IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received February 1, 2019, accepted February 23, 2019, date of publication February 26, 2019, date of current version March 13, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2901812

Noise Robust Speaker Recognition Based on Adaptive Frame Weighting in GMM for i-Vector Extraction

XINGYU ZHANG[1], XIA ZOU[1], MENG SUN[1], THOMAS FANG ZHENG[2], (Senior Member, IEEE), CHONG JIA[1], AND YIMIN WANG[3]

[1]Lab of Intelligent Information Processing, Army Engineering University, Nanjing 210007, China
[2]Research Institute of Information Technology, Tsinghua University, Beijing 100084, China
[3]College of Communication Engineering, Army Engineering University, Nanjing 210007, China

Corresponding author: Meng Sun (sunmengccjs@ gmail.com)

This work was supported by the Natural Science Foundation of Jiangsu Province for Excellent Young Scholars, under Grant BK20180080.

ABSTRACT Even though speaker recognition has gained significant progress in recent years, its performance is known to be deteriorated severely with the existence of strong background noises. Inspired by a recently proposed clean-frame selection approach, this work investigates a relatively elegant weighting method when computing the Baum-Welch statistics of Gaussian mixture models (GMMs) in i-vector extraction. By introducing weighting parameters to the frames of enrollment/testing utterances, the optimization problem is redefined and solved. New updating rules are derived by incorporating weights to the computation of posterior probabilities, mean vectors, and covariance matrices of the GMM. The experiments conducted on the Speakers in the Wild (SITW) database show that the proposed algorithm has significantly improved the performance of i-vector-based speaker recognition systems in noisy environments. Compared with the GMM i-vector baseline, the equal error rate is reduced from 5.75 to 4.72 and the minimum value of cost function ($C_{\text{det}}^{\text{min}}$) is reduced from 0.4825 to 0.4505. Slight but significant superiority is also observed over the method with an additional feature enhancement frontend by using deep neural networks.

INDEX TERMS Gaussian mixture models, frame weighting, Baum-Welch statistics, i-vector, robust speaker recognition.

I. INTRODUCTION

A key property of a good automatic speaker recognition (ASR) system is being able to model the uncertainty and variation of the utterances of the same speaker. Towards this goal, a lot of statistical approaches, such as Vector Quantization (VQ), Gaussian mixture models (GMMs) [1] and i-vectors [2] have been proposed and investigated extensively over the past decades. Those classical approaches have obtained satisfying performance on clean conditions and have been applied in miscellaneous scenarios where a speaker's identity needs to be recognized by voice. Recently, with the rise of deep learning, deep Neural Networks (DNNs) and deep Convolution Neural Networks (CNNs) have also been explored to speaker recognition by extracting deep features [3]–[5] Among those deep models, deep speaker embedding has shown promising performance using limited enrollment and testing data [6]. Compared with deep learning methods, GMM holds relatively low computing complexity and does not require additional backend software (e.g. Caffe, TensorFlow). It is thus easy to be implemented in miscellaneous computing platforms. Therefore, it is still meaningful to improve the performance of the GMM+i-vector recipe in real-world scenarios with the existence of many kinds and SNR levels of background noises.

In realistic scenarios, due to the mismatch between training data (usually clean) and testing data (usually contaminated by unknown noises with different intensities), the recognition performance of the system will degrade drastically. To deal with this problem, researchers have proposed many methods to enhance the quality of features at different levels, such as speech signal enhancement by statistical processing [7]–[9] or by deep learning [10]–[12], DNN based cepstral feature de-noising [13], i-vector de-noising [14], multi-task adversarial network (MAN) for extracting noise-invariant bottleneck (BN) features [15] etc. Backend classifiers have also

The associate editor coordinating the review of this manuscript and approving it for publication was Shiqiang Wang.

been investigated by parallel model combination [16], robust variants of the Probabilistic Linear Discriminative Analysis (PLDA) model [17], [18], multi-style training [19], [20], etc. These techniques usually work as plug-in tools to augment the performance of GMM+i-vector. As a substitution of GMM+i-vector, our work could also be further improved by the techniques above.

Besides improving the quality of contaminated features, feature selection by discarding noisy frames is another idea to improve the performance of speaker recognition. The work in [21] showed that, voice segments drowned in noises, could play a negative role, which could be improved by conducting frame selection. SoftSAD was proposed in [22] to perform soft selection of speech frames by weighting the frames in the calculation of Baum-Welch statistics. Though the motivation of this work was to choose "speech-like" frames from clean speech, it provided meaningful grounding for our work by showing the benefit of using frame-weighted Baum-Welch statistics. The work in [23] studied nonlinear frame-likelihood weighting method which proved the practicability of the frame dependent non-linear weighting method. Our work would extend above ideas to a more practical aspect of noise-robustness on speaker recognition.

Recently, an algorithm, Noise Invariant Frame Selection (NIFS), was proposed in [24], where the input testing utterance was artificially-noised to help choose noise invariant frames. Experiments showed the effectiveness of this algorithm on speaker recognition in noisy environments, except under low signal-to-noise ratios (SNRs). Another inefficiency of the algorithm was that a key threshold needs to be chosen to judge which frames to retain and which ones to remove. In this work, to improve the noise robustness of speaker recognition and to get rid of the difficulty of choosing such a threshold, we introduce frame weighting to the computation of Baum-Welch (BW) statistics of GMM in i-vector extraction. Experiments have demonstrated the effectiveness of the proposed approach, compared with several recently proposed algorithms. It is worth noting that our algorithm can also be used in other speech processing tasks where noise robustness is required during i-vector extraction, e.g. spoofing detection in [25] and speaker identification in [26].

