Multimed Tools Appl (2015) 74:5131–5140
DOI 10.1007/s11042-013-1587-5

Enhancing GMM speaker identification by incorporating SVM speaker verification for intelligent web-based speech applications

Ing-Jr Ding · Chih-Ta Yen

Published online: 12 July 2013

Abstract Speech applications, which operate a system by voice commands, facilitate web access for disabled and visually impaired users. Human-computer interactions, such as speaking and listening to web applications, provide options for developing a multimodal interaction tool in the accessible design of an intelligent web. Speaker identification and verification are essential functionalities for intelligent web programs with speech applications. This paper proposes an enhanced Gaussian mixture model (GMM) method by incorporating the information derived from the support vector machine (SVM), called EGMM-SVM, for web-based applications with speaker recognition. The EGMM-SVM improves the accuracy of the estimated likelihood scores between the speech frame and the GMM. In EGMM-SVM, SVM plays a crucial role in transmitting the quality information of the utterances from a test speaker, through the GMM when performing GMM likelihood calculations. The experimental results show that speaker recognition by using the developed EGMM-SVM with an accurate operation mechanism for Gaussian distribution derivations yields a higher recognition rate than does a conventional GMM without any considerations on the quality of test speech utterances.

Keywords EGMM-SVM · Gaussian mixture model · Support vector machine · Speaker recognition · GMM likelihood score

1 Introduction

Web-based social media services enable users to easily access information and web services anywhere and anytime through network connections. A user-friendly interface facilitates convenient access for users, and multimodal interfaces make web-based social media technology easier to use, more efficient, and more acceptable to end-users. Applications allowing people to interact on the web through speech provide the most practical and natural mode of communication [9]. Intelligent web-based speech applications typically include

I.-J. Ding · C.-T. Yen (✉)
Department of Electrical Engineering, National Formosa University, No.64, Wunhua Rd., Huwei Township, Yunlin County 632, Taiwan, Republic of China
e-mail: chihtayen@gmail.com

popular automatic speech recognition (ASR) and speaker recognition, which provide security when using voice commands to access personal, business or public information online.

Recently, speaker recognition techniques belonging to a type of audio-based identity recognition have provided effective security in surveillance, remote homecare, and web service applications. Compared with facial recognition [15, 17], fingerprint recognition [1, 7] and gesture recognition [10], which are categorized as video-based identity recognition techniques, speaker recognition adopts the biomedical features of acoustic data and acts as an auxiliary recognition technique that reflects the auditory aspect of the reality in the context. Speaker recognition technology is rapidly becoming as developed as speech recognition [8], and numerous computational techniques for speaker recognition have been observed in recent years [2, 4, 5, 11, 12, 14, 18–20]. Nevertheless, the most crucial problem in speaker recognition is recognition accuracy.

Speaker recognition may be further divided into two categories: speaker identification and speaker verification. Speaker identification is used to determine the identity of a person. A speaker verification system verifies the identity of a person based on his or her uttered voice, and evaluates whether the speaker is acceptable or not. When GMM-based speaker recognition is adopted [16] for speech-pattern recognition, it performs more effectively in speaker identification. When SVM-based speaker recognition is employed [3] as a speech pattern classifier, it provides a more favorable option for executing speaker verification. This study focused on GMM-based speaker identification tasks.

Although the GMM approach is the optimal choice for performing speaker identification, the recognition accuracy of the overall speaker recognition system is still inferior to that of a human listener. Recent studies have shown improvement in the performance of GMM speaker recognition [4, 11, 12, 18, 19]. The enhancement of GMM speaker recognition is further categorized into two types of techniques: model-based and feature-based improvement approaches. Model-based improvement methods aim to enhance the GMM classification model when training the GMM [11, 12]. A new algorithm was proposed in [11] for speaker verification applications in discriminative training of the GMM with diagonal covariances under a large-margin criterion. For training a large-scale generative model of speaker and session variability, Kenny et al. (2007) presented a corpus-based approach to GMM speaker verification, in which maximum-likelihood II criteria were used [12].

