SIP (2016), vol. 5, e10, page 1 of 9 © The Authors, 2016.

doi:10.1017/ATSIP.2016.10

ORIGINAL PAPER

Ensemble based speaker recognition using unsupervised data selection

CHIEN-LIN HUANG[1], JIA-CHING WANG[1] AND BIN MA[2]

This paper presents an ensemble-based speaker recognition using unsupervised data selection. Ensemble learning is a type of machine learning that applies a combination of several weak learners to achieve an improved performance than a single learner. A speech utterance is divided into several subsets based on its acoustic characteristics using unsupervised data selection methods. The ensemble classifiers are then trained with these non-overlapping subsets of speech data to improve the recognition accuracy. This new approach has two advantages. First, without any auxiliary information, we use ensemble classifiers based on unsupervised data selection to make use of different acoustic characteristics of speech data. Second, in ensemble classifiers, we apply the divide-and-conquer strategy to avoid a local optimization in the training of a single classifier. Our experiments on the 2010 and 2008 NIST Speaker Recognition Evaluation datasets show that using ensemble classifiers yields a significant performance gain.

Keywords: Speaker recognition, Ensemble classifier, Unsupervised data selection

Received 19 May 2015; accepted 12 April 2016

I. INTRODUCTION

Nowadays, the demand continues to increase for speaker recognition technology in such applications as telephony, security, and communication. For example, the application of voice mining is used to monitor the communications, which is popularly adopted by intelligence agencies, government and law Enforcement. Speaker recognition is a kind of biometric verification such as fingerprint, iris, and face recognition. The major components of speaker recognition, which finds the identity information of a speaker from speech signals, include feature analysis, statistical modeling, and verification decision.

Most speaker recognition systems use cepstrum-based features such as Mel-frequency cepstral coefficients (MFCC) [1] or perceptual linear prediction [2] cepstral coefficients, which provide an estimate of short-term energy as a function of frequency. Gaussian mixture model (GMM) has been commonly applied for statistical modeling in speaker recognition applications with speaker adaptation techniques. To solve speaker data sparseness and channel mismatch problems, maximum *a posteriori* (MAP) has been widely used to adapt the speaker model from the universal background model (UBM) [3]. To compensate the channel and session effects, eigenchannel is applied in speaker recognition [4]. Recently, i-vector technique is proposed to estimate total variability for speaker adaptation [5].

Different from speech recognition with HMM modeling [6], the common speaker recognition methods are based on GMM framework. The advantage of the GMM-based approach is that speaker recognition can be performed in a completely text-independent manner [7] and all speech frames without any transcription and segmentation are used to estimate speaker information and build GMMs. However, one disadvantage of such a GMM modeling approach is that the acoustic variability of phonetic events is not taken into account during comparisons with different speakers [7]. To solve this problem, many previous studies focused on using specific constrained groups of data to improve the speaker recognition performance.

A) Related works in ensemble-based speaker recognition

The generalization ability of an ensemble could be significantly better than that of a single learner. Zhang et al. intended to improve the performance of the speaker recognition system by introducing a novel method combining optimizing annular region-weighted distance k-nearest neighbor with BagWithProb ensemble learning schemes [8]. In the DataBoost-UP algorithm, the data (i-vectors) is synthesized using the utterance partitioning technique

[1]Department of Computer Science and Information Engineering, National Central University, Taiwan 32001, Republic of China
[2]Human Language Technology, Institute for Infocomm Research (I2R), Singapore 138632, Singapore

Corresponding author:
C.-L. Huang
Email: chiccocl@gmail.com

instead of random generation of attribute values in the minimum and maximum interval. Both the minority (target speaker) and majority (background speakers) classes are oversampled to prevent overemphasis on the hard instances of the minority class. The DataBoost-UP is used to create an ensemble of SVM classifiers [9]. Sturim et al. presented text-constrained Gaussian mixture models to close the gap between text-dependent and text-independent speaker verification. Speech is segmented into acoustic units such as words or phones, and then GMM–UBM verifiers are trained and tested using only speech from constrained groups of units [10]. Park and Hazen proposed speaker identification using domain-dependent automatic speech recognition (ASR) to provide phonetic segmentation. A combination of classifiers is used to reduce identification errors [7]. Baker et al. studied GMM modeling using multilingual broad phonetics to construct syllabic events and segmentations for speaker verification [11]. Bocklet and Shriberg described a speaker recognition approach using syllable-level constraints for cepstral frame selection. Complementary information and improvement can be found by combining eight subsystems including syllable onsets, syllable nucleii, syllable codas, syllables following pauses, one-syllable words, and three other kinds of syllables [12]. Sanchez et al. studied the performances between constraint-dependent and constraint-independent approaches for training UBMs and joint factor analysis. They explored unit-based constraints, which are regions constrained by specific syllables, phones, or sub-phone regions [13]. In addition, unsupervised clustering was applied to speaker recognition to compensate the domain mismatch between training, enrollment, and testing data in [14]. Attempts of ensemble of speaker recognition systems have been made in [15].

All of the above work segmented and selected data for more detailed speaker model construction based on prosody, syllable, or phoneme analysis. Although these approaches showed improvements in speaker recognition, many shortcomings remain in them. For example, the quality of the feature frame selection is obviously influenced by the accuracy of ASR or prosody estimation systems. Furthermore, prior or auxiliary knowledge is required for such constrain-based approaches as language information. According to these reasons, we do not have experimental comparisons. Although there is no comparison with the existing work on the ensemble of speaker recognition, the performance of the proposed method is consistently better than the baseline.

B) Proposed framework

In this study, we propose an ensemble learning using unsupervised data selection, which considers acoustic variability in the model training, speaker enrollment and testing. The speech data are segmented into several subsets of speech frames without any auxiliary information or pre-processor (ASR or prosody estimator systems) and then ensemble classifiers are trained using these subsets in a divide-and-conquer manner. The ensemble framework is similar to neural networks or mixture of experts [16]. In such a way, we can avoid the local optimization training when a single conventional classifier is adopted.

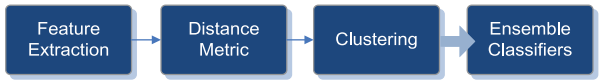

Fig. 1. The pipeline of the proposed ensemble based speaker recognition using unsupervised data selection.

