposed system and that of the traditional nearest neighbor approach. The experiment is performed on the registered users to check false rejection rate and on unregistered users to check false acceptance rate. The experiment shows that with 0% of false acceptance, the proposed system can decrease 30.05% of false rejection cases, compared to the traditional nearest neighbor approach.

The proposed method is presented in II and the experimental results are shown in III. Discussion is presented in IV.

2. Speaker recognition based on personal identification voice and trapezoidal fuzzy similarity function

The speaker recognition for an access control system involves identifying which speaker among the registered ones matches the best with the test user and then verifying if the test user is that speaker. In order to do this task, two main processes that have to be included are the data collection to construct speaker models, and then the testing process.

Focus on the access control applications which require fast processing and as few burdens for users as possible, the number of time that one user has to utter his or her PIV during the registration (enrolment) process is limited to three times. From the enrolment data, speakers' features are extracted and then vectors containing mean values of each feature are stored as speaker reference models. At the same time, the system stores the TFS function for each speaker to be used in the decision process. Fig. 1 shows the data collection and making reference models for identification and verification tasks. Each registered speaker has two reference models: Model 1 is used for speaker identification (SI) task and Model 2 is used for speaker verification (SV) task.

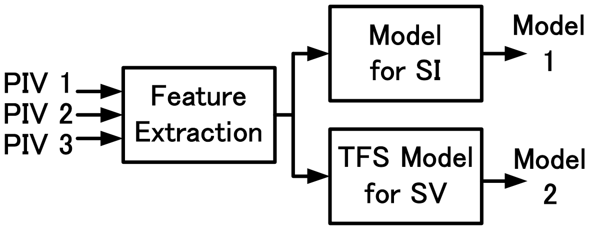

Fig. 1. Data collection to construct reference models for each speaker

For the testing process, the test user is required to utter the PIV that she/he used in the registration process. The system extracts the features, then calculates the distances from the feature vector of the test user to those of all registered speakers and designates the speaker with the minimum distance as the best match. After that, the system consults the TFS model of that speaker in the database and calculates the fuzzy similarity between the PIV of the speaker and that of the test user in order to decide whether to accept the test user or to reject. This process is presented in Fig. 2.

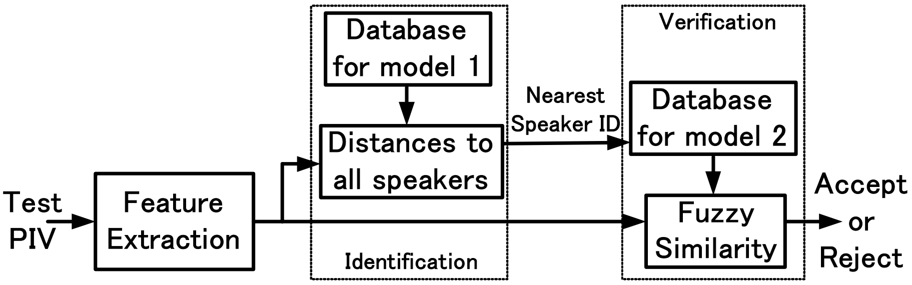

Fig. 2. Proposed recognition system

Suppose that the number of registered speakers available in the database *S* is *N*, that is:

(1) $$S = \{s_{i,m}\},\ i = 1,\ 2, ...,\ N,\quad m \in \{1,\ 2,\ 3\},$$

where *i* is the speaker ID, m is the order of the data recording and $s_{i,m}$ is PIV utterance of the speaker *i* at the *m*-th record.

2.1. Voice feature analysis

Although no speech feature is infallible at distinguishing speakers, the speech spectrum has been shown to be very effective for speaker identification [10]. Besides the application of the traditional acoustic features such as length ($f_{i,m,1}$), energy ($f_{i,m,2}$), zero-crossing rate ($f_{i,m,3}$), formant frequency 1 and 2 ($f_{i,m,4}$), ($f_{i,m,5}$), this work introduces the Bounded Variation Quantity (BVQ) ($f_{i,m,6}$) feature [3], and proposes the Energy Increase Frequency (EIF) ($f_{i,m,7}$), and the Voice Valley Number (VVN) ($f_{i,m,9}$). Although these acoustic features change each time users pronounce their PIV, they do not vary much. On the other hand, these features change very much from user to user. The effectiveness of the features used for speaker identification is examined by the discriminate ability for a certain database.

Before extracting voice features, the pre-processing of all the data is necessary. This process consists of cutting off all the irrelevant starting and ending parts and is done based on the frame energy threshold. After the feature extraction, values of all features have to be normalized in order to be used in the identification and verification process.