Feature-based improvement methods for GMM speaker recognition focus on emitting the unexpected noise of the input test speech signal, or developing an acoustic feature that facilitates characterization of a speaker's information [4, 18, 19]. You et al. developed a Bhattacharyya-based GMM-distance to measure the distance between two GMM distributions, allowing the speaker's information to be exploited not only from the mean vectors of GMM but also from the covariance matrices [18, 19]. In addition, several feature extraction and channel compensation techniques in a GMM speaker recognition system were analyzed and discussed in [4].

Although model-based and feature-based enhancements to GMM speaker recognition increase the recognition accuracy of the system, those approaches cannot ensure that a satisfactory recognition performance is maintained when substandard test data for recognition are encountered. Related research on the validation and evaluation of test data is rarely seen in the field of GMM speaker identification. Inadequate test data with an ambiguous class tendency would jeopardize the recognition performance of a GMM speaker recognition system. For general speech-pattern recognition techniques, including GMM-based speaker identification, the quality of test data is most essential for recognition accuracy. To address this problem, the SVM that is popularly adopted in speaker verification for evaluating and verifying the availability of the data from a test speaker was used. This paper proposes an enhanced GMM method with the support of the SVM, called EGMM-SVM, for speaker identification. In

EGMM-SVM, the SVM is integrated into the conventional GMM-based speaker identification scheme to evaluate the availability of the test data. The information derived from SVM verification is evaluated by the GMM classifier when performing the likelihood calculation between the speech frame and the GMM speaker models. The proposed EGMM-SVM speaker identification with the assistance of the SVM speaker verification offers several advantages:

- It decreases unreliable recognition decisions in conventional GMM-based speaker identification methods by incorporating the SVM for assessing test data;
- It provides a new scheme, combining the SVM speaker verification and GMM speaker identification for practical speaker recognition applications;
- It achieves more robust recognition using the improved likelihood estimate of GMM classifiers, especially in adverse conditions, in which the test data are of extremely inferior quality.

2 GMM-based speaker recognition

As mentioned, modeling schemes are the mainstream techniques for speaker recognition, and the modeling of speech patterns implemented in the Gaussian mixture model is by far the most popular and widely used scheme. The operational architecture of GMM speaker recognition and the modeling methodology of GMM speaker models are introduced in the following sections.

2.1 Operation architecture of GMM speaker recognition

Figure 1 illustrates the overall operational architecture of a GMM-based speaker recognition system, in which there are two primary processing phases in the speaker recognition framework: the training\establishment phase of GMM speaker models, and the test\recognition phase of GMM classifiers. When performing speaker recognition in a practical application, the input utterances acquired from a speaker are segmented into the frame sequence from which acoustic features are extracted to determine the degree of likelihood for all trained GMM speaker models through the operation of GMM classifiers. The recognizing operation is then completed and the decision to categorize the test speaker as one of all speaker classes can be made after accumulating the degree of likelihood estimates for all of the GMM speaker models in a predefined time period.

2.2 GMM and speaker modeling

In this work, a GMM is adopted in the development of a speaker recognition system [16]. Mathematically, a GMM is a weighted sum of M Gaussians, denoted as

$$\lambda = \{w_i, \mu_i, \Sigma_i\}, \quad i = 1, 2, ..., M, \quad \sum_{i=1}^{M} w_i = 1, \tag{1}$$

where w_i is the weight, μ_i is the mean and Σ_i is the covariance.

To determine the GMM parameters for a certain speaker class, the E-M algorithm suggested in [6] is readily applicable. Before running the E-M algorithm, it is crucial to initialize the model by assigning starting values to the parameters. These can be realized by a binary splitting vector quantization algorithm [13]. With the parameter settings of the initial