Figure 1 shows the pipeline of the proposed ensemble-based speaker recognition using unsupervised data selection. Basically, there are three elements before we do ensemble training and testing. First, at the feature extraction stage, we aim at extracting discriminative and effective acoustic features by applying long-term feature (LTF) analysis. Second, at the distance metric stage, we explore two categories of distance metrics, including vector-based and likelihood-based distance metrics, to measure the similarity between data. Finally, the clustering algorithm can be naturally employed at the clustering stage. We conducted experiments on the 2010 and 2008 NIST Speaker Recognition Evaluation (SRE) datasets.

C) Outline of the paper

The rest of this paper is organized as follows. In Sections II–IV, the pipeline of the proposed method, namely, feature extraction, distance metric, and clustering for ensemble-based speaker recognition, are described. In Section V, we describe our experiment setup and protocol, and introduce the performance evaluation metrics. We present the experiment results as well as a discussion of the results in Section VI. Finally, we conclude this work in Section VII.

II. FEATURE EXTRACTION

At the first stage of ensemble-based speaker recognition pipeline is feature extraction. Feature extraction is an important process to estimate a numerical representation from speech samples and to characterize the speakers. Many kinds of feature analysis have been proposed for speaker recognition in previous studies. The conventional short-term spectral features, such as MFCC, are useful acoustic features for speaker recognition. Many efforts have been devoted to improving the effectiveness of MFCC, such as reducing the dimensionality, enhancing discriminative ability [17], and characterizing speakers with temporal features [18]. Due to the importance of phase in human speech, features are extracted by integrating MFCC and phase information for speaker identification and verification [19]. In deep neural network (DNN) speech recognition [20], experiments show the gain of DNN is almost entirely attributed to DNN's feature vectors that are concatenated from several consecutive speech frames within a relatively long context window. In this study, we aim at extracting discriminative and effective acoustic features for speaker recognition, by applying LTF analysis to enhance

https://doi.org/10.1017/ATSIP.2016.10 Published online by Cambridge University Press

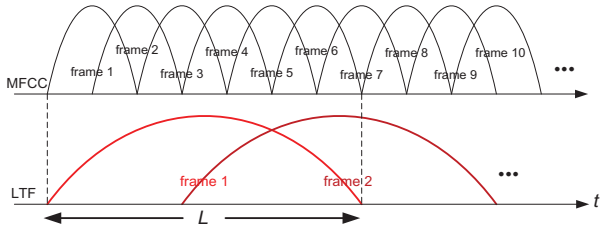

Fig. 2. Illustration of speaker discriminative feature analysis using the mean of short-term spectral features in a long-term window.

the discriminative capability of short-term spectral features as shown in Fig. 2.

We applied LTF analysis [21] as the feature extraction based on the traditional MFCCs of a short-time spectral analysis of 16 ms. We extracted 36 MFCCs consisting of 12 coefficients in addition to the first and second derivatives. Speech signals were divided into 18 sub-bands between 250 and 3500 Hz using the Mel-filter bank to make spectral contents that resemble those of telephone channels. LTF is used to average several short-time spectral features in a long-time window and capture the spectral statistics over a long period of time. The overlapping long-term windows are applied on the short-term features, reducing short-term MFCC frames J to LTF frames K, with $K = (J - L)/Z + 1$. L denotes the size of the long-term window and Z is the step of the long-term window shift. Since the mean of multiple short-term spectral features is used, LTF can simultaneously take account of short-term frequency characteristics and long-term resolution. This transformation results in a more compact feature vector for statistical modeling. According to the previous study [21], the optimal values of L and Z were 4 and 2, respectively.

III. DISTANCE METRIC

The second stage of ensemble-based speaker recognition pipeline is distance metric calculation. The distance metric calculation of ensemble-based speaker recognition is similar to the speaker diarization scheme [22, 23]. The similarity of between them is to search for homogeneous segments. The differences between them are purposes. In speaker diarization, speaker segmentation is applied to extract the longest possible homogenous segments in a conversation. In ensemble-based speaker recognition, the distance metric calculation is used to measure similarity of short feature frames in speech of a speaker. The distance metrics are used for acoustic clustering of the speech data. We explore two distance metrics, the vector-based and likelihood-based distance metrics, to measure the similarity and construct partitioning clusters for ensemble learning.

A) Vector-based distance metrics

In this study, we use the LTF to analysis acoustic characteristics on the longer range. A feature frame can be viewed as a data point in an n-dimensional vector space. The data points with similar acoustic characteristics tend to cluster together. Thus, Euclidean and Mahalanobis distance metrics are reasonable solutions for the clustering. We applied Euclidean distance to measure the length of the path connecting two feature vectors. The Euclidean distance between vectors $\mathbf{x}$ and $\mathbf{y}$ in an n-dimensional space is given by

$$d(\mathbf{x},\mathbf{y})_{Euc} = \sqrt{\sum_{i=1}^{n} |\mathbf{x}_i - \mathbf{y}_i|^2}. \quad (1)$$

The other common distance measure is the Mahalanobis distance metric. The Mahalanobis distance metric considers correlations of data, and thus the similarity is estimated by

$$d(\mathbf{x},\mathbf{y})_{Mah} = \sqrt{(\mathbf{x}-\mathbf{y})^T \mathbf{A}^{-1}(\mathbf{x}-\mathbf{y})}, \quad (2)$$

where $\mathbf{A}$ is the covariance matrix. In addition, the cosine measure is a type of vector-based distance metric used to estimate the similarity between vectors $\mathbf{x}$ and $\mathbf{y}$ as

$$d(\mathbf{x},\mathbf{y})_{Cos} = \frac{\sum_{i=1}^{n} \mathbf{x}_i \times \mathbf{y}_i}{\sqrt{\sum_{i=1}^{n} \mathbf{x}_i^2} \times \sqrt{\sum_{i=1}^{n} \mathbf{y}_i^2}}, \quad (3)$$

where n is the dimension of the feature vector. The cosine distance is suitable to measure the similarity between the data points with strong directional scattering patterns. For instance, the cosine distance is popularly used on the applications of information retrieval [24, 25] and i-vector-based speaker recognition [5].

