Separated Modeling of Speaker Dependent and Independent Vocal Characteristics for Text-Independent Speaker Verification

Wei-Ho Tsai and Cin-Hao Ma

***Abstract*—This study proposes an approach to model a speaker's voice by assuming that the voice can be decomposed into a speaker-dependent (SD) component and a speaker-independent (SI) component. Using the Universal Background Model (UBM) to approximate the characteristics of the SI component, we derive the SD component by finding the optimal combination of the two models that matches the given speech data. Results of the experiments conducted using the 2001 NIST speaker recognition evaluation corpus demonstrate the superiority of the proposed approach over the conventional speaker verification based on GMM-UBM.**

***Index Terms*—GMM-UBM, speaker dependent, speaker independent, speaker verification.**

I. Introduction

Speaker verification has long been an important research topic [1]-[14]. It is aimed to determine if a speaker is whom he or she claimed to be. Speaker verification mainly can be used in the three fields: authentication, surveillance and forensics. In the applications of authentication, speaker verification is also known as biometric person authentication. It serves as a key to access to a secure system. Compared to the conventional key or credit card that can be stolen or lost, or compared to the password that can be easily misused or forgotten, biometric person authentication is more safe and convenient. In the applications of surveillance, speaker verification can serve as a filter to extract the relevant information on the target person, and thereby provide personalized services. In the forensic context, suspect's voice can be examined through the use of speaker verification, which helps to verify if he/she was talking.

Speaker verification can be divided into two categories, namely text-dependent and text-independent, where the former requires that a speaker says the enrolled texts exactly, while the latter verifies a speaker's identity without constraint on the speech content. Text-independent speaker verification is generally more convenient than text-dependent speaker verification, since users can speak at their wills to the text-independent speaker verification system. However, the problem of text-independent speaker verification is more difficult than that of text-dependent speaker verification by the fact that the former needs to handle the great diversity of speech content.

To date, GMM-UBM (Gaussian Mixture Model - Universal Background Model) [5] is the most prevalent approach to text-independent speaker verification, out of its generalization ability to handle acoustic patterns not covered in a target speaker's training speech. In the GMM-UBM framework, however, since a target speaker's model is derived from a UBM trained using a large number of non-target speakers' speech, it may carry more voice information on non-target speakers than the target speaker. As a tendency, the system could falsely reject a target speaker if the resulting target speaker's model is similar to the UBM. The aim of this work is to derive a target speaker's model that captures voice characteristics more closely related to the target speaker. This is done by assuming that any speaker's voice can be decomposed into speaker-dependent (SD) and speaker-independent (SI) parts. Using the UBM to approximate the generation of the SI part, we model the SD part by finding the optimal combination of the two models that forms the given speech data.

The rest of this paper is organized as follows. Section II details the proposed approach of separated modeling of SD and SI vocal characteristics for text-independent speaker verification. Section III presents our experiments using the 2001 NIST speaker recognition evaluation corpus. In Section IV, we conclude this work and indicate some possible future direction.

II. Methodology

As shown in Fig. 1, it is assumed that a specific person's speech is the result of a speaker-dependent (SD) excitation source passing through a speaker-independent (SI) vocal system, in analogy to the well-known source–filter model of speech production. The SI vocal system is related to some general rules of pronunciation, while the SD excitation source is related to the speaking style of each individual person. Our aim is, thus, to characterize the SD excitation source as a speaker-specific model.

Let $\mathbf{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T\}$ be the cepstrum-based feature vectors extracted from a target person's speech, and $\mathbf{S} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_T\}$ and $\mathbf{V} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_T\}$ be the unobservable counterparts of the SD excitation source and SI vocal system underlying $\mathbf{X}$, respectively. Then, in the cepstral domain, $\boldsymbol{x}_t = \boldsymbol{s}_t + \boldsymbol{v}_t$, $1 \le t \le T$. Suppose that the distributions of $\mathbf{S}$ and $\mathbf{V}$ can be, respectively, represented by GMMs $\lambda_s = \{w_{s,i}, \boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{s,i} \mid 1 \le i \le I\}$, and $\lambda_v = \{w_{v,j}, \boldsymbol{\mu}_{v,j}, \boldsymbol{\Sigma}_{v,j} \mid 1 \le j \le J\}$, where $w_{s,i}$ and $w_{v,j}$

Manuscript received July 9, 2014; revised September 11, 2014. This work was supported in part by the Ministry of Science and Technology, Taiwan under Grant NSC 101-2221-E-027-128-MY2

The authors are with the Department of Electronic Engineering, National Taipei University of Technology, No.1, Sec. 3, Chunghsiao E. Rd. Taipei City, 10608, Taiwan (e-mail: whtsai@ntut.edu.tw, t101419012@ntut.edu.tw).