This paper focuses on improving the noise robustness of GMM+i-vector based ASR system. Specifically, an algorithm is proposed to incorporate frame weights to the calculation of Baum-Welch statistics in the procedures of i-vector extraction The main contributions of this paper are listed as follows.

1) An improved Baum-Welch algorithm to train GMM with data weighting, which outperformed a similar approach presented in [27].
2) A straightforward approach to evaluate the noise robustness of speech frames without additional manual efforts on choosing a threshold in [24].

The remaining part of this paper is organized as follows. In Section II, a convention recipe of noise robust ASR by using GMM and i-vector extraction is presented, as well as on which part our algorithm will improve. The detailed mathematical derivation of the algorithm to solve GMM with data weighting is given in Section III. In Section IV, how the weights are calculated is stated, which serves as inputs to the algorithms derived in Section III The experimental results of the proposed algorithm and its comparison with other state-of-the-art ones are discussed in Section V. The conclusion is given in Section VI.

II. i-VECTOR BASED NOISE ROBUST ASR AND THE ROLE OF OUR WORK

In this section, we introduce the noise robust ASR system by using the recipe of GMM+i-vector and explain on which part our work will improve. First of all, a conventional recipe of noise robust ASR by using GMM and i-vector extraction is presented in Fig.1, which has several components: training, enrollment and testing [28], as is illustrated in Fig.1.

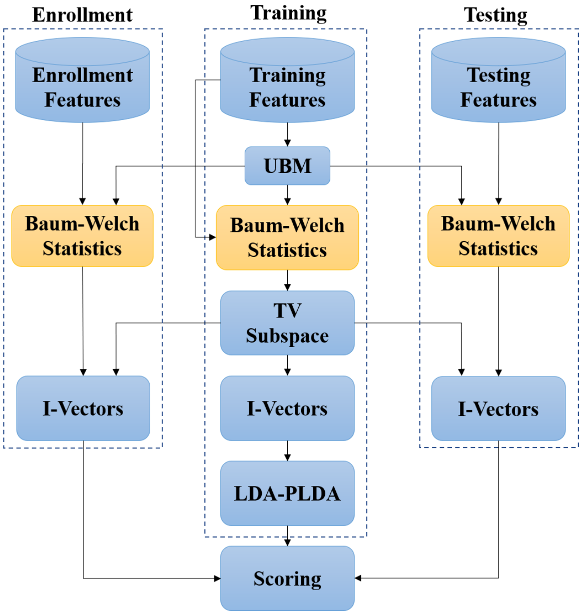


FIGURE 1. A conventional recipe of noise robust automatic speaker recognition by using GMM and i-vector. The dashed boxes from left to right refer to enrollment, training and testing, respectively. The proposed algorithm of this paper is going to improve the highlighted boxes of Baum-Welch statistics.

In the training stage (see the dashed middle box of Fig.1), a universal background model (UBM) is firstly trained on the features extracted from training data (Training Features in Fig. 1) which is usually collected from a large number of speakers. Baum-Welch (BW) statistics of the training stage are subsequently obtained by adapting the UBM to the training features of each training speaker A total variability (TV) subspace model is then learned from the BW statistics of all the training speakers and is deployed to yield a low and fixed dimensional latent factor, i.e. i-vectors. With a further dimension reduction by using Linear Discriminative Analysis (LDA), a Probabilistic LDA (PLDA) is learned as the classifier to facilitate speaker recognition. The routine developed in [28] is utilized for the LDA-PLDA step.

In the enrollment/testing stages (see the dashed left/right box of Fig.1), their corresponding BW statistics are generated by adapting the UBM to the enrollment/testing features. The learned TV subspace model is then applied on the corresponding BW statistics to extract i-vectors of enrollment/testing speakers.

Finally, the i-vectors from the enrollment speakers and the testing speakers are sent to the PLDA classifier which outputs the log-likelihood ratios as scores for evaluation.

The algorithm proposed in this paper is going to improve the three highlighted parts, Baum-Welch statistics, in the recipe depicted by Fig. 1. Data weighting will be introduced to the modeling of GMM when computing Baum-Welch statistics to reflect the degree to which the data is contaminated by noises. The rigid mathematical derivation of the algorithm will be presented in the next section.

III. WEIGHTED GMM IN i-VECTOR EXTRACTION

In this section, we consider different frames with different weights when computing BW statistics of GMM in i-vector extraction. The weights are denoted by $\{\alpha_1, \cdots \alpha_i, \cdots \alpha_N\}$ for frames $\{x_1, \cdots x_i, \cdots x_N\}$ where $\{\alpha_i \geq 0, i = 1, \cdots N\}$. The definitions of the weights will be discussed in Section IV.