B) Likelihood-based distance metric

Besides the vector-based distance metrics, we can also use the likelihood estimation for the similarity measure. We explore two likelihood-based similarity measures. One is the log-likelihood distance metric. The other is delta-Bayesian information criterion (BIC) estimation. We treat each cluster as a Gaussian model $\lambda = \{\mathbf{u}, \mathbf{\Sigma}\}$ in the log-likelihood estimation. The log-likelihood score is estimated by

$$\log(L(\mathbf{x}|\lambda_k)) = \log\left(\frac{1}{(2\pi)^{n/2}|\Sigma_k|^{n/2}} e^{-1/2(\mathbf{x}-\mathbf{u}_k)^T \Sigma_k^{-1}(\mathbf{x}-\mathbf{u}_k)}\right), \quad (4)$$

where $L(\mathbf{x}|\lambda_k)$ is the likelihood of acoustic feature $\mathbf{x}$ given the model λ_k. The mean vector $\mathbf{u}_k \in \Re^n$ and the covariance matrix $\mathbf{\Sigma}_k \in \Re^n$ are applied for each Gaussian; n is the dimension of acoustic feature vector $\mathbf{x}$ and k is the label of the cluster.

The other likelihood-based similarity measurement is the BIC which can be used for speaker clustering [26]. The BIC value shows how well the data $\mathbf{x}$ fit the model λ_k estimated by

$$BIC(\lambda_k) = \log(L(\mathbf{x}|\lambda_k)) - \frac{\varepsilon}{2}\delta_k \log(n_x), \quad (5)$$

where ε is a design parameter, δ_k is the number of free parameters in λ_k, and n_x is the number of feature vectors in $\mathbf{x}$. The similarity between data $\mathbf{x}$ and $\mathbf{y}$ is given by

https://doi.org/10.1017/ATSIP.2016.10 Published online by Cambridge University Press

the delta-BIC score. The delta-BIC score is widely used for audio segmentation, model selection, and speaker clustering [27, 28]. Based on Gaussian assumption, the delta-BIC score between $\mathbf{x}$ and $\mathbf{y}$ is estimated by

$$d(\mathbf{x},\mathbf{y})_{Delta} = N\log|\boldsymbol{\Sigma}| - N_x\log|\boldsymbol{\Sigma}_x| - N_y\log|\boldsymbol{\Sigma}_y| - \varepsilon P, \quad (6)$$

where $N = N_x + N_y$ is the total number of frames. $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_y$ represent the covariance matrices of $\mathbf{x}$ and $\mathbf{y}$, respectively. $\boldsymbol{\Sigma}$ is the covariance matrix of the aggregate of $\mathbf{x}$ and $\mathbf{y}$. P is a penalty factor given by

$$P = \frac{1}{2}\left(n + \frac{1}{2}n(n+1)\right)\log N \quad (7)$$

with different penalty factors, we can perform various model selection criterions such as AIC and MDL [27].

IV. CLUSTERING FOR ENSEMBLE BASED SPEAKER RECOGNITION SYSTEMS

The last stage of ensemble-based speaker recognition pipeline is clustering. In this study, we investigate two clustering algorithms for the unsupervised data selection and the combination of ensemble classifiers.

A) Unsupervised clustering

The unsupervised data selection can be achieved in various ways. We explore two data clustering algorithms based on the partitioning and hierarchical techniques in this study. One popular partitioning technique is the K-means clustering algorithm that partitions data into K clusters in which each data belongs to the cluster with the nearest mean. The K-means clustering algorithm aims to assign every speech frames in a cluster to its respective acoustic characteristics. For example, we can find that the gender information is identified if we set the number of clusters is two. We implement the K-means clustering algorithm with multiple random starting points and an iteratively optimized objective function in this study.

Moreover, we explore the hierarchical technique to build a hierarchy of clusters. There are two strategies for hierarchical clustering. One is agglomerative and the other is divisive. The agglomerative hierarchical clustering is a bottom-up manner in which each observation starts on its own cluster. Pairs of clusters are merged and move up the hierarchy. The divisive hierarchical clustering is a top-down manner in which all data start from one cluster. Splits are performed recursively, and data move down the hierarchy. We conducted the divisive hierarchical clustering in this study. To compare with the K-means clustering algorithm, the termination condition of the hierarchical method is specified by the desired number of K clusters.

B) Data normalization and selection

With clustering algorithms, the feature warping [29] is performed using clustered feature vectors. A transformation function $\varphi(.)$ is applied to convert features according to a lookup table. A lookup table is devised so as to map a rank order determined from the sorted cepstral feature elements to a warped feature using the desired warping target distribution. The feature warping is a kind of normalization process used to map a feature stream to a standard normal distribution. This process effectively Gaussianises the distribution of selected feature vectors so as to better fit to Gaussian assumptions in the model training and testing. The similar technique such as histogram equalization (HEQ) is commonly used in image processing and speech recognition [30, 31].

For the training of ensemble clusters of the unsupervised data selection, the UBM training dataset is utilized. The created clusters are then used to split the following data into subsets: UBM training, score normalization, speaker enrollment, and testing. The ensemble-based speaker recognition systems are trained and tested with the corresponding subsets. Because the selection of number K may lead to a data sparsity problem in the training and the testing of speaker recognition, we study different numbers of K in our experiments.

C) Combination of ensemble classifiers

We usually consult several experts before making an important decision in daily life. Ensemble-based systems weigh several opinions and combine them to reach a final decision instead of a single-expert system [32, 33]. Figure 3

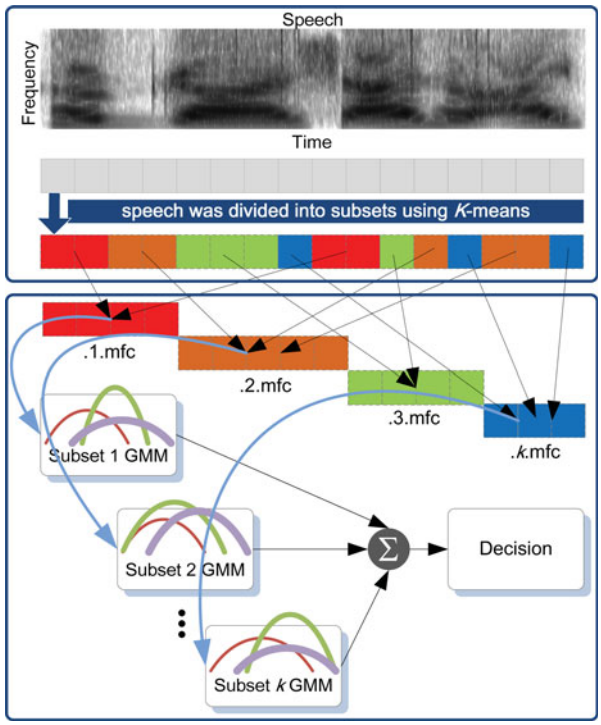

Fig. 3. Testing procedure of ensemble classifiers using unsupervised data selection.

https://doi.org/10.1017/ATSIP.2016.10 Published online by Cambridge University Press

illustrates the proposed ensemble classifiers for speaker recognition based on the divide-and-conquer strategy. The original speech data are segmented into several data subsets from which ensemble-based speaker recognition systems are trained and tested by non-overlapping segmentations.