DOI: 10.7763/IJIEE.2015.V5.515

are mixture weights; $\boldsymbol{\mu}_{s,i}$ and $\boldsymbol{\mu}_{v,j}$ are the mean vectors; and $\boldsymbol{\Sigma}_{s,i}$ and $\boldsymbol{\Sigma}_{v,j}$ are the covariance matrices. The optimal speaker-specific model λ_s in a maximum likelihood sense is

$$\begin{aligned}\lambda_s^* &= \arg\max_{\lambda_s} \Pr(\mathbf{X} \mid \lambda_s, \lambda_v) \\ &= \arg\max_{\lambda_s} \left\{ \prod_{t=1}^{T} \sum_{i=1}^{I} \sum_{j=1}^{J} w_{s,i} w_{v,j} \Pr(\boldsymbol{x}_t \mid i, j) \right\},\end{aligned} \tag{1}$$

where $\Pr(\boldsymbol{x}_t \mid i,j)$ accounts for the possible combination of the SD excitation source and SI vocal system that can form an instant speech feature $\boldsymbol{x}_t$. If we further assume that **S** and **V** are statistically independent, the probability $\Pr(\boldsymbol{x}_t|i,j)$ can be computed by

$$\Pr(\boldsymbol{x}_t \mid i, j) = \int_{-\infty}^{\infty} \mathcal{N}(\boldsymbol{s}; \boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{s,i}) \mathcal{N}(\boldsymbol{x}_t - \boldsymbol{s}; \boldsymbol{\mu}_{v,j}, \boldsymbol{\Sigma}_{v,j}) d\boldsymbol{s}, \tag{2}$$

where $\mathcal{N}(\cdot)$ denotes a Gaussian density function.

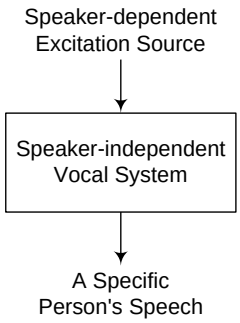

Fig. 1. The proposed model of generating a specific person's speech.

Although **V** is unobservable, its stochastic characteristics could be estimated using speech data from a large number of speakers. Specifically, we can pool all available feature vectors from non-target persons' speech to train λ_v via Expectation-Maximum (EM) algorithm [15], where λ_v is equivalent to the so-called UBM. Given λ_v, Eq. (1) can then be solved using the EM algorithm, which starts with an initial model λ_s and iteratively estimates a new model λ_s' such that $\Pr(\mathbf{X}|\lambda_s',\lambda_v) \geq \Pr(\mathbf{X}|\lambda_s,\lambda_v)$. The goal of increasing the probability $\Pr(\mathbf{X}|\lambda_s',\lambda_v)$ can be achieved by maximizing the auxiliary function

$$Q(\lambda_s, \lambda_s') = \sum_{t=1}^{T} \sum_{i=1}^{I} \sum_{j=1}^{J} \Pr(i, j \mid \boldsymbol{x}_t, \lambda_s, \lambda_v) \log \Pr(i, j, \boldsymbol{x}_t \mid \lambda_s', \lambda_v), \tag{3}$$

where

$$\Pr(i, j, \boldsymbol{x}_t \mid \lambda_s', \lambda_v) = w_{s,i}' w_{v,j} \Pr(\boldsymbol{x}_t \mid i, j), \tag{4}$$

and

$$\Pr(i, j \mid \boldsymbol{x}_t, \lambda_s, \lambda_v) = \frac{w_{s,i} w_{v,j} \Pr(\boldsymbol{x}_t \mid i, j)}{\sum_{m=1}^{I} \sum_{n=1}^{J} w_{s,m} w_{v,n} \Pr(\boldsymbol{x}_t \mid m, n)}. \tag{5}$$

Letting $\nabla Q(\lambda_s,\lambda_s') = 0$ with respect to each parameter to be re-estimated, we can show that [16]

$$w_{s,i}' = \frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{J} \Pr(i, j \mid \boldsymbol{x}_t, \lambda_s, \lambda_v), \tag{6}$$

$$\boldsymbol{\mu}_{s,i}' = \frac{\sum_{t=1}^{T} \sum_{j=1}^{J} \Pr(i, j \mid \boldsymbol{x}_t, \lambda_s, \lambda_v) \cdot E\{\boldsymbol{s}_t \mid \boldsymbol{x}_t, i, j\}}{\sum_{t=1}^{T} \sum_{j=1}^{J} \Pr(i, j \mid \boldsymbol{x}_t, \lambda_s, \lambda_v)}, \tag{7}$$