A. WEIGHTED GMM AND PARAMETER ESTIMATION

In speaker recognition, a GMM can be utilized to model the probability density of spectral features extracted from utterances of a particular speaker. For a D-dimensional feature vector, x_i, the probability density is given by,

$$Pr\,(x_i;\theta) = \sum_{k=1}^{K} w_k \mathcal{N}\,(x_i; m_k, \Sigma_k), \tag{1}$$

where $\mathcal{N}\,(x_i; m_k, \Sigma_k)$ is a D dimensional Gaussian distribution, m_k the mean vector, Σ_k the diagonal covariance matrix, w_k the weight of the k-th Gaussian with the constraint $\sum_{k=1}^{K} w_k = 1$, K the total number of Gaussians and $\theta = \{w_k, m_k, \Sigma_k\}_{k=1}^{K}$ the set of the GMM parameters.

Given N feature vectors extracted from an utterance, the Maximum-Likelihood Estimation (MLE) of θ is going to maximize the following data likelihood,

$$L\,(x_i;\theta) = \prod_{i=1}^{N} Pr\,(x_i;\theta). \tag{2}$$

Thanks to the property of the exponential family, log-likelihood, $J\,(\theta)$ is usually used as the optimization goal,

$$\max_{\theta} J\,(\theta) = \max_{\theta} \sum_{i=1}^{N} \log Pr(x_i;\theta). \tag{3}$$

By introducing weighting parameter α_i for each feature vector x_i, the log-likelihood objective function becomes,

$$\max_{\theta} J\,(\theta) = \max_{\theta} \sum_{i=1}^{N} \alpha_i \log \sum_{k=1}^{K} w_k \mathcal{N}\,(x_i; m_k, \Sigma_k). \tag{4}$$

It is obviously seen from (4) that the objective function is consistent with the conventional model in (3) when no frame weighting is conducted where $\alpha_i = 1$.

In order to optimize (4), inspired by the derivation of the conventional Expectation Maximization (EM) algorithm [29] an auxiliary function $Q\left(\theta;\hat{\theta}\right)$,

$$Q\left(\theta;\hat{\theta}\right) = \sum_{i=1}^{N} \alpha_i \sum_{k=1}^{K} \hat{\beta}_{ik}\,(\log w_k + \log \mathcal{N}\,(x_i; m_k, \Sigma_k)) + C \tag{5}$$

is firstly constructed by introducing an intermediate variable,

$$\hat{\beta}_{ik} = \frac{\hat{w}_k N\left(x_i; \hat{m}_k, \hat{\Sigma}_k\right)}{\sum_{j=1}^{K} \hat{w}_j N\left(x_i; \hat{m}_j, \hat{\Sigma}_j\right)}, \tag{6}$$

where C is a nonnegative constant term,

$$C = \sum_{i=1}^{N} \alpha_i \sum_{k=1}^{K} \hat{\beta}_{ik} \log \frac{1}{\hat{\beta}_{ik}} \geq 0, \tag{7}$$

and $\hat{\theta} = \{\hat{w}_k, \hat{m}_k, \hat{\Sigma}_k\}_{k=1}^{K}$ is the parameter estimation of the previous iteration in EM.

It is straightforward to show that $Q\left(\hat{\theta};\hat{\theta}\right) = J\left(\hat{\theta}\right)$ and $J\,(\theta) \geq Q\left(\theta;\hat{\theta}\right)$, given $\sum_{k=1}^{K} \hat{\beta}_{ik} = 1$ and $\hat{\beta}_{ik} \geq 0$. Therefore, for every iteration, one only needs to maximize $Q\left(\theta;\hat{\theta}\right)$ (w.r.t. θ) to increase the value of $J\,(\theta)$ until convergence.

Given the fact that $Q\left(\theta;\hat{\theta}\right)$ is a concave function of θ the stationary point would be the optimization solution,

$$\frac{\partial Q\left(\theta;\hat{\theta}\right)}{\partial m_k} = 0, \quad \frac{\partial Q\left(\theta;\hat{\theta}\right)}{\partial \Sigma_k} = 0. \tag{8}$$

Hence, m_k and Σ_k are calculated by,

$$m_k = \frac{\sum_{i=1}^{N} \alpha_i \hat{\beta}_{ik} x_i}{\sum_{i=1}^{N} \alpha_i \hat{\beta}_{ik}}, \tag{9}$$

and

$$\Sigma_k = \text{diag}\left(\frac{\sum_{i=1}^{N} \alpha_i \hat{\beta}_{ik} (x_i - m_k)(x_i - m_k)^T}{\sum_{i=1}^{N} \alpha_i \hat{\beta}_{ik}}\right), \tag{10}$$

where *diag* refers to the diagonalization operator by which only the diagonal entries in a matrix are retained. Lagrange multiplier is subsequently applied to optimize the function with respect to w_k

$$\max_{w_k} Q_{new} = \max_{w_k} \left(Q + \lambda\left(\sum_{k=1}^{K} w_k - 1\right)\right). \tag{11}$$

By solving $\partial Q_{new} / \partial w_k = 0$, the updated w_k is obtained by,

$$w_k = \frac{\sum_{i=1}^{N} \alpha_i \hat{\beta}_{ik}}{\sum_{i=1}^{N} \alpha_i}. \tag{12}$$