We consider two factors for building the ensemble-based speaker recognition. One is to cluster and select data based on acoustic variability. The other is to combine the results of ensemble classifiers. In this study, the frame counts (FCs) of the subsets are used as the weights for a combination of ensemble classifiers. With conventional GMM–UBM architecture, the speaker recognition decision is based on the log-likelihood ratio (LLR) between target speaker GMM λ_{SPK} and UBM λ_{UBM}.

$$\Lambda = \frac{1}{N}\sum_{t=1}^{N}[\log p(x_t|\lambda_{SPK}) - \log p(x_t|\lambda_{UBM})], \quad (8)$$

where N means the total frames. If the score exceeds threshold $\Lambda > \theta$, then the claimed speaker will be accepted, or else rejected. To exploit the ensemble classifiers in the GMM–UBM architecture, the proposed LLR score $\tilde{\Lambda}$ considering the FC is then estimated as follows:

$$\tilde{\Lambda} = \frac{1}{N}\sum_{k=1}^{K} n_k(X) \times \left[\log p_k(X|\lambda_{SPK_k}) - \log p_k(X|\lambda_{UBM_k})\right], \quad (9)$$

where $n_k(X)$ is the number of frames in classifier k and satisfies $\sum_{k=1}^{K} n_k(X)/N = 1$. In other words, the contribution of ensemble classifier k is zero if the FC $n_k(X)$ is zero. Equations (8) and (9) indicate that LLR was calculated only on the test data. x_t in equation (8) and X in equation (9) represent the test data. Given the test data X and subset k GMM, we can estimate the likelihood, $p_k(X|\lambda_{UBM_k})$, and then know $n_k(X)$. Base on the same idea, in the ensemble method of i-vector, cosine scores of subsets are combined with the average weighted sum which considering the FC.

V. EXPERIMENT PROTOCOL

The NIST SRE data were collected from different types of channel as telephones and microphones. We evaluated the system on the core condition of the 2010 NIST SRE in the tel–tel condition (det5) [34]. In this section, we apply three speaker recognition systems based on MAP, eigenchannel, and i-vector for evaluating the proposed approach.

A) Baseline systems

The NIST SRE-2004, SRE-2005, and SRE-2006 one-side data were used to train gender-dependent UBMs. The speaker adaptation techniques are used to solve speaker data sparseness and channel mismatch problems. MAP [3] is a popular approach to adapt speaker model from UBM. To further consider various channel factors, the eigenchannel adaptation [4] provides a good solution for channel mismatch. The eigenchannel assumes the means of the speaker's model are given by $\mathbf{m}_{SPK} = \mathbf{m}_{UBM} + \mathbf{Uh}$, while $\mathbf{m}_{UBM}$ denotes the supervector of the concatenation of UBM means, $\mathbf{U}$ is a rectangular low-rank matrix in which the columns are the numbers of directions of channel variability, and $\mathbf{h}$ is a normally distributed random vector that is learned from samples. The SRE-2004, SRE-2005, and SRE-2006 data were used to derive the eigenchannel estimation. The channel factor was set to 40 in this study.

The fast-scoring technique was applied by approximating likelihood values using the top five mixture components [35]. The outputs of MAP and eigenchannel systems were normalized with ZT-norm to further compensate for the nuisance effects, in which T-norm [36] is first applied and then Z-norm [37] speaker models are tested by imposters' speech utterances. With T-norm, the input test speech utterance is evaluated against cohort models to obtain normalization scores using mean and standard deviation. With Z-norm, a speaker model is tested against imposter speech utterances to obtain the mean and standard deviation scores of normalization. For run-time efficiency, Z-norm can be estimated in an offline mode. In this study, 50 speakers are randomly selected from the NIST SRE-2004, SRE-2005, and SRE-2006 one-side data for Z-norm and non-overlapped 50 speakers for T-norm.

Furthermore, the i-vector system has become one of the state-of-the-art techniques in speaker verification applications [5]. The i-vector estimation assumes the speaker and channel-dependent GMM mean supervector is given by $\mathbf{m}_{SPK} = \mathbf{m}_{UBM} + \mathbf{Tw}$, while $\mathbf{T}$ is a rectangular low-rank matrix representing R bases spanning subspace with important variability in the GMM mean supervector space, and $\mathbf{w}$ is a normally distributed random vector of size R that is learned from the samples. We termed the vector weighting $\mathbf{w}$ as i-vector and selected the dimension $R = 200$ for the speaker recognition evaluation. To minimize the effect of within-speaker covariances, we applied the within-class covariance normalization (WCCN) transform in i-vector space to find the transformed vector $\hat{\mathrm{w}} = B^T\mathbf{w}$. The transform matrix $\mathbf{B}$ is derived from the Cholesky decomposition of $\mathbf{W} = \mathbf{BB}^T$, where $\mathbf{w}$ is the within-speaker covariance matrix estimated by

$$\mathbf{W} = \frac{1}{S}\sum_{s=1}^{S}\sum_{i=1}^{N_s}(\mathbf{x}_i^s - \mathbf{u}_s)(\mathbf{x}_i^s - \mathbf{u}_s)^T \quad \mathbf{u}_s = \frac{1}{N_s}\sum_{i=1}^{N_s}\mathbf{x}_i^s, \quad (10)$$

where S is the number of speakers, each having N_s i-vectors. Switchboard II, SRE-2004, SRE-2005, and SRE-2006 data were used to derive the estimation of $\mathbf{T}$. WCCN was estimated only on SRE-2004, SRE-2005, and SRE-2006 data. In addition, we apply the simple technique of normalizing i-vector to the unit length by capturing their directions, $\bar{\mathbf{w}} = \hat{\mathbf{w}}/\|\hat{\mathbf{w}}\|$.