$$\boldsymbol{\Sigma}_{s,i}' = \frac{\sum_{t=1}^{T} \sum_{j=1}^{J} \Pr(i, j \mid \boldsymbol{x}_t, \lambda_s, \lambda_v) \cdot E\{\boldsymbol{s}_t \boldsymbol{s}_t^{\mathrm{Tr}} \mid \boldsymbol{x}_t, i, j\}}{\sum_{t=1}^{T} \sum_{j=1}^{J} \Pr(i, j \mid \boldsymbol{x}_t, \lambda_s, \lambda_v)} - \boldsymbol{\mu}_{s,i} \boldsymbol{\mu}_{s,i}^{\mathrm{Tr}}, \tag{8}$$

where Tr denotes the transpose, and

$$E\{\boldsymbol{s}_t \mid \boldsymbol{x}_t, i, j\}$$

$$= \frac{\int_{-\infty}^{\infty} \boldsymbol{s} \mathcal{N}(\boldsymbol{s}; \boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{s,i}) \mathcal{N}(\boldsymbol{x}_t - \boldsymbol{s}; \boldsymbol{\mu}_{v,j}, \boldsymbol{\Sigma}_{v,j}) d\boldsymbol{s}}{\Pr(\boldsymbol{x}_t \mid \boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{s,i}, \boldsymbol{\mu}_{v,j}, \boldsymbol{\Sigma}_{v,j})} \tag{9}$$

$$= \left(\boldsymbol{\Sigma}_{s,i}^{-1} + \boldsymbol{\Sigma}_{v,j}^{-1}\right)^{-1} \left[\boldsymbol{\Sigma}_{v,j}^{-1} (\boldsymbol{x}_t - \boldsymbol{\mu}_{v,j}) + \boldsymbol{\Sigma}_{s,i}^{-1} \boldsymbol{\mu}_{s,i}\right],$$

$$E\{\boldsymbol{s}_t \boldsymbol{s}_t^{\mathrm{Tr}} \mid \boldsymbol{x}_t, i, j\}$$

$$= \frac{\int_{-\infty}^{\infty} \boldsymbol{s} \mathcal{N}(\boldsymbol{s}; \boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{s,i}) \mathcal{N}(\boldsymbol{x}_t - \boldsymbol{s}; \boldsymbol{\mu}_{v,j}, \boldsymbol{\Sigma}_{v,j}) d\boldsymbol{s}}{\Pr(\boldsymbol{x}_t \mid \boldsymbol{\mu}_{s,i}, \boldsymbol{\Sigma}_{s,i}, \boldsymbol{\mu}_{v,j}, \boldsymbol{\Sigma}_{v,j})} \tag{10}$$

$$= E\{\boldsymbol{s}_t \mid \boldsymbol{x}_t, i, j\} E\{\boldsymbol{s}_t^{\mathrm{Tr}} \mid \boldsymbol{x}_t, i, j\} + \left(\boldsymbol{\Sigma}_{s,i}^{-1} + \boldsymbol{\Sigma}_{v,j}^{-1}\right)^{-1}.$$

When testing an unknown speech utterance, the system determines whether or not it is produced by the claimed speaker using

$$\log \Pr(\mathbf{Y} \mid \lambda_s, \lambda_v) - \log \Pr(\mathbf{Y} \mid \lambda_v) \underset{\text{No}}{\overset{\text{Yes}}{\gtrless}} \theta, \tag{11}$$

where **Y** is the utterance's cepstrum-based feature vectors, and θ is a pre-set threshold. Fig. 2 shows the block diagram of the proposed speaker-verification system.

III. Experiments

A. Speech Data

Our experiments were conducted using the *2001 NIST speaker recognition evaluation corpus* [17]. The corpus contains 174 and 60 speakers in the defined "evaluation set" and "development set", respectively. The evaluation set was further divided into two subsets, one for training and one for testing. The training subset contains 2-min speech recorded from each speaker, and the testing subset contains 2038

speech utterances. Each utterance was evaluated against one target speaker and ten imposter speakers. Thus, there were a total of 2038 positive trials and 20380 negative trials. The development set, composed of 2-min speech per speaker, was used to create the UBM. Speech feature vectors, each consists of 15 Mel-Frequency Cepstral Coefficients (MFCCs) [18] were computed for every 16 ms using a 32 ms Hamming window. In addition, cepstral mean subtraction was applied to the MFCCs to reduce channel effects.