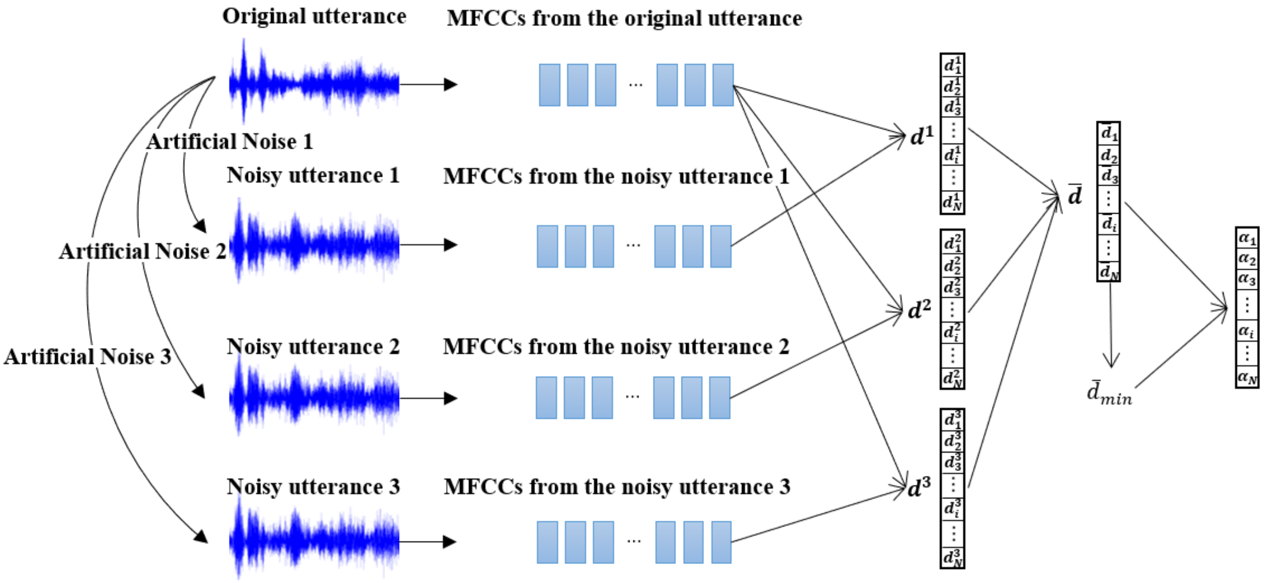

Original utterance
MFCCs from the original utterance
Artificial Noise 1
Noisy utterance 1
MFCCs from the noisy utterance 1
Artificial Noise 2
Artificial Noise 3
Noisy utterance 2
MFCCs from the noisy utterance 2
Noisy utterance 3
MFCCs from the noisy utterance 3
d^1
d^2
d^3
d_1^1 d_2^1 d_3^1 d_i^1 d_N^1
d_1^2 d_2^2 d_3^2 d_i^2 d_N^2
d_1^3 d_2^3 d_3^3 d_i^3 d_N^3
$\bar{d}$
$\bar{d}_1$ $\bar{d}_2$ $\bar{d}_3$ $\bar{d}_i$ $\bar{d}_N$
$\bar{d}_{min}$
α_1 α_2 α_3 α_i α_N

FIGURE 2. The algorithm of calculating different weights for different frames. $\bar{d}_{\mathbf{i}}$ is the averaged Euclidean distance over the three-noise cases of the *i*-th frame. $\bar{d}_{min}$ is the minimum among $\{\bar{d}_1\bar{d}_2,\ldots,\bar{d}_N\}$. The weights α_i can be obtained by using Eqn. (17).

In the next EM iteration, $\hat{\beta}_{ik}$ is firstly computed by (6) with the updated w_k, m_k and Σ_k, and (9) (10) (12) are subsequently conducted to update $\theta = w_k, m_k, \Sigma_k\}_{k=1}^{K}$.

B. BMODIFICATIONS OF THE I-VECTOR EXTRACTION PROCEDURES

With a pre-trained speaker and channel independent super-vector $\mu_{KD\times 1}$ from the means of the universal background model (UBM), the i-vector $\omega_{R\times 1}$ is extracted by solving the following equation,

$$M = \mu + T\omega, \tag{13}$$

where $M_{KD\times 1}$ is the super-vector computed by concatenating the means of the GMM which has been adapted to the features from a specific speaker as presented in (9), $T_{KD\times R}$ a low-rank matrix to model the speaker and channel variability and $\omega_{R\times 1}$ the i-vector which is a random vector following a standard normal distribution Matrix T models the total variability subspace and has been learned from the training data by using the EM algorithm as presented in [2].

For the speech frames of a specific speaker from the training/enrollment/testing data, by using the weighed GMM presented in subsection A, a weighted version of M will be obtained, which boils down to computing the following modified zeroth and first order BW statistics,

$$N_k = \sum_{i=1}^{N} \beta_{ik}\alpha_i \quad \text{and } F_k = \sum_{i=1}^{N} \beta_{ik}\alpha_i x_i. \tag{14}$$

Also, the centralized first order statistics are required later

$$\tilde{F}_k = \sum_{i=1}^{N} \beta_{ik}\alpha_i\,(x_i - \mu_k), \tag{15}$$

where μ_k is the k-th sub-vector of μ. It is straightforward to see that, $\tilde{F}_k/N_k$ is the k-th sub-vector of the centralized statistics $M - \mu$ With uniform weighting, i.e. $\alpha_i = 1, \forall i$, (15) is consistent with the corresponding step of the conventional i-vector extraction method, given the fact that $\sum_{k=1}^{K} \beta_{ik} = 1$.