As we discussed earlier, the speech data were divided into several subsets using unsupervised data selection. Table 1 summarized the dataset (or parameters) used for ensemble classifiers based on different evaluation systems.

https://doi.org/10.1017/ATSIP.2016.10 Published online by Cambridge University Press

Table 1. Data (or parameters) used for ensemble classifiers based on different evaluation systems.

Data\system	MAP	Eigenchannel	i-vector
UBM	X	X	X
Enrollment	X	X	X
Testing data	X	X	X
ZT-norm	X	X	
U		X	
T			X
WCCN			X

B) Performance evaluation

Two types of errors, false acceptance and false rejection, can occur in speaker verification. Equal error rate (*EER*) reports the system performance when the false acceptance $P_{FalseAlarm|NonTarget}$ and false rejection rates $P_{Miss|Target}$ are equal. The minimum Detection Cost Function (*DCF*) is a weighed sum of miss detection and false alarm rates as defined in NIST SRE-2010 [33], and shown as follows:

$$\begin{aligned} DCF &= C_{Miss} \times P_{Miss|Target} \times P_{Target} \\ &\quad + C_{FalseAlarm} \times P_{FalseAlarm|NonTarget} \\ &\quad \times (1 - P_{Target}), \end{aligned} \tag{11}$$

where $C_{\mathbf{Miss}} = 10$, $C_{\mathbf{FalseAlarm}} = 0.001$, and $P_{\mathbf{Target}} = 1$ were defined in SRE-2010. The speaker verification results were reported in terms of $1000 \times DCF$ for SRE-2010 in this study. The following results were given on the *EER* and the minimum *DCF* point.

VI. RESULTS AND ANALYSIS

We evaluated the robustness of the ensemble classifiers using unsupervised data selection from several viewpoints. LTF4 was used for all the experiments, which is with four long-term windows.

A) Unsupervised data selection

To determine the effect of unsupervised data selection and the ensemble classifiers, we first compared K-means (K) and hierarchical (H) clustering algorithms based on Mahalanobis distance metric, and weighting schemes using equal weighting (EW) and FCs. The summarized results were shown in Table 2. Four subsets ($k = 4$) were used for the ensemble classifiers with MAP and ZT-norm. The mixture number of UBMs was 256. The baseline system was trained and tested with all data, which means it is the conventional single classifier. The results showed that the K-means clustering algorithm outperformed the hierarchical and baseline systems. The combination of ensemble classifiers with a weighting scheme of FCs was better than EW.

Base on K-means clustering algorithm and a weighting scheme of FCs, we conducted ensemble systems using different similarity metrics to compare with the baseline system. We explored five similarity measures, log-likelihood measure, delta-BIC measure, cosine measure, Euclidean, and Mahalanobis distance measure, to construct partitioning clusters. The summarized results were shown in Table 3. We found that ensemble-based speaker recognition showed improvements with suitable data selection scheme, such as cosine and Mahalanobis distance metrics. Since we conventionally focus on minimizing *DCF* score, the best performance was shown on the Mahalanobis distance metric. Comparing the baseline system, the ensemble-based speaker recognition contributed to 8.51% relative *EER* reduction from 10.81 to 9.89, and 15.29% relative *DCF* reduction from 0.85 to 0.72.

We evaluated the conversational telephone English speech of the SRE-2008 core task based on the optimized setting obtained from the SRE-2010 data using version 3 of the NIST SRE-2008 answer keys. Four subsets were used for the ensemble classifiers with MAP and ZT-norm. The mixture number of UBMs was 256. We had the same improvements shown in Table 4. Comparing with the baseline system, the ensemble-based speaker recognition using the Mahalanobis distance metric contributed to 11.52% relative *EER* reduction from 7.55 to 6.68, and 16.87% relative

Table 2. Results of ensemble classifiers using different clustering and weighting schemes on MAP and ZT-norm systems on NIST SRE-2010.

	Male		Female		All	
Systems	EER (%)	$_{1000x}$DCF	EER (%)	$_{1000x}$DCF	EER (%)	$_{1000x}$DCF
Baseline	9.97	0.75	12.68	0.91	10.81	0.85
H+EW	10.76	0.93	11.83	0.77	11.46	0.95
H+FC	10.48	0.92	11.43	0.78	10.87	0.94
K+EW	9.35	0.64	10.99	0.76	10.03	0.75
K+FC	9.08	0.63	10.42	0.74	9.89	0.72

Table 3. Results of ensemble classifiers using different distance metrics on MAP and ZT-norm systems on NIST SRE-2010.

	Male		Female		All	
Distance metrics	EER (%)	$_{1000x}$DCF	EER (%)	$_{1000x}$DCF	EER (%)	$_{1000x}$DCF
Baseline	9.97	0.75	12.68	0.91	10.81	0.85
Likelihood	10.76	0.90	11.55	0.81	11.16	0.90
Cosine	8.78	0.68	9.89	0.80	9.40	0.78
Euclidean	10.48	0.90	11.55	0.82	11.02	0.91
Mahalanobis	9.08	0.63	10.42	0.74	9.89	0.72
Delta-BIC	10.48	0.70	12.39	0.86	11.30	0.84

Table 4. Results of ensemble classifiers using different distance metrics on MAP and ZT-norm systems on NIST SRE-2008.

	Male		Female		All	
Distance metrics	EER (%)	$_{100x}$DCF	EER (%)	$_{100x}$DCF	EER (%)	$_{100x}$DCF
Baseline	7.03	2.95	7.79	3.50	7.55	3.32
Cosine	7.17	2.78	7.60	3.48	7.45	3.25
Mahalanobis	6.13	2.42	7.08	2.91	6.68	2.76

https://doi.org/10.1017/ATSIP.2016.10 Published online by Cambridge University Press

DCF reduction from 3.32 to 2.76 on the SRE-2008 data. Motivated by the advanced channel and speaker adaptation techniques, we further extend the proposed ensemble-based speaker recognition to eigenchannel and i-vector approaches.

B) Ensemble-based eigenchannel systems

Experiment results of the ensemble-based eigenchannel system with ZT-norm are shown in Fig. 4. We conducted experiments on four different numbers of UBM mixtures including 128, 256, 512, and 1024 with four subsets. Four subsets were used for ensemble classifiers with the Mahalanobis distance metric. In Fig. 4, the blue and dashed line showed the eigenchannel approach. The red and solid line showed the proposed ensemble classifiers with the eigenchannel adaptation.