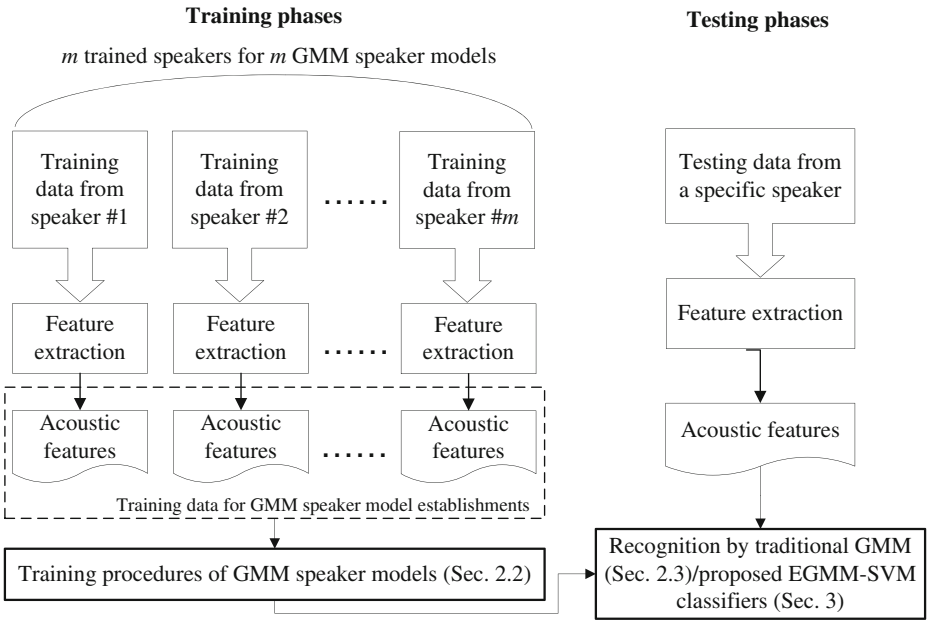


Fig. 1 Operational architecture of a GMM-based speaker recognition task

model, the E-M process starts iteratively, maximizing the likelihood estimate of the training data from the speaker by adjusting the initial model parameters. The expectation and maximization steps in the E-M process are repeated so that the parameter set as $\lambda = \{w_i, \mu_i, \Sigma_i\}, \quad i = 1, 2, ..., M$ of the GMM converges to an equilibrium state.

2.3 GMM classifier for likelihood score calculations

After completing the training of the GMM, the speaker recognition procedure can then be executed based on these trained GMM. Note that the speaker identification used here is a GMM classifier consisting of multiple GMM speaker models, which are categorized into two types: the valid speaker models and the imposter models. The classifier operates with a decision window (or its equivalent, over an interval) covering n acoustic feature vectors of D dimensions, $X=\{x_i|i=1,2,...,n\}$, combined with n GMM speaker models, $\lambda_1, \lambda_2, \ldots, \lambda_n$.

During the recognition phase, the class of X is determined by maximizing *a posteriori* probability $P(\lambda_s|X)$ [16],

$$\widehat{s} = \underset{s=\{1,2,...,n\}}{\operatorname{argmax}} P\left(\lambda_s \middle| X\right) = \arg \max_{s=\{1,2\}} \frac{f\left(X \middle| \lambda_s\right)}{f(X)} \cdot P(\lambda_s) \tag{2}$$

Note that

$$f\left(x_i \middle| \lambda_s\right) = \sum_{j=1}^{M} w_j \cdot b_{s_j}(x_i), \tag{3}$$

and

$$b_{s_j}(x_i) = \frac{1}{(2\pi)^{D/2} \cdot \left|\Sigma_{s_j}\right|^{1/2}} \cdot \exp\left\{-\frac{1}{2}\left(x_i - \mu_{s_j}\right)^T \left(\Sigma_{s_j}\right)^{-1}\left(x_i - \mu_{s_j}\right)\right\}. \quad (4)$$

However, in real implementation, Eq. (1) is replaced by

$$\hat{s} = \arg \max_{s=\{1,2,\ldots,n\}} \sum_{i=1}^{n} \log f\left(x_i \middle| \lambda_s\right), \quad (5)$$

for simplicity. At the end of the recognition procedure, the signal X is then recognized as one of the n speaker classes indicated by $\hat{s}$.

3 Enhanced GMM by the information from SVM (EGMM-SVM)

In a practical speaker recognition application, the operational performance of the GMM classifier has a definitive influence on the accuracy of speaker recognition. An excellent GMM classifier with outstanding recognition performance is necessary. The operational performance of a GMM classifier depends strongly on the quality of the test utterances obtained from the speaker. The higher the degree of discrimination is in the test utterance, the more qualified the utterance would be. However, when performing speaker recognition in practical online applications, the test data acquired from a speaker are usually viewed as substandard if the data lack in distinguishability. To address this problem and increase the recognition accuracy of GMM speaker recognition, the test data are first verified using an SVM mechanism. The appraised data derived from the SVM are then accounted for when the GMM classifier is performing. The overall speaker recognition process includes SVM speaker verification and GMM speaker identification, which are depicted in Fig. 2.