2.1.1. Bounded Variation Quantity (BVQ) ($f_{i,m,6}$)

Bounded variation quantity $V_{s_{i,m}}(t,\tau)$ of a signal $s_{i,m}$ at time t for an observed duration τ is known to be an effective feature for pattern recognition [3] and is calculated as $V_{s_{i,m}}(t,\tau)$:

$$f_{i,m,6} = V_{s_{i,m}}(t,\tau) = \sum_{i=0}^{\tau/\Delta} \left| s_{i,m}(t-\Delta i) - s_{i,m}(t) \right|, \tag{2}$$

where Δ is sampling time-interval, and τ/Δ is the number of checking points. When use BVQ to evaluate the similarity between two signals, the two signals have to have the same number of checking points. The observed time τ is chosen to be the length of the input signal, or $\tau = f_{i,m,1}$, Δ is adjusted such that $\tau/\Delta = \text{const}$. We use 1325 check points to calculate BVQ. This number of check points is confirmed experimentally.

2.1.2. Energy Increase Frequency (EIF) ($f_{i,m,7}$)

When a user utters the same PIV at different times, the way he/she stresses the word (increasing or decreasing energy of the voice) does not vary much. Thus, the frequency of the increasing energy EIF of the user’s pronounced PIV may not change significantly.

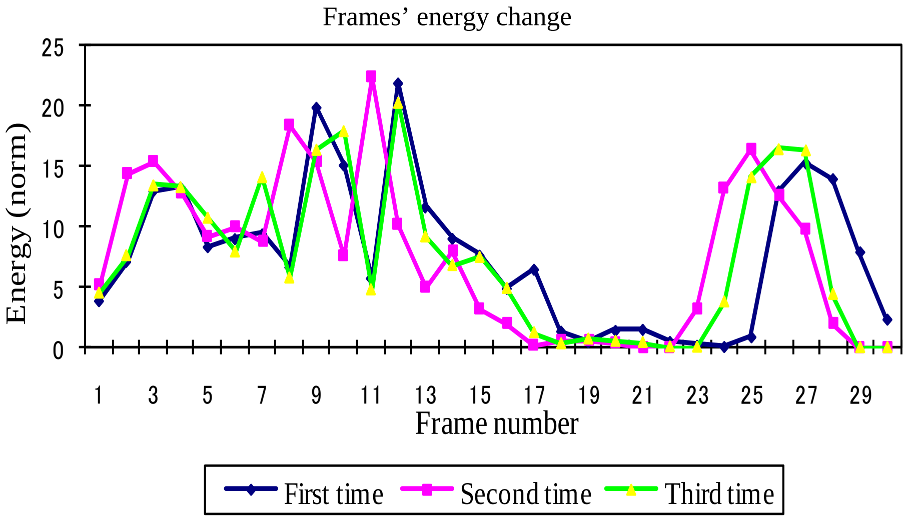

Fig. 3. EIF of "U-e-ha-ra-yu-ki-ko" utterance from a user

The observations show that by calculating the energy of each small frame, it is possible to find out how a user stresses a word. Fig. 3 shows the energy variation by frame (256 samples per frame) of a user with the utterance "U-e-ha-ra-yu-ki-ko" in three times. If we divide the input signal into frames of 256 samples, the number of frames is:

(3) $$\mathrm{FN} = \left\lfloor \frac{f_{i,m,l}}{256} \right\rfloor .$$

EIF is defined as follows:

(4) $$\mathrm{EIF} \underset{\mathrm{def}}{\square} f_{i,m,7} = \sum_{l=1}^{\mathrm{FN}-1} \begin{cases} 1 & \text{if } E_l < E_{l+1} \\ 0 & \text{otherwise} \end{cases} ,$$

where E_l is the energy of l-th frame. With the above definition, EIF value of the user in Fig. 3 is EIF=11, 11, 11 at three different sessions.

2.1.3. Voice Valley Number (VVN) ($f_{i,m,8}$)

To analyze the waveform in more detail, the input signal is smoothed by applying a filter such as Butterworth. Fig. 4 shows the waveform of the utterance "Ro-bo-to" after being smoothed. It is observed that the number of valleys of the wave varies depending on the chosen PIV and the way he/she speaks. However, from time to time, these valleys change slightly even for a single person. To prevent this intra-variation, amplitudes and time thresholds are used to define "valley". A portion of signal is called valley if and only if it satisfies the following condition:

(5) $$\begin{gathered} t_2 - t_1 > \tau > 0, \\ \begin{cases} y(t) = \delta, & t = t_1 \quad \text{or} \quad t = t_2, \\ y(t) < \delta, & t \in (t_1, t_2), \end{cases} \end{gathered}$$

where δ is the amplitude threshold and τ is the time threshold. These two thresholds are optimized to make the number of valleys stable for each user and to maximize the discriminate ability of the feature for a certain database. VVN is defined as the number of voice valley that satisfies the condition in (5). For this definition, the VVN of the user in Fig. 4 is equal to 2.