Compared with results shown in Table 2, large gains were obtained using the eigenchannel technique. Eigenchannel with ZT-norm showed the effect of good channel compensation and score normalization. Our proposed ensemble-based approach can be further used for improving the overall performance. Basically, we can found that the *DCF* score decreased when the number of UBM mixture increased. In Fig. 4, UBM with 256 mixtures achieved the lowest *DCF* scores.

We achieved 19.64% relative *DCF* reduction from 0.56 to 0.45 (or 16.67% relative *DCF* reduction from 0.54 to 0.45) using the ensemble-based eigenchannel system with the UBMs of 256 mixtures. To further explore the relations between the number of mixture in UBM and the number of data subsets in ensemble, we conducted experiments with five different numbers of data subsets (2, 4, 8, 16, and 32) on four different numbers of UBM mixtures (128, 256, 512, and 1024) shown in Fig. 5. Due to data sparsity, a smaller number of subsets in the ensemble should be applied if a larger size of UBM mixtures is adopted. As a result, we can find that UBM of 128 mixtures with eight subsets, UBM of 256 and 512 mixtures with four subsets, and UBM of 1024 mixtures with two subsets achieved the lowest *DCF* scores. The best performance was located on UBM of 256 mixtures with four subsets. Based on these best setting, we further applied ensemble classifiers on the i-vector-based speaker recognition system in the following experiments.

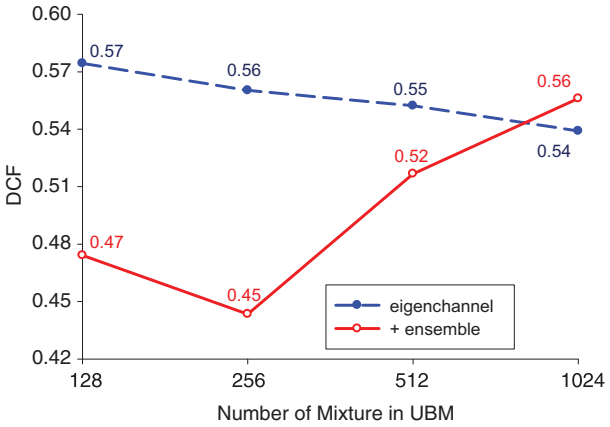

Fig. 4. DCF curves of eigenchannel with ZT-norm systems with different numbers of UBM mixtures on NIST SRE-2010.

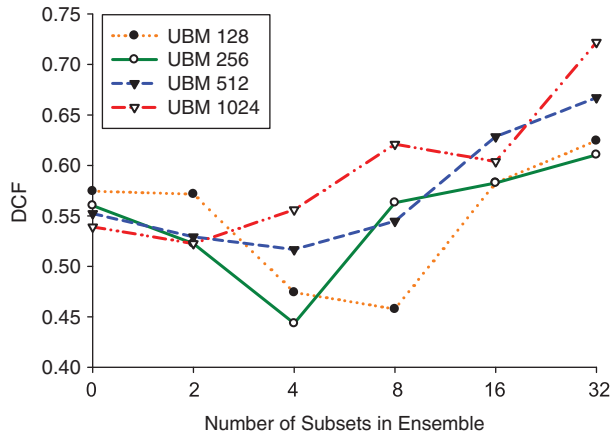

Fig. 5. DCF curves of eigenchannel with ZT-norm systems with different numbers of UBM mixtures and data subsets on NIST SRE-2010.

C) Ensemble-based i-vector systems

The evaluation results of SRE-2010 were shown in Table 5 based on the i-vector system. The proposed ensemble-based systems using unsupervised data selection outperformed the conventional i-vector approach. Comparing the baseline system, the ensemble-based i-vector system contributed to 9.36% relative *EER* reduction from 4.38 to 3.97, and 5.26% relative *DCF* reduction from 0.57 to 0.54% on the SRE-2010 data. Experimental results of SRE-2008 data were shown in Table 6. The experiment confirmed that ensemble classifiers consistently improved the speaker recognition performance. Since original speech data were segmented into several data subsets according to acoustic characteristics on training and testing, we were able to train and test a more robust speaker recognition system.

Fusion of LTFs showed further improvement. We apply the same kind of MFCC features with the different size of

Table 5. Results of I-Vector system with and without ensemble classifiers on NIST SRE-2010.

	Male		Female		All	
Systems	EER (%)	$_{1000x}$DCF	EER (%)	$_{1000x}$DCF	EER (%)	$_{1000x}$DCF
i-vector	3.97	0.56	4.79	0.49	4.38	0.57
+Ensemble	3.68	0.53	4.23	0.50	3.97	0.54

Table 6. Results of i-vector system with and without ensemble classifiers on NIST SRE-2008.

	Male		Female		All	
Systems	EER (%)	$_{100x}$DCF	EER (%)	$_{100x}$DCF	EER (%)	$_{100x}$DCF
	EER (%)	$_{100x}$DCF	EER (%)	$_{100x}$DCF	EER (%)	$_{100x}$DCF
i-vector	3.59	1.75	4.23	1.65	3.78	1.72
+Ensemble	3.44	1.74	3.54	1.63	3.54	1.69

https://doi.org/10.1017/ATSIP.2016.10 Published online by Cambridge University Press

Table 7. fusion of ensemble based I-Vector system with LTFs on NIST SRE-2010 and SRE-2008.

Systems	SRE-2010		SRE-2008	
	EER (%)	$_{1000x}$DCF	EER (%)	$_{100x}$DCF
Baseline	4.38	0.57	3.78	1.72
LTF4	3.97	0.54	3.54	1.69
LTF6	3.90	0.62	3.85	1.76
LTF8	4.24	0.65	3.72	1.85
Fusion	3.67	0.56	3.50	1.64