Finally, the i-vector for this speaker is obtained by,

$$\omega = (I + T^{'}\Sigma^{-1}NT)^{-1}T^{'}\Sigma^{-1}\tilde{F}, \tag{16}$$

where $I_{R\times R}$ is an identity matrix, $N_{KD\times KD}$ a diagonal matrix with diagonal blocks $\{N_k I_{D\times D}, k = 1, \cdots, K\}$, $\tilde{F}_{KD\times 1}$ a super-vector obtained by concatenating $\tilde{F}_k$'s, $\Sigma_{KD\times KD}$ a diagonal covariance matrix estimated during factor analysis training (see [30]) which models the residual variability not captured by the total variability matrix T.

IV. WEIGHTS DEFINITION TO IMPROVE NOISE ROBUSTNESS

In Section III, weighted GMM for i-vector extraction has been presented. In this section, we describe how to calculate different weights for different frames as shown in Fig.2 Inspired by [24], in our proposed algorithm, different types of noises are added to the original testing speech to explore the noise-robustness of different frames. In order to make straightforward comparison w.r.t. [24], in the following experiments, three types of noises, i.e. *white*, *babble*, and *pink*, are chosen the same as in [24]. Then the Euclidean distances of MFCCs between the frames of the resulted noisy speech and the frames of the original noisy speech are calculated. The three distances are averaged for each frame, as depicted in Fig.2. The weight for the i-th frame x_i is subsequently defined by,

$$\alpha_i = e^{-(\bar{d}_i - \bar{d}_{min})}, \tag{17}$$

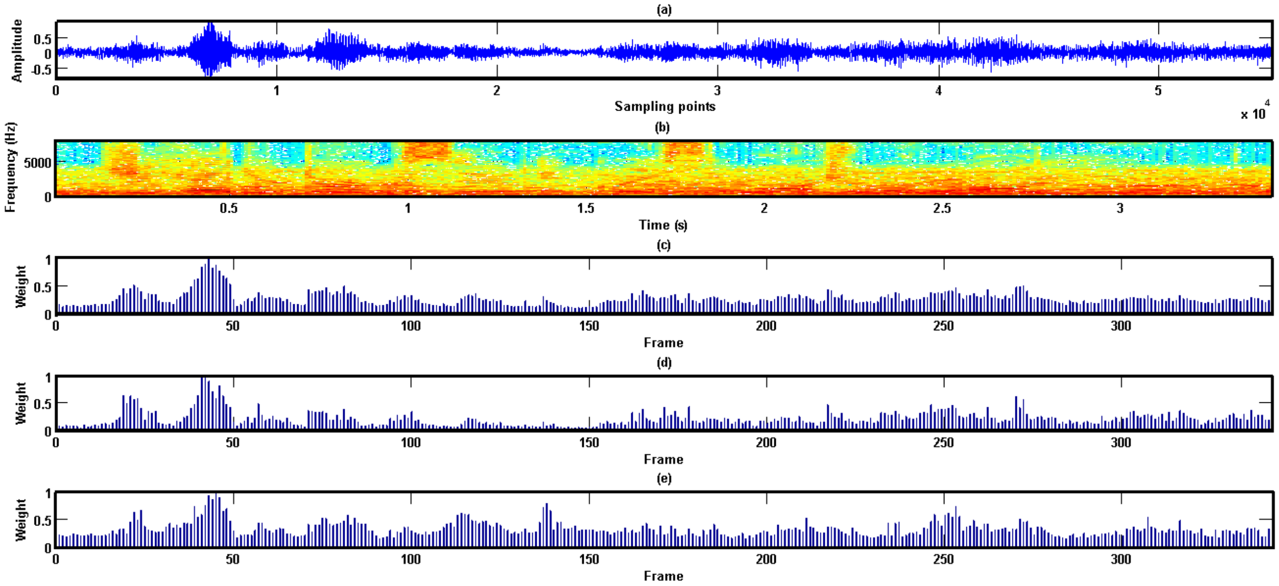


FIGURE 3. **(a) Speech waveform. (b) Spectrogram. (c)(d)(e) Weights calculated by (17) (18) (19), respectively.**

where $\overline{d}_i$ is the averaged Euclidean distance over the three-noise cases of the i-th frame and $\overline{d}_{min}$ the minimum among the average values $\{\overline{d}_1\overline{d}_2, \ldots, \overline{d}_N\}$.

It is worth mentioning that the impacts of different noise combinations, different SNRs and the number of noises on the robustness of the selection of frames are analyzed in [24]. Although the performance varies a bit with these factors, every combination and their average performance perform better than the baselines. Hence, we will not study this topic here.