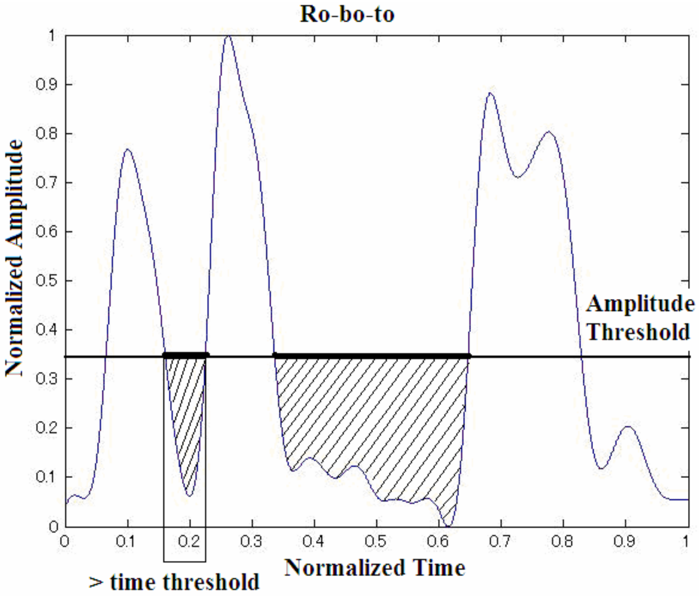

Fig. 4. Smoothed waveform of "Ro-bo-to" utterance

2.2. Speaker reference model for identification based on the Nearest Neighbor Approach

Based on the features extracted from the PIV a user utters three times during the registration process, the system builds a speaker reference model that is used in the identification process. A speaker reference model is an 8-dimension feature vector, in which the element in each dimension is the average of a feature evaluated at three different times. The reference model for speaker *i* is:

$$\vec{F}_i = \left(f^a_{i,1}, f^a_{i,2}, ..., f^a_{i,8} \right),$$

$$f^a_{i,j} = \frac{1}{3}\sum_{m=1}^{3} f_{i,m,j}, \quad j = 1, 2, ..., 8, \qquad (6)$$

Where $f_{i,m,j}$ is the value of feature *j*-th at the *m*-th recording during the enrolment of speaker *i*. The proposed system uses 8 features, *j*=1, 2, …, 8. This speaker model is stored in the system to be used in the identification process.

For the testing process, the test user is required to utter the PIV that she/he used in the registration process. The system extracts the 8 afore mentioned features, building a feature vector $\vec{x} = \left\{ x_j;\ j = 1, 2, ..., 8 \right\}$ and calculate the distances between this feature vector to all speaker models in the database. Then the Euclidean distance $d\left(\vec{x}, \vec{F}_i\right)$ from the feature vector of the test user $\vec{x}$ to that of speaker model *i* in the database $\vec{F}_i$ is denoted as d_i and calculated as:

(7) $$d_i = d\left(\vec{x}, \vec{F}_i\right) = \sqrt{\sum_{j=1}^{8}\left(x_j - f_{i,j}^{a}\right)^2} \ .$$

The system then assigns the speaker s_i^* as the one to whom the test user's feature vector is closest:

(8) $$s_i^* = \arg\min_{s_i \in S}\left(d_i\right).$$

2.3. Speaker verification based on Trapezoidal fuzzy similarity

After the system identifies the speaker s_i^* as the one being the most similar to the test user, it is necessary to verify whether the test user is really speaker s_i^*. In this process, a Trapezoidal Fuzzy Similarity (TFS) function is proposed to score the similarity between the PIV uttered by the test user and that of the speaker s_i^*.

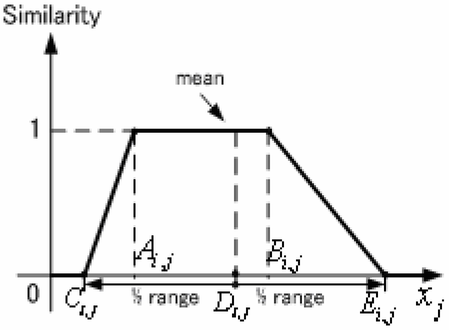

Fig. 5. Trapezoidal fuzzy similarity

The TFS function is constructed to estimate the speaker models for all the registered speakers in the database, and then the fuzzy similarity between the PIV of the test user and that of the referenced speaker is calculated. The speaker model of a user *i* based on TFS is shown in Fig. 5, where $A_{i,j}$, $B_{i,j}$, $D_{i,j}$ represent the minimum, maximum and mean values of the enrolment data for feature *j* and user *i*; and "range" represents the estimated variation of that feature for a specific user. It is estimated from the statistical data of one user recorded 242 times at different times in different days. An example for the length feature is shown in Fig. 6.