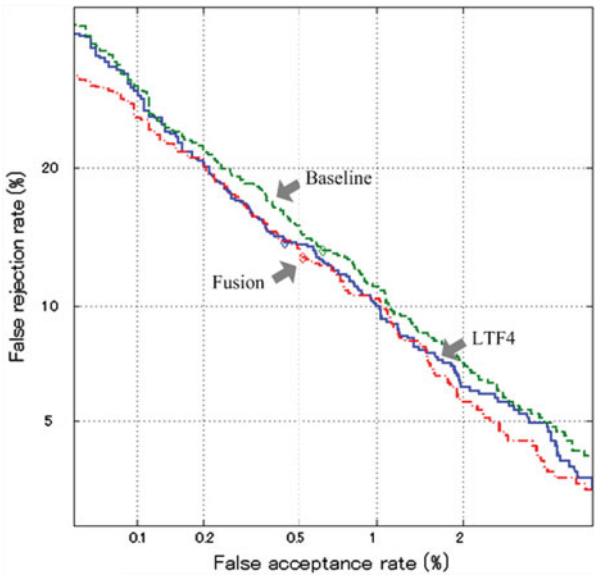

Fig. 6. DET curves showing improvements of conventional i-vector system, ensemble-based system, fusion of LTF system on SRE-2010.

the long-term windows, $L = 4, 6, 8$ frames, namely LTF4, LTF6, and LTF8 [16]. Table 7 showed the fusion results of SRE-2010 and SRE-2008 considering features of LTF4, LTF6, and LTF8. Fusion weights were selected as 0.5, 0.3, and 0.2, respectively. The results showed that the fusion was complementary. The evaluations of SRE-2010 were plotted with the DET curves in Fig. 6. Regarding the i-vector systems, the scoring method used cosine similarity. We apply LTF on i-vector estimation and i-vector is used for unsupervised clustering. In addition, we used ensemble-based system for fusion of LTFs.

VII. CONCLUSION

We studied the ensemble method using unsupervised data selection for effective speaker recognition. Unlike previous constrain approaches, we had no auxiliary information requirement. The speech data were divided into several subsets using K-means clustering algorithm with the Mahalanobis distance metric and the FC weighting scheme. There are many clustering algorithms. In this study, we compared K-means and HAC to discover the effect of clustering algorithms and unsupervised data selection. With the divide-and-conquer strategy, ensemble classifiers were used to avoid the local optimization training on the single classifier. We studied feature extraction techniques using long-term and temporal information for effective speaker recognition, and trained and evaluated the ensemble classifiers based on the selected data subsets. Using the LTF and the ensemble method decreases the amount of data for training, because the LTF provides the more compact feature and the ensemble method divides data into subsets. Three speaker recognition experiments based on MAP, eigenchannel, and i-vector on the NIST SRE-2010 and SRE-2008 datasets were conducted. Based on the experiment results, we confirm that the ensemble classifiers with unsupervised data selection consistently improve the speaker recognition performance on different evaluation tasks and systems.

REFERENCES

[1] Davis, S.; Mermelstein, P.: Comparison of parametric representations of monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Audio Speech Lang. Process., 28 (1980), 357–366.

[2] Hermansky, H.: Perceptual linear prediction (PLP) analysis of speech. J. Acoust. Soc. Am., 87 (4) (1990), 1738–1752.

[3] Bimbot, F. et al.: A tutorial on text-independent speaker verification. EURASIP J. Appl. Signal Process., 4 (2004), 430–451.

[4] Kenny, P.; Boulianne, G.; Ouellet, P.; Dumouchel, P.: Joint factor analysis versus eigenchannels in speaker recognition. IEEE Trans. Audio Speech Lang. Process., 15 (4) (2007), 1435–1447.

[5] Dehak, N.; Kenny, P.; Dehak, R.; Dumouchel, P.; Ouellet, P.: Front-end factor analysis for speaker verification. IEEE Trans. Audio Speech Lang. Process., 19 (4) (2011), 788–798.

[6] Huang, C.-L.; Wu, C.-H.: Generation of phonetic units for mixed-language speech recognition based on acoustic and contextual analysis. IEEE Trans. Comput., 56 (9) (2007), 1225–1233.

[7] Park, A.; Hazen, T.J.: ASR dependent techniques for speaker identification, in *Proc. Seventh Int. Conf. on Spoken Language Processing*, Denver, Colorado, USA, 2002, 1337–1340.

[8] Zhang, Y.; Tang, Z.-M.; Li, Y.-P.; Qian, B.: Ensemble learning and optimizing KNN method for speaker recognition, *in Proc. Fourth Int. Conf. on Fuzzy Systems and Knowledge Discovery (FSKD)*, Haikou, Hainan, China, 2007, 285–289.

[9] Sreenivasa Rao, K.; Sarkar, S.: Robust Speaker Recognition in Noisy Environments, Springer International Publishing, 2014. doi:10.1007/978-3-319-07130-5.

[10] Sturim, D.E.; Reynolds, D.A.; Dunn, R.B.; Quatieri, T.F.: Speaker verification using text-constrained Gaussian mixture models, in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Orlando, Florida, USA, 2002, 677–680.

[11] Baker, B.; Vogt, R.; Sridharan, S.: Gaussian mixture modelling of broad phonetic and syllabic events for text-independent speaker verification, in *Proc. Ninth European Conf. on Speech Communication and Technology*, Lisbon, Portugal, 2005, 2429–2432.

[12] Bocklet, T.; Shriberg, E.: Speaker recognition using syllable-based constraints for cepstral frame selection, in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, 2009, 4525–4528.

[13] Sanchez, M.H.; Ferrer, L.; Shriberg, E.; Stolcke, A.: Constrained cepstral speaker recognition using matched UBM and JFA training, in *Proc. 13th Annu. Conf. Int. Speech Communication Association (Interspeech)*, Florence, Italy, 2011, 141–144.

https://doi.org/10.1017/ATSIP.2016.10 Published online by Cambridge University Press

[14] Shum, S.; Reynolds, D.; Garcia-Romero, D.; McCree, A.: Unsupervised clustering approaches for domain adaptation in speaker recognition systems, in *Proc. Odyssey*, Joensuu, Finland, 2014.

[15] Garcia-Romero, D.; Zhou, X.; Espy-Wilson, C.Y.: Multicondition training of Gaussian PLDA models in i-vector space for noise and reverberation robust speaker recognition, in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Kyoto, Japan, 2012, 4257–4260.

[16] Jacobs, R.A.; Jordan, M.I.; Nowlan, S.J.; Hinton, G.E.: Adaptive mixtures of local experts. *Neural Comput.*, 3 (1) (1991), 79–87.

[17] Gales, M.J.F.: Semi-tied covariance matrices for hidden Markov models. IEEE Trans. Speech Audio Process., 7 (1999), 272–281.