3.1 Analysis of test data by SVM

This section introduces the SVM classification schemes that were adopted for analyzing the availability of a test utterance and for evaluating the differentiation degree of the utterance. In most applications, the SVM is used as a data classifier [3]. The SVM is based on the theory of the structural risk minimization of statistics. The SVM classifies new input data by using a separating hyperplane. To determine whether an input speech datum belongs to the valid speaker set, the SVM first attempts to locate the SVM model for the valid speaker set in the SVM database. The separating hyperplane of the SVM model for the valid speaker set then classifies the input speech datum as either valid or invalid (the imposter). In this study, the trained SVM model for speaker verification was established in a supervised-mode environment where two categories of training speakers, valid speakers and imposters (those not in the group of valid speakers), were collected, and the class label for each training sample was known before training the SVM.

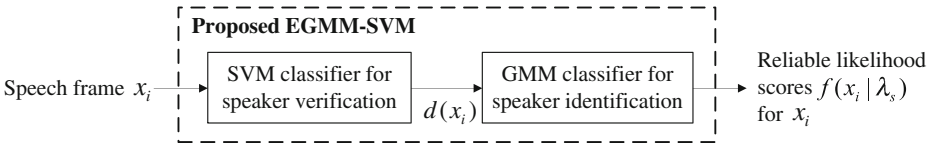


Fig. 2 Proposed EGMM-SVM for GMM-based speaker identification

Suppose a set of labeled training points is (x_1,y_1), (x_2,y_2),…,(x_n,y_n). Each training point x_i belongs to either of two classes and is assigned a label,$y_i \in \{-1, 1\}$, for i=1,2,...,n. Based on these training data, the hyperplane is

$$w \cdot x + b = 0, \tag{6}$$

which is defined by the pair (w,b), such that the point x_i can be separated according to the function

$$f(x_i) = sign(w \cdot x_i + b) = \begin{cases} 1, & if \ \ y_i = 1 \\ -1, & if \ \ y_i = -1 \end{cases} \tag{7}$$

The set S is linearly separable if a pair (w,b) exists such that the inequalities

$$\begin{cases} (w \cdot x_i + b) \geq 1, & if \ \ y_i = 1, \\ (w \cdot x_i + b) \leq -1, & if \ \ y_i = -1, \end{cases} \qquad i = 1, 2, ..., n, \tag{8}$$

are valid for all elements of set S. Equation (3) can be rewritten as one set of inequalities as follows:

$$y_i(w \cdot x_i + b) - 1 \geq 0, \quad \forall i. \tag{9}$$

This study used the SVM model to determine the quality of each test datum. A trained SVM hyperplane was selected to separate the valid speakers from the imposters and to verify the ith test speech frame x_i. The index $d_{SVM}(x_i)$, indicating the distance between the speech frame x_i and the SVM separating hyperplane, could effectively govern the degree of availability of the ith speech frame x_i. Figure 3 clearly shows the meaning of the index $d_{SVM}(x_i)$ in the SVM separation hyperplane classification space.

3.2 Proposed EGMM-SVM

The quality of the test data obtained from a speaker for the GMM classifier calculation in the recognition phase immediately affects the classification accuracy of the GMM classifier in the online operational phase. As mentioned, inaccurate GMM recognition calculation caused by inadequate test data with an indefinite class tendency is alleviated when the SVM classifier evaluates the test data before performing the GMM classification. Incorporating an SVM classifier into the GMM-based speaker recognition process to estimate the availability of test data before recognition calculation further enhances the robustness of GMM speaker recognition.