The statistical data shows that the feature distribution of one person is very close to a normal distribution. Therefore, it is assumed that the feature values for one person follow a normal distribution, and the range is estimated through the standard deviation σ. The value of "range" is chosen as 6σ because for a normal distribution, almost all values (99.7%) lie in the interval $[\mu - 3\sigma, \mu + 3\sigma]$. This range is used to absorb the intra-variation of a feature for a person. $C_{i,j}$, $E_{i,j}$ are calculated from $D_{i,j}$ with the distance of a half of range, and $h_{i,j}$ is the height of the trapezoid:

$$\begin{aligned}
&A_{i,j} = \min\left\{ f_{i,m,j} \right\}, \quad m = 1,\ 2,\ 3,\\
&B_{i,j} = \max\left\{ f_{i,m,j} \right\}, \quad m = 1,\ 2,\ 3,\\
&D_{i,j} = \frac{1}{3}\sum_{m=1}^{3} f_{i,m,j},\\
&C_{i,j} = D_{i,j} - \frac{1}{2}\text{range},\\
&E_{i,j} = D_{i,j} + \frac{1}{2}\text{range},\\
&\text{range} = 6\sigma.
\end{aligned} \tag{9}$$

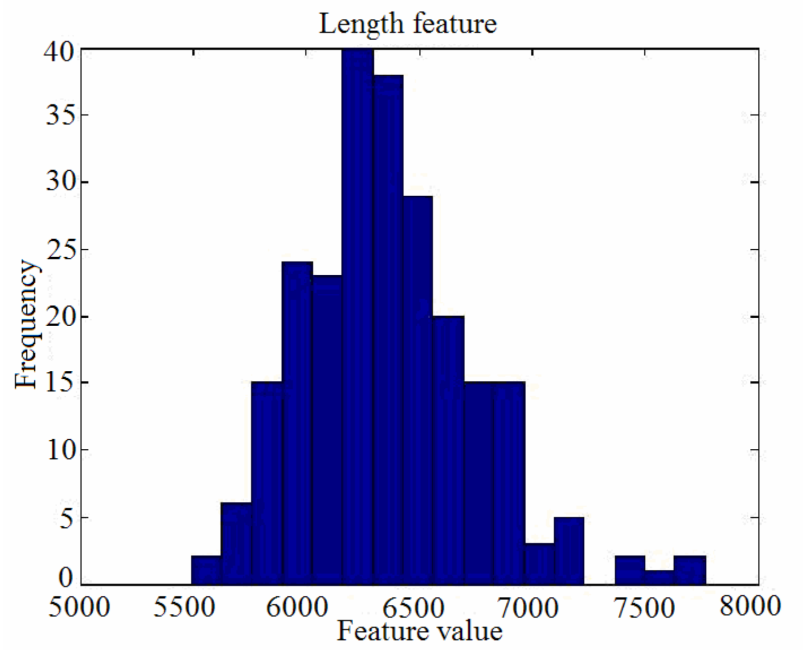

Fig. 6. Statistical data of the length feature of a person recorded 242 times at different moments

The system stores a set of parameters $\{A_{i,j}, B_{i,j}, C_{i,j}, E_{i,j}\}$, $j = 1, 2, \ldots, 8$. The fuzzy similarity between the test user having x_j as the value of feature j with the referenced speaker s_i^* regarding feature j is:

(10)
$$\mathrm{FS}^*_{i,j} = \begin{cases} \dfrac{x_j - C^*_{i,j}}{A^*_{i,j} - C^*_{i,j}} & \text{if } C^*_{i,j} < x_j < A^*_{i,j} \\ 1 & \text{if } A^*_{i,j} \le x_j \le B^*_{i,j} \\ \dfrac{E^*_{i,j} - x_j}{E^*_{i,j} - B^*_{i,j}} & \text{if } B^*_{i,j} < x_j < E^*_{i,j} \\ 0 & \text{otherwise} \end{cases},$$

where $A^*_{i,j}, B^*_{i,j}, C^*_{i,j}, E^*_{i,j}$ are the parameters of the TFS speaker model for feature *j* of s^*_i. Following the same process for the other features, we get a set of fuzzy similarity scores. The final fuzzy similarity score is calculated based on fuzzy *T*-norm as the following:

(11)
$$\mathrm{FS}^*_i = \underset{j=1}{\overset{8}{T}} \left(\mathrm{FS}^*_{i,j} \right).$$

Either the product *T*-norm or the minimum *T*-norm would give the same result; the latter, however, has shorter computation time. Thus, the minimum *T*-norm is chosen to calculate fuzzy similarity.