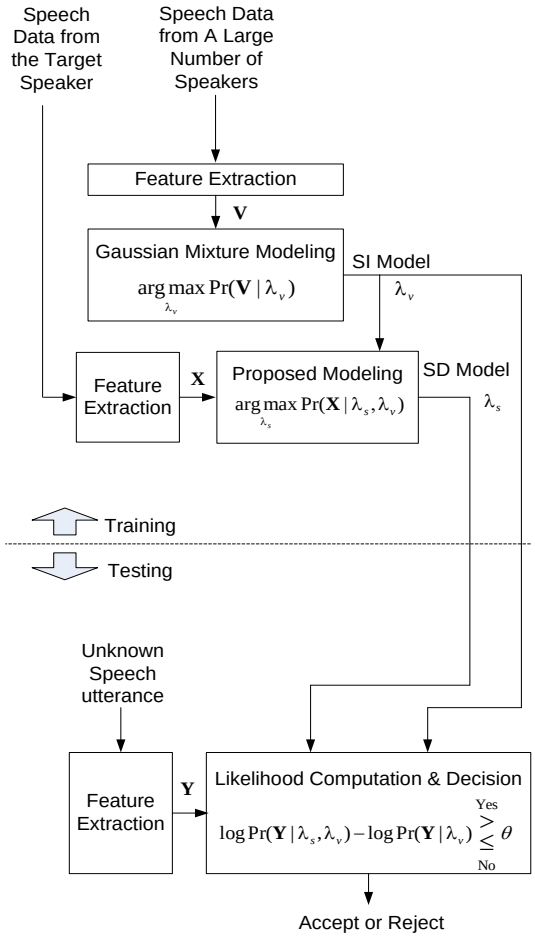

Fig. 2. Block diagram of the proposed speaker-verification system.

B. Experiment Results

First, the performance of the proposed system was assessed with respect to various numbers of mixture densities used in λ_s and λ_v. Table I shows the resulting minimum values of the detection cost function (DCF) evaluated according to [17]. We can see from Table I that the best performance of our system was achieved when using 32 and 256 mixture densities in λ_s and λ_v, respectively.

TABLE I: MINIMUM VALUES OF DCF OBTAINED WITH THE PROPOSED SYSTEM

No. of Mixture Densities in λ_v	No. of Mixture Densities in λ_s				
	16	32	64	128	256
64	0.0496	0.0438	0.0493	0.0496	0.0511
128	0.0483	0.0432	0.0491	0.0490	0.0504
256	0.0450	**0.0421**	0.0484	0.0488	0.0493
512	0.0466	0.0429	0.0481	0.0476	0.0471

Then, we compared our system with the GMM-UBM system. Following [7], the number of mixture densities used in the GMM-UBM system was set to 1024, and only mean vectors were adjusted during the adaptation. Fig. 3 shows the DET curves obtained with the GMM-UBM system and our system using 32 and 256 mixture densities in λ_s and λ_v, respectively. It is clear from Fig. 3 that the proposed system consistently yields lower false alarms probability and miss probability than those of the GMM-UBM system. In particular, our system significantly reduces the probability of falsely rejecting a target speaker. This confirms the superiority of the proposed system over the GMM-UBM system.

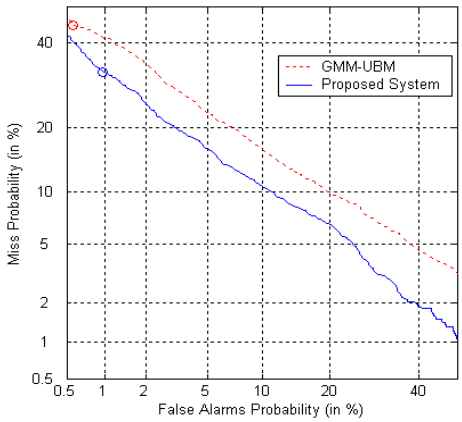

Fig. 3. DET curves obtained with the GMM-UBM system and our system using 32 and 256 mixture densities in λ_s and λ_v, respectively.

IV. CONCLUSION

In this study, we have assumed that each speaker's voice can be decomposed into an SD part and an SI part and used the UBM to characterize the SI part. We have also derived the formulae to model the SD part by finding the optimal combination of the two models that forms the specific speaker's voice. Our experiments conducted using the 2001 NIST speaker recognition evaluation corpus demonstrate the superiority of our approach over the popular one based on GMM-UBM.

Although the proposed separated modeling of SI and SD vocal characteristics for text-independent speaker verification is feasible, it does not consider the cases that speech utterances are corrupted by noises and channel distortion. Thus, in the future, we will try to derive the formulae to model the SD part with the integration of environmental factors. A number of techniques [19]-[25] could be integrated into our formulation.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers, for their careful reading of this paper and their constructive suggestions.