In conventional GMM-based speaker recognition, the likelihood score of certain speech frame is determined by Eq. (3). However, Eq. (3) does not show information about the quality of the speech frame x_i revealed. For speaker recognition techniques, including GMM, the quality of test data for GMM classification calculation is the most crucial consideration. Inadequate test data with an ambiguous class inclination would most likely lead to an unreliable estimate of GMM likelihood scores, which inevitably jeopardizes the recognition performance of a speaker recognition system. To address this problem, an EGMM-SVM method is proposed. EGMM-SVM provides an effective formula for estimating GMM likelihood scores as follows:

$$f\left(x_i \middle| \lambda_s\right) = \frac{d_{SVM}(x_i)}{C} \cdot \sum_{j=1}^{M} w_j \cdot b_{s_j}(x_i), \tag{10}$$

where $d_{SVM}(x_i)$ could be used to effectively govern the degree of availability of the ith speech

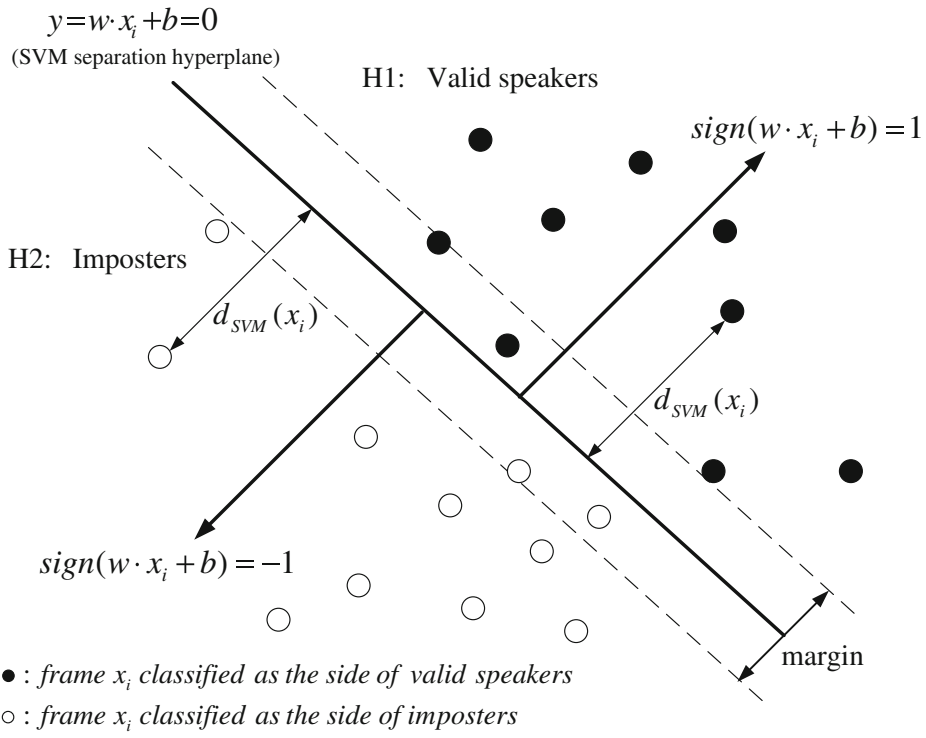


Fig. 3 Index $d_{SVM}(x_i)$ was derived from the SVM classification space for use in the GMM classifier calculation

frame, and x_i is the distance between the speech frame x_i and the SVM separating hyperplane; C is a constant denoting the scaling factor of $d_{SVM}(x_i)$.

In Eq. (10), $\sum_{j=1}^{M} w_j \cdot b_{s_j}(x_i)$ denotes the likelihood score of the frame x_i for a certain GMM, λ_s, and the score could be precisely governed by the index $d_{SVM}(x_i)$. In Eq. (10) the accuracy of the likelihood score $f(x_i|\lambda_s)$ could be effectively regulated by the index $d_{SVM}(x_i)$. When the quality of ith speech frame x_i is in doubt because it falls within the scope of the SVM margin and the distance from the SVM hyperplane is small [i.e., $d_{SVM}(x_i)$ is small], then the unreliable GMM likelihood score is calculated due to this inadequate test data that have indefinite class inclination. In this case, the GMM likelihood score should be less referenced. Conversely, when a large $d_{SVM}(x_i)$ is calculated, the ith speech frame x_i is well-qualified, and facilitates distinction between the valid and imposter speaker classes. In this case of standard test data, more references should be given to the well-estimated GMM likelihood scores.