Based on the final score of the fuzzy similarity between the test user with the referenced speaker s^*_i on the database, we can verify whether the PIV of the test user and that of the referenced speaker is similar or not, thus deciding to accept or reject the test user. The final decision is done based on the following rule:

(12)
$$\begin{aligned} \mathrm{FS}^*_i &> 0 \rightarrow \text{accept}, \\ \mathrm{FS}^*_i &= 0 \rightarrow \text{reject}. \end{aligned}$$

3. Experimental results on speaker recognition

3.1. Database and description of the evaluation method and experiments

The experiments are performed on a voice database of 24 people. This database consists of utterances from 21 Japanese and 3 non-Japanese male speakers recorded at four different sessions. Each speaker is requested to speak the same word in all the four sessions, which is considered as his/her PIV. To estimate the variation of voice features, a PIV data uttered at 242 different sessions of a person is recorded. Through this statistical data, a variation range is estimated and then applied to calculate the TFS reference model for all the users. All the data is recorded in quiet classrooms at Tokyo Institute of Technology with a Olympus Voice-Trek V-61 recorder at 8 kHz mode.

The experiment is designed to evaluate two points: 1) the effectiveness of the proposed feature set; 2) the performance of the speaker recognition system, in which identification and verification processes are involved. The use of the database and the evaluation of the above two points are described below.

3.1.1. Effectiveness of the proposed feature set

The effectiveness of the proposed feature set is evaluated based on discriminate ability [2], [20] and calculation time. The feature with higher discriminate ability, that is, the one which can correctly distinguish more speakers, and with faster calculation time is considered better. In this work, feature's discriminate abilities are defined based on the mean value and the width range calculated from the statistical distribution of a user's data recorded 242 times, which constitute an extension of [2]. All the data recorded from 24 speakers at four different sessions is used to calculate the discriminate ability. To equally evaluate the discriminate ability of different features, all feature values are normalized to the interval [0, 1] as the following expression:

$$f'_{i,m,j} = \frac{f_{i,m,j}}{\max_j(f_{i,m,j})}\,. \tag{13}$$

In the definition given in [2] (Fig. 7a), the discriminate ability of one feature between user 1 and user 2 is defined as:

$$d_{1,2} = \frac{a}{b}\,, \tag{14}$$

where a is the distance between the maximal feature value $B_{1,j}$ of user 1 and minimal value $A_{2,j}$ of user 2, and b is the distance between the minimal feature value $A_{1,j}$ of user 1 and maximal feature value $B_{2,j}$ of user 2. In this work, the use of the distribution's width as the range of feature value for each user (Fig. 7b) is proposed, in other words, every user has the feature value range equal to the distribution's width. Each user's vector has a different mean point, and from that mean point, the minimal and maximal points of the feature value are calculated. Therefore, the discriminate value between user 1 and user 2 is

$$d'_{1,2} = \frac{a'}{b'} = \frac{C_{2,j} - E_{1,j}}{E_{2,j} - C_{1,j}}\,. \tag{15}$$

Then, the discriminate ability of feature j for a specific database S is

$$D_j = \sum_{i=1}^{N}\sum_{k=i+1}^{N} d_{i,k}\,. \tag{16}$$

The higher the discriminate ability is, the better one feature can classify speakers inside the database. If two features have the same discriminate ability, the feature with shorter calculation time is considered to be better.

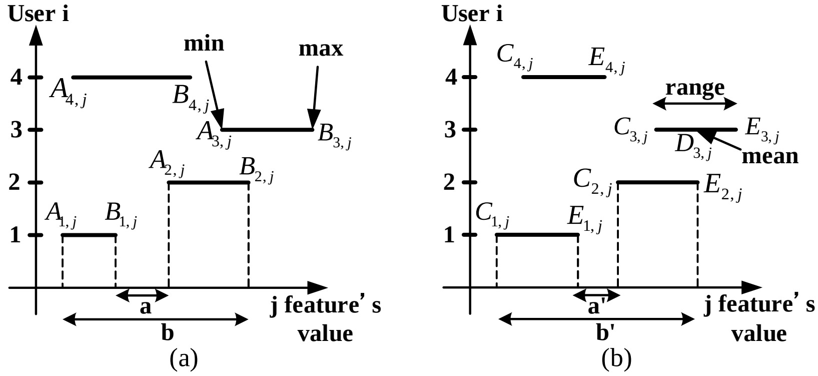

Fig. 7. Discriminate ability in case of four users with four sessions each in (a) conventional definition (b) extended definition

In addition, by building a discriminate table for each feature and then calculating the summary table, it is possible to evaluate whether the feature set can distinguish one speaker from the others or not.