REFERENCES

[1] A. E. Rosenberg, "Automatic speaker verification: A review," *Proceedings of IEEE*, vol. 64, no. 4, pp. 475-487, 1976.
[2] G. R. Doddington, "Speaker recognition—Identifying people by their voices," *Proceedings of IEEE*, vol. 73, no. 11, pp. 1651-1664, 1985.

[3] A. Higgins, L. Bahler, and J. Porter, "Speaker verification using randomized phrase prompting," *Digital Signal Processing*, vol. 1, no. 2, pp. 89-106, 1991.
[4] D. A. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech Communication*, vol. 17, pp. 91-108, 1995.
[5] C. S. Liu, H. C. Wang, and C. H. Lee, "Speaker verification using normalized log-likelihood score," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 1, pp. 56-60, 1996.
[6] Q. Li, B. H. Juang, C. H. Lee, Q. Zhou, and F. K. Song, "Recent advancements in automatic speaker authentication," *IEEE Robotics & Automation Magazine*, vol. 6, no.1, pp. 24-34, 1999.
[7] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, 2000.
[8] D. A. van Leeuwen, A. F. Martin, M. A. Przybocki, and J. S. Bouten, "NIST and NFI-TNO evaluations of automatic speaker recognition," *Computer Speech and Language*, vol. 20, pp. 128-158, 2006.
[9] Y. Mami and D. Charlet, "Speaker recognition by location in the space of reference speakers," *Speech Communication*, vol. 48, pp. 127-141, 2006.
[10] W. M. Campbell, J. P. Campbell, T. P. Gleason, D. A. Reynolds, and W. Shen, "Speaker verification using support vector machines and high-level features," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2085-2094, 2007.
[11] M. A. Przybocki, A. F. Martin, and A. N. Le, "NIST speaker recognition evaluations utilizing the mixer corpora—2004, 2005, 2006," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 1951-1959, 2007.
[12] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.
[13] M. S. Chavan and S. V. Chougule, "Speaker features and recognition techniques: A review," *International Journal of Computational Engineering*, vol. 2, no. 3, pp. 720-728, 2012.
[14] X. Anguera Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356-370, 2012.
[15] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1-38, 1977.
[16] R. C. Rose, E. M. Hofstetter, and D. A. Reynolds, "Integrated models of signal and background with application to speaker identification in noise," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 245–257, 1994.
[17] Itl. [Online] Available: http://www.itl.nist.gov/iad/mig/tests/spk/2001/
[18] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, speech, and Signal Processing*, vol. 28, pp. 357-366, 1980.
[19] M. Ji, T. J. Hazen, J. R. Glass, and D. A. Reynolds, "Robust speaker recognition in noisy conditions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1711-1723, 2007.
[20] A. Panda and T. Srikanthan, "Psychoacoustic model compensation for robust speaker verification in environmental noise," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 945-953, 2012.
[21] X. Zhao and D. Yuan, "Variational bayesian joint factor analysis models for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3,pp. 1032-1042, 2012.
[22] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435-1447, 2007.
[23] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification,", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788-798,2011.
[24] R. Wei and M. W. Mak "Boosting the performance of I-Vector based speaker verification via utterance partitioning," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 1012-1022, 2013.
[25] T. Hasan and J. H. L. Hansen, "Maximum likelihood acoustic factor analysis models for robust speaker verification in noise," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 381-391, 2014.

Wei-Ho Tsai received his B.S. degree in electrical engineering from National Sun Yat-Sen University, Kaohsiung, Taiwan, in 1995. He received his M.S. and Ph.D. degrees in communication engineering from National Chiao-Tung University, Hsinchu, Taiwan, in 1997 and 2001, respectively. From 2001 to 2003, he was with Philips Research East Asia, Taipei, Taiwan, where he worked on speech processing problems in embedded systems. From 2003 to 2005, he served as a postdoctoral fellow at the Institute of Information Science, Academia Sinica, Taipei, Taiwan. He is currently a professor in the Department of Electronic Engineering & Graduate Institute of Computer and Communication Engineering, National Taipei University of Technology, Taiwan. His research interests include spoken language processing and music information retrieval. Dr. Tsai is a life member of ACLCLP and a member of IEEE.

Cin-Hao Ma received the B.S degree in electronic engineering from National Taipei University of Technology, Taipei, Taiwan, in 2012. He is pursuing the Ph.D. degree in computer and communication engineering at National Taipei University of Technology currently. His research interests include signal processing and multimedia applications.