In Fig3, the waveform of an utterance and its corresponding spectrogram together with the weights extracted by Eqn. (17) are shown. It can be seen from the figure that high weights correspond to formants in the spectrogram and high-energy parts of the signal, which makes sense For speaker recognition tasks, formants and voiced sounds contribute more than other parts to the recognition results. Increasing their weights will help improve the robustness of the system One may argue that other statistics could be taken to replace $\overline{d}_i$. We hereby replace $\overline{d}_i$ with the maximal or minimal value of the distances regarding the three noise types, which is denoted by $\hat{d}_i$ (for maximum) or $\check{d}_i$ (for minimum), respectively. Correspondingly, to facilitate normalization, $\overline{d}_{min}$ in (17) should be replaced by the minimal value of $\hat{d}_i$'s (denoted by $\hat{d}_{min}$) or the minimal value of $\check{d}_i$'s (denoted by $\check{d}_{min}$), respectively. Therefore, the formulae to calculate frame weights are switched to,

$$\alpha_i = e^{-\left(\hat{d}_i - \hat{d}_{min}\right)}, \tag{18}$$

and

$$\alpha_i = e^{-\left(\check{d}_i - \check{d}_{min}\right)}, \tag{19}$$

respectively. Fig.3 shows the weights of the frames of an utterance calculated by (17), (18) and (19), respectively. From Fig.3 it is straightforward to see that the contours of the weights from the three schemes are quite similar, which has also induced their similar performance on EER values in our experiments[1]. Besides this similarity, the amounts of weights decease monotonously in the order of the minimum scheme (19), the mean scheme (17) and the maximum scheme (18), which is as expected. The best performance of the mean scheme (17) in our experiments implies that the scheme is able to reflect the degree to which a frame is contaminated by noises more adequately than the maximum scheme (18) and the minimum scheme (19). Therefore, the weighting scheme is always taken as (17) in the following experimental setting.

V. EXPERIMENTS AND RESULTS

A. AEXPERIMENTAL SETUP

In this work, $D = 39$ dimensional MFCC features with 13 MFCCs, 13 Δ and 13 $\Delta\Delta$ were utilized. Each frame of an utterance was processed by a 25 ms Hamming window with 10 ms shifts. A first-order high pass pre-emphasis filter with $\alpha = 0.97$ was applied. 27 Mel-channels were used in the filter-bank. $K = 2048$ Gaussians were taken in GMM. $R = 400$ dimensional i-vectors were extracted.

The experiments were carried out on the SITW speaker recognition database, TIMIT VoxCeleb1 and VoxCeleb2. SITW contains hand-annotated speech samples from open-source media for the purpose of benchmarking text-independent speaker recognition technology on single and multi-speaker audio acquired across unconstrained or "wild"

[1] For readers' reference, EERs on SITW involving the three weighting schemes: 4.72% with (17), 4.77% with (18) and 4.74% with (19).

conditions, which consists of 2800 recordings of 299 speakers, with an average of eight different sessions per person [31] TIMIT was designed for speech recognition, especially for phoneme recognition, which consists of 6300 recordings of 630 speakers and served as a good choice for training models [32] VoxCeleb1 is a large-scale text-independent speaker identification dataset collected in the wild, which was extracted from videos uploaded to YouTube and consists of 153516 utterances of 1251 speakers. The dataset is gender balanced, with 55% of the speakers male. The speakers span a wide range of different ethnicities, accents, professions and ages. The nationality and gender of each speaker is also provided. Crucially, all are degraded with realworld noise, consisting of background chatter, laughter, overlapping speech, room acoustics, and there is a range in the quality of recording equipment and channel noise [33] VoxCeleb2 contains over 1 million utterances for over 6,000 celebrities, extracted from videos uploaded to YouTube. The dataset is fairly gender balanced, with 61% of the speakers male. The speakers span a wide range of different ethnicities, accents, professions and ages [34] VoxCeleb1 and VoxCeleb2 provide a large amount of data to model the diversity of speaker characteristics.

Our algorithm presented in Section III and IV was evaluated on the *core-core* subset of SITW, which involves single speaker files and focuses on the solution of single speaker recognition problem. It is worth noting that the *core-core* condition in SITW SRC is similar to NIST SRE but more challenging since the utterances in SITW was recorded in real-world environment with reverberation and noises.

Equal Error Rate (EER) and C_{det}^{min} were computed to compare the algorithms. The primary metric is based on the cost function C_{det} with modified parameters in SITW [31],

$$C_{det} = C_{miss} \times P_{miss} \times P_{tar} + C_{fa} \times P_{fa} \times (1 - P_{tar}), \quad (20)$$

where the prior target probability P_{tar} is set to 0.01 and costs for missing detection and false alarm are set to 1 ($C_{miss} = C_{fa} = 1, P_{tar} = 0.01$) The primary metric, C_{det}^{min} is the minimum value of C_{det} for the range of thresholds. The EER is the rate at which both acceptance and rejection errors are equal.

B. RESULTS AND DISCUSSION

In order to prove the efficiency of the proposed algorithm, the following algorithms were chosen for comparison, GMM+i-vector baseline [28] NIFS [24], Denoised i-vector [14], a *fixed weighted-data* EM algorithm (FWD-EM) [27] and DNN feature enhancement (DNN-FE) [13].

In order to evaluate the performance of the proposed algorithm trained on both small-scaled dataset and large-scaled dataset, we designed three groups of experiments on SITW.

For the first group of experiments, only TIMIT was used to train UBM and PLDA, which contains 6300 utterances from 630 speakers. The performance is given in Table 1.

TABLE 1. The results for methods trained on ''clean'' TIMIT.