3.1.2. Evaluation of the proposed recognition system

An auto access control system based on voice recognition aims to accept the registered speakers and reject un-registered ones. Accordingly, the experiments are divided into two tests in order to confirm the performance for users who are registered in database and for users who are not registered in the database.

For the former, cross validation is performed on the data recorded from 24 speakers at four different sessions. The data from three sessions are used for training and constructing speaker models and the data of the remaining session is used for testing. The training data is selected randomly from four sessions, and thus creates 4^{24} combinations. It is, however, time-consuming to test on all trials (combinations), thus we limit the number of trials to a level that is acceptable in calculation time and error margin. The experiment is done on 100 000 trials. For 99% confidence with $10\,000$ trials, we can obtain an error margin of $1.29/\sqrt{n} = 1.29/\sqrt{10\,000} = 1.29\%$ (n is number of samples or trials). The performance on this test is evaluated based on the average percentage of the three cases: the system 1) wrongly rejects a registered user (false positive errors); 2) accepts a registered one but recognize him/her as another person in the database; 3) truly accepts a registered user. In this test, the worst situation happens when a registered user is wrongly rejected (false positive errors).

For the latter, cross validation is also performed on the same data of 24 speakers. 20 speakers are randomly selected out of 24 speakers in order to train and three sessions among four are randomly chosen to construct the database of registered speakers. The data from the remaining 4 speakers are used for testing. This way of selecting data gives ${}^{24}C_{20} \times {}^{4}C_{3} = 42\,504$ combinations. Similarly, the experiment is performed on $11\,442$ trials to obtain an error margin of

$1.29/\sqrt{11442}=1.20\%$ at 99% confidence. The performance on this test is evaluated based on the average percentage of two cases: the system 1) wrongly accepts an unregistered user; 2) truly rejects an unregistered user. The occurrence of the former case (false negative errors) must be reduced.

At the same time, the performance of the proposed speaker recognition based on TFS is compared with that of the method based on nearest neighborhood approach using threshold values.

3.2. Experimental results on discriminate ability and calculation time

Table 1 shows the discriminate ability and calculation time of features for the database of 24 speakers. The features are Length, Zero Crossing Rate (ZCR), Bounded Variation Quantity (BVQ), Energy, Formant Frequency 1 (FF1), Formant Frequency 2 (FF2), Energy Increasing Frequency (EIF), and Voice Valley Number (VVN). The feature extraction code and speaker recognition one are programmed in Matlab 7.0 on a 1.8 GHz, Intel Core Duo CPU.

Table 1. Discriminate ability and searching order

Feature	Discriminate Ability	Calculation Time (ms)	Searching Order
Length	273.74	0.007	1
BVQ	213.34	0.323	2
Energy	204.43	0.323	3
VVN	203.41	1.458	4
FF2	192.69	28.323	5
EIF	169.56	1.625	6
FF1	144.84	28.323	7
ZCR	144.17	1.146	8

It is shown that for this database of 24 users, the proposed two features BVQ and VVN have high discriminate ability, higher than conventional acoustic features such as FF1, and FF2, whereas the proposed EIF is higher than FF1, ZCR.

3.3. Performance comparison of speaker recognition based on the Nearest neighbor approach and the Proposed trapezoidal fuzzy similarity

In the first test, 3 sessions are randomly selected and speaker models are constructed based on these data. 10 000 trials are generated with different random combinations of three sessions out of four recording ones for the database. In the second test, 20 speakers are randomly chosen to construct the database of the registered speakers, and 3 among four sessions are also randomly chosen to build the reference speaker models for those 20 speakers. 11 442 trials are generated with different random combinations of 20 speakers out of 24 ones for the database.

Table 2. Speaker recognition using the nearest neighbor approach based on threshold

Criterion	Without threshold	Threshold = 0.12
False Positive Error (Wrongly Rejection – Test 1)	3.13%	51.3%
False Negative Error (Wrongly Acceptance – Test 2)	100.00%	0%

In the nearest neighbor (NN) approach, there is a trade-off between the results for false positive error (wrong rejection of the registered users in the first test) and false negative error (wrong acceptance of the un-registered users in the second test). Table 2 shows the speaker recognition results for the nearest neighbor approach. Without any threshold, the recognition system based on NN approach leads to 3.13% of false positive error in average for test 1 with 10 000 trials and 100% of false negative error in average for test 2 with 11 442 trials. When the threshold for NN distance is set to 0.12 such that no false negative error exists, the average false positive error is 51.3%.

Instead of using a threshold value to decide whether to accept or reject a user, the TFS model is proposed to score the similarity between the closest reference speaker and the test user, and then decide to accept or reject the test user based on that similarity. On average, the proposed system shows no false negative error in the second test which means it rejects un-registered users 100% of the time as shown in Table 3. Table 4 shows the detailed results of the first test for the proposed recognition system. The proposed system has 21.25% of false positive error, 30.05% lower than the results for the NN-threshold approach. However, among the registered users who are accepted, some of them are wrongly recognized as a different one in the database.