[18] Reynolds, D. et al.: Beyond cepstra: exploiting high-level information in speaker recognition, in *Workshop on Multimodal User Authentication*, Santa Barbara, CA, 2003.

[19] Wang, L.; Ohtsuka, S.; Nakagawa, S.: High improvement of speaker identification and verification by combining MFCC and phase information, in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, 2009, 4529–4532.

[20] Pan, J.; Liu, C.; Wang, Z.; Hu, Y.: Ensemble learning and optimizing KNN method for speaker recognition, in *Proc. Eighth Int. Symp. on Chinese Spoken Language Processing (ISCSLP)*, Hong Kong, 2012, 301–305.

[21] Huang, C.-L.; Su, H.; Ma, B.; Li, H.: Speaker characterization using long-term and temporal information, in *Proc. 12th Annu. Conf. Int. Speech Communication Association (Interspeech)*, Makuhari, Chiba, Japan, 2010, 370–373.

[22] Anguera Miro, X.; Bozonnet, S.; Evans, N.; Fredouille, C.; Friedland, G.; Vinyals, O.: Speaker diarization: a review of recent research. IEEE Trans. Audio Speech Lang. Process., 20 (2012), 356–370.

[23] Senoussaoui, M.; Montreìal, Q.C.; Kenny, P.; Stafylakis, T.; Dumouchel, P.: A study of the cosine distance-based mean shift for telephone speech diarization. IEEE Trans. Audio, Speech Lang. Process., 22 (2014), 217–227.

[24] Huang, C.-L.; Wu, C.-H.: Spoken document retrieval using multilevel knowledge and semantic verification. IEEE Trans. Audio Speech Lang. Process., 15 (8) (2007), 2551–2560.

[25] Huang, C.-L.; Ma, B.; Li, H.; Wu, C.-H.: Speech indexing using semantic context inference, in *Proc. 13th Annu. Conf. Int. Speech Communication Association (Interspeech)*, Florence, Italy, 2011, 717–720.

[26] Schwarz, G.: Estimating the dimension of a model. *Ann. Stat.*, 6 (2) (1978), 461–464.

[27] Wu, C.-H.; Hsieh, C.-H.: Multiple change-point audio segmentation and classification using an MDL-based gaussian model. IEEE Trans. Audio Speech, Lang. Process., 14 (2) (2006), 647–657.

[28] Tang, H.; Chu, S.M.; Hasegawa-Johnson, M.; Huang, T.S.: Partially supervised speaker clustering. IEEE Trans. Pattern Anal. Mach. Intell., 34 (5) (2012), 959–971.

[29] Pelecanos, J.; Sridharan, S.: Feature warping for robust speaker verification, in *Proc. 2001: A Speaker Odyssey*, Crete, Greece, 2001, 213–218.

[30] Torre, A.; Peinado, A.M.; Segura, J.C.; Perez-Cordoba, J.L., Bentez, M.C.; Rubio, A.J.: Histogram equalization of speech representation for robust speech recognition. IEEE Trans. Speech Audio Process., 13 (3) (2005), 355–366.

[31] Huang, C.-L.; Tsao, Y.; Hori, C.; Kashioka, H.: Feature normalization and selection for robust speaker state recognition, in *Proc. Oriental COCOSDA*, Hsinchu, Taiwan, 2011, 102–105.

[32] Polikar, R.: Ensemble based systems in decision making. IEEE Circuits Syst. Mag., 6 (3) (2006), 21–45.

[33] Rokach, L.: Ensemble-based classifiers. Artif. Intell. Rev., 33 (2010), 1–39.

[34] The NIST year 2010 speaker recognition evaluation plan, 2010. [Online] Available: http://www.nist.gov/

[35] Reynolds, D.A.; Quatieri, T.F.; Dunn, R.B.: Speaker verification using adapted Gaussian mixture models. Digit. Signal Process., 10 (2000), 19–41.

[36] Auckenthaler, R.; Carey, M.; Lloyd-Thomas, H.: Score normalization for text-independent speaker verification systems. Digit. Signal Process., 10 (2000), 42–54.

[37] Li, K.P.; Porter, J.E.: Normalizations and selection of speech segments for speaker recognition scoring, in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, New York, New York, USA, 1988, 595–598.

Chien-Lin Huang received the Ph.D. degree in Computer Science and Information Engineering from National Cheng Kung University, Taiwan, in 2008. He is a speech scientist at Voicebox Technologies currently. Before, he was a scientist in Japan NICT and Singapore I2R, respectively. Chien-Lin's research focuses on speech recognition, speaker recognition, and speech retrieval. He is an active member of speech and language processing communities. He has co-authored over 40 technical papers and holds 2 U.S. patents.

Jia-Ching Wang received the Ph.D. degree in Electrical Engineering from National Cheng Kung University, Tainan, Taiwan. He currently works at Department of Computer Science and Information Engineering, National Central University, Jhongli, Taiwan, as an associate professor. He was an honorary fellow at Department of Electrical and Computer Engineering, University of Wisconsin-Madison, WI, USA, during 2008 and 2009. His research interests include multimedia signal processing and associated VLSI architecture design. He is an honorary member of Phi Tau Phi Scholastic Honor Society and a senior member of IEEE.

Bin Ma received the B.Sc. degree in Computer Science from Shandong University, China, in 1990, the M.Sc. degree in Pattern Recognition & Artificial Intelligence from the Institute of Automation, Chinese Academy of Sciences (IACAS), China, in 1993, and the Ph.D. degree in Computer Engineering from The University of Hong Kong, in 2000. He was a Research Assistant from 1993 to 1996 at the National Laboratory of Pattern Recognition in IACAS. In 2000, he joined Lernout & Hauspie Asia Pacific as a Researcher working on speech recognition. From 2001 to 2004, he worked for InfoTalk Corp., Ltd, as a Senior Researcher and a Senior Technical Manager for speech recognition. He joined the Institute for Infocomm Research, Singapore in 2004 and is now working as a Senior Scientist and the Lab Head of Automatic Speech Recognition. He has served as a Subject Editor for Speech Communication in 2009–2012, the Technical Program Co-Chair for INTERSPEECH 2014, and is now serving as an Associate Editor for IEEE/ACM Transactions on Audio, Speech, and Language Processing. His current research interests include robust speech recognition, speaker & language recognition, spoken document retrieval, natural language processing and machine learning.

https://doi.org/10.1017/ATSIP.2016.10 Published online by Cambridge University Press