Methods	EER [%]	C_{det}^{min}
GMM + i-vector	12.95	0.8237
NIFS	12.69	0.8225
Denoised i-vector	11.95	0.8213
FWD-EM	11.24	0.8090
DNN-FE	9.65	0.7885
Our algorithm	9.54	0.7832

For the second group of experiments UBM and PLDA were trained with contaminated TIMIT. To reduce mismatch between training and enrollment/testing conditions, an effective approach is to augment the training dataset by noisy samples. ''Noised'' train data was generated artificially using a MATLAB tool provided in the REVERB challenge [35]. In contrast to ''clean'' data, the ''noised'' data was obtained by distorting 50% of the audio recording. Babble noise and reverberation were added to match real-world conditions. The evaluation values of this group of experiments are listed in Table 2.

TABLE 2. The results for methods trained on ''noised'' TIMIT.

Methods	EER [%]	C_{det}^{min}
GMM + i-vector	12.58	0.8223
NIFS	12.55	0.8221
Denoised i-vector	11.54	0.8195
FWD-EM	10.22	0.8032
DNN-FE	9.35	0.7798
Our algorithm	9.12	0.7697

For the third group of experiments, UBM and PLDA models were trained with VoxCeleb1 and VoxCeleb2. Note that there are 60 speakers in VoxCeleb1 and 118 speakers in VoxCeleb2 that overlap with our evaluation dataset, i.e. SITW. The utterances from the 178 speakers were firstly removed from the dataset prior to training. Finally, a total of 1236567 utterances from 7185 speakers were used to train the models. The performance of the model is given in Table 3.

TABLE 3. The results for methods trained on VoxCeleb1 and VoxCeleb2.

Methods	EER [%]	C_{det}^{min}
GMM + i-vector	5.75	0.4825
NIFS	5.67	0.4804
Denoised i-vector	5.58	0.4725
FWD-EM	5.54	0.4717
DNN-FE	4.98	0.4588
Our algorithm	4.72	0.4505

Generally, all the algorithms demonstrated superiority over the GMM+i-vector baseline, as reported by other research groups and also seen from Table 1, 2 and 3. By utilizing the weighting scheme in GMM and taking the weighting scheme of (17), the proposed algorithm showed better performance than four other recently proposed algorithms for robust speaker recognition. By training the model with datasets

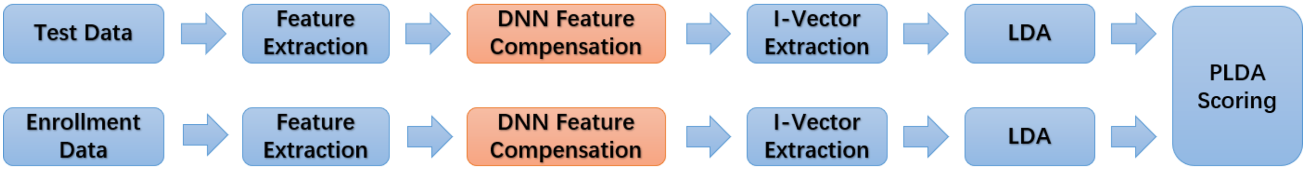


FIGURE 4. DNN feature compensation module in SR system during enrollment and test stage.

created by adding noise and reverberation to the clean TIMIT training set, the training and test mismatch issues are alleviated By increasing the scale of training data, the robustness of the system has been significantly improved and EER has been reduced from 12.95 to 5.75 Detailed comparison and discussion on the baselines are given as below.

1) NIFS ALGORITHM

By adding miscellaneous types of noises artificially, NIFS selects noise invariant frames from utterances, where a hard threshold is chosen and frames with energy lower than the threshold are all discarded. NIFS has an obvious problem where useful information could also have been discarded due to the subjectively chosen threshold. Compared to NIFS, our algorithm adopts a soft strategy and does not discard any frames but just giving noise invariant frames higher weights and the remaining frames lower weights. The experimental results on SITW demonstrated the superiority of the proposed algorithm over NIFS, as shown in Table 1, 2 and 3.

2) DENOISED i-VECTOR

Denoised i-vector is to estimate a clean i-vector given its noisy counterpart [14]. The method is based on the hypothesis that the probabilistic distribution of the noise is Gaussian and the mixture is additive in the i-vector space. Maximal A Posteriori is hereby computed as an estimation of the clean i-vector.

However, the two assumptions may not be satisfied in the real-world scenario. Compared to Denoised i-vector, our algorithm does not make any assumption on the possible distribution of i-vectors, which would not introduce any additional error to deteriorate the model's final performance, especially when the density of i-vectors is not Gaussian. Significant improvements were observed by comparing Denoised i-vector and our algorithm on the results of SITW in Table 1, 2 and 3.

3) FWD-EM

FWD-EM is another recently proposed algorithm to introduce weights to data points in GMM learning. By using some approximation and simplification, the authors move the exponential factor of the modified Gaussian distribution to the denominator of the covariance matrix. FWD-EM is subsequently derived for the case with fixed weight for each data point. FWD-EM solves the same problem as our algorithm but with slightly different objective function and the consequent updating rules. FWD-EM holds good probabilistic probability but introducing additional approximation and assumption; while our algorithm directly imposes data weights to the log likelihood without any probabilistic interpretation nor simplification. The experimental results on SITW showed that our algorithm outperformed FWD-EM significantly as seen in Table 1, 2 and 3. One interpretation to this outcome might be that the real-world data failed to fit the assumptions taken in the approximation and simplification when constructing their objective function of FWD-EM.