Table 3. Recognition result for the proposed system in the second test
with the confidence of 99% and error margin of 1.20%

Criterion	Average recognition result of 11442 trials
Wrongly accepted (False negative)	0.00%
Truly rejected	100.00%

Table 4. Recognition result for the first test with the confidence of 99%
and error margin of 1.29%

Criterion	Average Result of 10 000 trials	Standard deviation
Wrongly rejection (False positive)	21.25%	8.00%
Truly acceptance but wrongly recognization	16.19%	6.51%
Truly acceptance and truly recognization	62.56%	9.31%

The above result is the average result of 10000 trials, and the distribution of the accurately recognized cases is shown in Fig. 8.

4. Discussion

The proposed feature set is proved to be able to discriminate speakers for the database of 24 speakers. Experiments show that the two proposed features have comparatively higher discriminate abilities, comparing to the traditional features such as formant frequency, zero-cross rate and so on. The feature extraction process is also very fast, being possible for real-time applications.

With only three short utterances in the enrolment process, the proposed system can truly reject 100% of un-registered users and accept 78.75% of registered speakers. This result is reliable for the database of 24 speakers. However, it is necessary to test it on larger databases. Also, for practical applications, it is necessary to minimize the false positive and false negative errors to zero.It is observed that there are some speakers who are more difficult to be distinguished from others, or in other words, have a higher probability of being mis-recognized. To reduce this error, the capture of inter-speaker information as proposed in [11] is one of the possible solutions. Another problem comes from the data itself. Even if all the data is recorded in a quiet room, the presence of noise as well as echoes is unavoidable, and they affect the recognition performance.

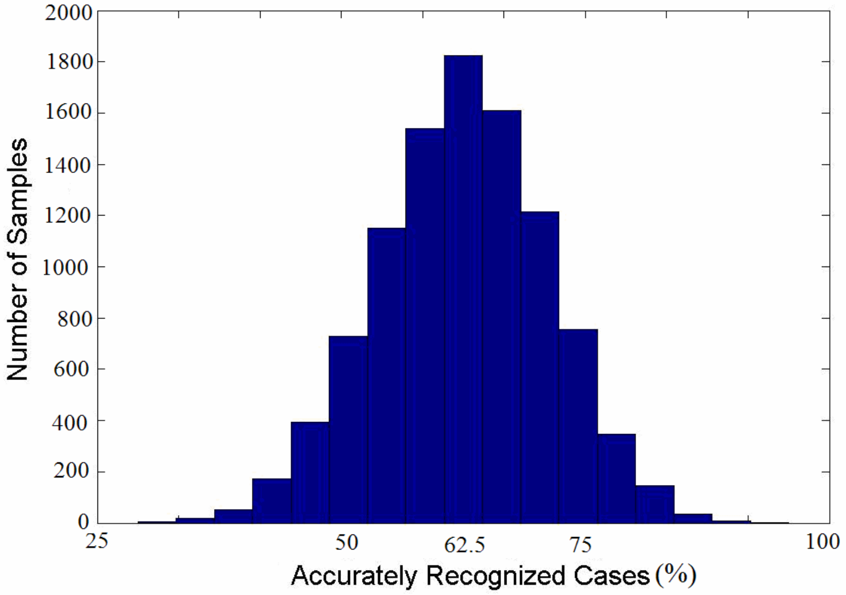

Fig. 8. Distribution of the truly accepted and recognized result on 10 000 trials

The proposed method shows to be effective in computation time on both the training process and the recognition process. Another advantage of the proposed method is that each time a new user is added to the database, it is not necessary to train the database again to find a new set of parameters such as in other methods using Gaussian mixture model or neural networks. Instead, it is just necessary to calculate the reference models for the newly added speaker.

5. Conclusion

A speaker recognition method based on Trapezoidal Fuzzy Similarity is proposed. A set of five conventional acoustic features and three new features is used to estimate speaker models based on PIVs uttered three times by each user. This feature set is evaluated and verified to be effective based on the ability of discriminating users inside the database. These speaker models are referenced to find out the registered user closest to the test user. However, because the number of utterance is limited, it is difficult to capture the variation of a person at different times. Thus, a statistical data created from a PIV uttered 242 times by one person is used to calculate the variation range used to build the Trapezoidal Fuzzy Similarity for each user. These models are used to evaluate the similarity between the closest speaker and the test user, so that the system can decide whether to accept or reject the test user. The proposed system can truly reject 100% of un-registered users and accept 78.75% of registered speakers, 30.05% higher than the nearest neighbor method using the threshold that can reject 100% of un-registered users.