4 Experiments and results

In this study, speaker recognition is designed to include speaker verification and speaker identification. The utterance from the test speaker is first evaluated for its validity and effectiveness in speaker verification processing, and then sent to speaker identification processing for an identity decision. The speaker recognition experiments were designed using an access control system

application in which the test speaker was requested to speak his or her name as the access key. Speaker recognition experiments contain two main phases: the training phase, in which SVM and GMM classification models are established, and the recognition phase for the performance evaluation of the proposed EGMM-SVM.

All the speech data were recorded in an office with a close-talking microphone. The speech signal was sampled at 44.1 kHz and recorded on the mono channel with 8-bit resolution. The analysis frames were 20-ms wide with a 10-ms overlap. For each frame, a 10-dimensional feature vector was extracted. The feature vector for each frame was a 10-dimensional cepstral vector.

The training data were collected from 27 male speakers. During speaker verification, 13 speakers were chosen as the valid speakers and 14 speakers were chosen as imposters. Each of the 27 speakers was asked to offer 20 utterances of his or her name in Mandarin as the training data for establishing the SVM. Training this SVM separation hyperplane involved 540 training utterances. In the speaker identification phase, the same 540 training utterances from the same 27 speakers in SVM training were used for GMM establishments. Twenty-seven GMM speaker models were trained, each of which represented the corresponding identity of the speaker.

In the recognition and test phase, each of the 27 speakers in the training phase was again requested to provide an additional 20 utterances of his or her name in Mandarin as test data, which were divided into 27 test databases, DB–1 to DB–27, each of which contained 20 utterances from a specific speaker. Table 1 reveals the comparative recognition accuracy between the conventional GMM without any evaluation scheme for the test data, and the EGMM-SVM with the support of SVM. Note that the parameter C denoting the scaling factor of $d_{SVM}(x_i)$ in Eq. (10) is a fixed constant, and its value is set in an empirical procedure to ensure that $\frac{d_{SVM}(x_i)}{C}$ is not larger than 1. The proposed EGMM-SVM approach in Table 1 shows a clear improvement in recognition performance. The EGMM-SVM achieves an average recognition rate of 87.75 %, which is more efficient than the average recognition rate of 75.5 % in a conventional GMM.

Table 1 A comparison of speaker recognition accuracy between the EGMM-SVM and the conventional GMM

Test data set	Recognition rates (%) Speaker recognition methods	
	EGMM-SVM	Conventional GMM
DB-1	100	90
DB-2	100	100
DB-3	70	40
…	…	…
DB-13	95	70
DB-14	85	50
DB-15	85	50
…	…	…
DB-26	90	90
DB-27	90	80
AVG.	87.75	75.5

5 Conclusion

This paper proposes an EGMM-SVM method to improve the conventional GMM used in web-based speaker recognition applications. The proposed EGMM-SVM is an enhanced-version of the GMM, which considers the quality of the test data by incorporating an SVM classifier when performing GMM likelihood score calculations. EGMM-SVM speaker recognition is a GMM-based speaker identification with the support of SVM speaker verification. Compared with the conventional GMM scheme, which does not consider the appropriateness of the speaker's test data, EGMM-SVM is more comprehensive and achieves more efficient performance in recognition accuracy.

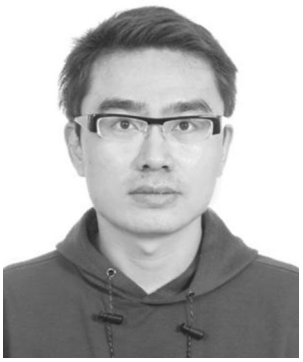
Acknowledgments This research is partially supported by the National Science Council (NSC) in Taiwan under grant NSC 101-2221-E-150-084.