The experiment data can be statistically processed in common, considering only the evaluation of a commonly averaged σ for the common Gaussian model. The fact, however, is that the deviation of the data from three times is not reliable enough, the common average σ which is again based on these unreliable variations, thus is not reliable. Furthermore, it is more difficult to take many data points (record many times) for many persons than to take a lot of data from only one person. Therefore, we do not consider the above statistical process, due to the reliability of the data as well as the feasibility on the real system.

Acknowledgements. This work is partially supported by MEXT (Japanese Ministry of Education, Culture, Sports, Science and Technology).

References

1. Campbell, J. P. Speaker Recognition: a Tutorial. – In Proc. of the IEEE, Vol. **85**, 1997, No 9, 1437-1462.
2. Arai, Y., K. Hirota. Fuzzy Hierarchical Pattern Recognition for Robotics Applications. – LNCS: Advanced Topics in Artificial Intelligence, Vol. **1342**, 1997, 274-281.
3. Hirota, K. The Bounded Variation Quantity (BVQ) and Its Application to Feature Extraction. – Pattern Recognition, Vol. **15**, 1982, No 2, 93-101.
4. Reynolds, D. A., R. C. Rose. Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models. – IEEE Transactions on Speech and Audio Processing, Vol. **3**, January 1995, No 1.
5. Markel, J. D., B. T. Oshika, A. H. Gray. Long-Term Feature Averaging for Speaker Recognition. – IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. **ASSP-25**, August 1977, No 4.
6. Gish, H., K. Karnofsky, M. Krasner, S. Roucos, R. Schwartz, J. Wolf. Investigation of Text-Independent Speaker Identification Over Telephone Channels. – In: Proc. of IEEE ICASSP, 1985, 379-382.
7. Rudasi, L., S. A. Zahorian. Text-independent Talker Identification with Neural Networks. – In: Proc. of IEEE ICASSP, May 1991, 389-392.
8. Bennani, Y., P. Gallinari. On the Use of TDNN-Extracted Features Information in Talker Identification. – In: Proc. of IEEE ICASSP, May 1991, 385-388.
9. Oglesby, J., J. Mason. Radial Basis Function Networks for Speaker Recognition. – In: Proc. of IEEE ICASSP, May 1991, 393-396.
10. Atal, B. Automatic Recognition of Speakers from their Voice. – IEEE Trans on ASSP, April 1976, 460-475.
11. Wang, L., K. Chen, H. Chi. Capture Inter-speaker Information with a Neural Network for Speaker Identification. – IEEE Transactions on Neural Networks, Vol. **13**, March 2002, No 2.
12. Hasan, M. R., M. Jamil, M. G. Rabbani, M. S. Rahman. Speaker Identification using Mel-Frequency Cepstral Coefficients. 3rd International Conference on Electrical & Computer Engineering, ICECE 2004, December 2004, 565-568.

13. Mikhael, W. B., P. Premakanthan. An Improved Speaker Identification Technique Employing Multiple Representations of the Linear Prediction Coefficients. – In: Proc. of ISCAS 2003, Vol. **2**, May 2003, 584-587.
14. Sambur, M. Selection of Acoustic Features for Speaker Identification. – IEEE Transactions on Acoustic, Speech, and Signal Processing, Vol. **23**, April 1975, Issue 2, 176-182.
15. Tran, D., M. Wagner. Fuzzy Nearest Prototype Classifier Applied to Speaker Identification. – In: Proc. of ESIT' 99, Abstract, 1999, p. 34.
16. Tran, D., M. Wagner. Fuzzy C-Means Clustering-Based Speaker Verification. – LNCS on Advances in Soft Computing – AFSS2002, Vol. **2275**, January 2002, 318-324.
17. Soong, F. K., A. E. Rosenberg, L. R. Rabiner, B. H. Juang. A Vector Quantization Approach to Speaker Recognition. – In: Proc. of ICASSP-85, Vol. **10**, April 1985, 387-390.
18. Higgins, A. L., L. G. Bahler, J. E. Porter. Voice Identification Using Nearest-Neighbor Distance Measure. – In: Proc. of ICASSP-93, Vol. **2**, April 1993, 375-378.
19. Matsui, T., S. Furui. Comparison of Text-Independent Speaker Recognition Methods Using VQ-Distortion and Discrete/Continuous HMM's. – IEEE Trans. on Speech and Audio Processing, Vol. **2**, July 1994, No 3, 456-459.
20. Hirota, K., Y. Arai, S. Hachisu. Moving Mark Recognition and Moving Object Manipulation in Fuzzy Controlled Logic. – CTAT, Vol. **2**, 1986, No 3, 399-418.