References

1. Bharkad S, Kokare M (2012) Hartley transform based fingerprint matching. J Inf Process Syst 8(1):85–100
2. Boujelbene SZ, Mezghani DBA, Ellouze N (2010) Improving SVM by modifying kernel functions for speaker identification task. Int J Digit Content Technol Appl 4(6):100–105
3. Burges CJC (1998) A tutorial on support vector machines for pattern recognition. Data Min Knowl Disc 2(2):121–167
4. Burget L, Matejka P, Schwarz P, Glembek O, Cernocky J (2007) Analysis of feature extraction and channel compensation in a GMM speaker recognition system. IEEE Trans Audio, Speech, Lang Process 15(7):1979–1986
5. Campbell WM, Campbell JP, Gleason TP, Reynolds DA, Shen W (2007) Speaker verification using support vector machines and high-level features. IEEE Trans Audio, Speech, Lang Process 15(7):2085–2094
6. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc 39:1–38
7. Fan CI, Lin YH (2012) Full privacy minutiae-based fingerprint verification for low-computation devices. J Converg 3(2):21–24
8. Gaikwad SK, Gawali BW, Yannawar P (2010) A review on speech recognition technique. Int J Comput Appl 10(3):16–24
9. Griol D, Molina JM, Corrales V (2011) The VoiceApp system: Speech technologies to access the semantic web. In: CAEPIA 2011. Lecture Notes in Computer Science, vol 7023, pp 393–402
10. Hussain A, Abbasi AR, Afzulpurkar N (2012) Detecting & interpreting self-manipulating hand movements for student's affect prediction. Hum-centric Comput Inf Sci 2(14):1–18
11. Jourani R, Daoudi K, Andre-Obrecht R, Aboutajdine D (2011) Speaker verification using large margin GMM discriminative training. In: Proceedings of International Conference on Multimedia Computing and Systems. Toulouse, France, pp 1–5
12. Kenny P, Boulianne G, Ouellet P, Dumouchel P (2007) Speaker and session variability in GMM-based speaker verification. IEEE Trans Audio, Speech, Lang Process 15(4):1448–1460
13. Linde Y, Buzo A, Gray RM (1980) An algorithm for vector quantizer design. IEEE Trans Commun 28:84–95
14. McLaren M, Vogt R, Baker B, Sridharan S (2010) Data-driven background dataset selection for SVM-based speaker verification. IEEE Trans Audio, Speech, Lang Process 18(6):1496–1506
15. Qian Z, Xu D (2009) Research advances in face recognition. In: Proceedings of IEEE Chinese Conference on Pattern Recognition, pp 1–5
16. Reynolds DA, Rose RC (1995) Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Transactions on Speech and Audio Processing 3(1):72–83
17. Satone MP, Kharate GK (2012) Face recognition based on PCA on wavelet subband of average-half-face. J Inf Process Syst 8(3):483–494
18. You CH, Lee KA, Li H (2009) An SVM kernel with GMM-supervector based on the Bhattacharyya distance for speaker recognition. IEEE Signal Proces Lett 16(1):49–52

19. You CH, Lee KA, Li H (2010) GMM-SVM kernel with a Bhattacharyya-based distance for speaker recognition. IEEE Trans Audio, Speech, Lang Process 18(6):1300–1312
20. Zhang M, Zou KQ (2008) The application of fuzzy clustering after improvement on speaker recognition. ICIC Express Lett 2(3):263–267

Ing-Jr Ding was born in Taipei, Taiwan, in 1975. He received the B.S. degree from Chang-Gung University in 1999, M.S. degree from National Central University in 2001, and Ph.D. degree from National Chiao-Tung University in 2008. He joined the Graduate Institute of Automation and Control at National Taiwan University of Science and Technology as a project assistant professor from March 2009 to July 2009. From August 2009 to July 2012, he served as an assistant professor in the Department of Electrical Engineering, National Formosa University. Since August 2012, he has been an associate professor in the Department of Electrical Engineering, National Formosa University. His research interests include speech recognition, artificial intelligence, and multimedia techniques. He is a member of IEICE.



Chih-Ta Yen was born in Taipei, Taiwan, in January 1974. He received his B.S. degree from the Department of Electrical Engineering at Tamkang University, Taiwan, in 1996, his M. S. degree from the Department of Electrical Engineering, National Taiwan Ocean University, Taiwan, in 2002, and his Ph.D. degree from the Department of Electrical Engineering at National Cheng Kung University, Taiwan, in 2008. He is currently an associate professor in National Formosa University in the area of communication technology at the Department of Electrical Engineering, Yunlin, Taiwan. His major interests are in the areas of multi-user optical communications, wireless communication systems, sensing systems, and satellite communications.