4) DNN-FE

Given the results of our experiments on SITW, DNN-FE turns out the most competitive one among the baseline algorithms. By following the recipe in [13], DNN-FE is utilized to enhance the cepstral features before i-vector extraction. The DNN is trained from parallel data of clean and noise corrupted speech which are aligned in the frame level. The training data is from VoxCeleb1 and VoxCeleb2. To generate training data for DNN-FE, "noised" training data is generated using the MATLAB tool provided in the REVERB challenge. In contrast to "clean" data, the "noised" data is obtained by distorting 50% of audio recording. Babble noise and reverberation are added to match real-world conditions. The features involved in DNN-FE are 39-dimensional MFCC features with 25ms window and 10ms shift. To predict clean MFCCs of frame, the sizes of the DNN is 429×2048×2048×2048×39, where the input feature is 11 consecutive frames each of which is represented by a 39- dimensional MFCC extracted from noisy speech, specifically 5 pasts and 5 futures along with the current frame, as shown in Fig.4 The target is a 39-dimensional MFCC from the corresponding frame of the clean speech. There are three hidden layers with 2048 nodes per layers in the DNN. Each layer has a linear transform and a nonlinear activation function. The input vector is linearly transformed by $W^{(1)}$ and $b^{(1)}$ first, then goes through a sigmoid activation function to form the output of hidden layer which is then forward further to the subsequent layers till the output layer of DNN. A linear activation function in the output layer is used to formulate the regression task. Mean Square Error (MSE) on MFCCs is taken as the objective function in DNN. Stochastic gradient descent algorithm is used to train the network parameters. The trained network could predict clean features by de-nosing the noisy features. Fig.5 shows that the trained DNN performs as a plug-in tool in the enrollment and testing stage

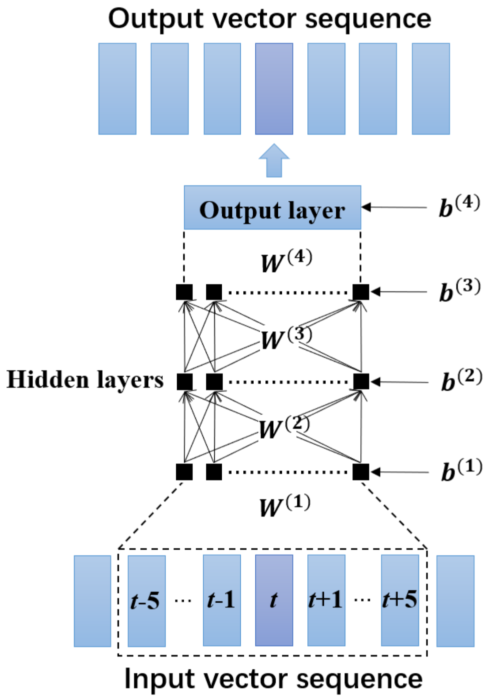


FIGURE 5. DNN structure for feature compensation.

of speaker recognition. The role of DNN-FE is to transform noisy MFCCs to clean ones towards better overall performance on PLDA scoring.

In fact, DNN tries to convert the contaminated MFCCs to corresponding clean versions; while our algorithm selects the relatively clean frames. Given their similarity on improving the quality of the input MFCC features, our algorithm performed similarly with DNN-FE by only showing marginal advantage on SITW. However, without a deep learning frontend, our algorithm holds relatively low computing complexity and does not require additional backend deep software.

VI. CONCLUSION

In the experience of human's listening to identify a speaker's identity, we do not need all the information we heard. In fact, some segments in speech, which are stronger than background noises, could have made positive contributions to the recognition performance. Our proposed algorithm actually modeled this phenomenon by introducing weight parameters to the frames of the input speech. With a modified objective function, new updating rules of Gaussian posteriori probabilities were derived and utilized in i-vector extraction. Experiments demonstrated the effectiveness of the proposed algorithm w.r.t. existing ones. Future work would be investigating more elegant ways to measure the noise-robustness of the input frames.

ACKNOWLEDGMENT

(Xingyu Zhang and Xia Zou contributed equally to this work.)

REFERENCES

[1] D. A. Reynolds and D. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

[2] N. Dehak, P. J. Kenny, R. Dehak, D. Pierre, and O. Pierre, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 4, pp. 788–798, May 2011.

[3] A. Torfi, J. Dawson, and N. M. Nasrabadi, "Text-independent speaker verification using 3D convolutional neural networks," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, San Diego, CA, USA, Jul. 2018, pp. 1–6.

[4] P. Matějka *et al.*, "Analysis of DNN approaches to speaker identification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 5100–5104.

[5] S. Ranjan and J. H. L. Hansen, "Improved gender independent speaker recognition using convolutional neural network based bottleneck features," in *Proc. INTERSPEECH*, Stockholm, Sweden, 2017, pp. 1009–1013.

[6] C. Li *et al.* (May 2017). "Deep speaker: An end-to-end neural speaker embedding system." [Online]. Available: https://arxiv.org/abs/1705.02304

[7] M. R. Islam, M. F. Rahman, and M. A. G. Khan, "Improvement of speech enhancement techniques for robust speaker identification in noise," in *Proc. 12th Int. Conf. Comput. Inf. Technol.*, Dhaka, Bangladesh, Dec. 2009, pp. 255–260.

[8] A. El-Solh, A. Cuhadar, and R. A. Goubran, "Evaluation of speech enhancement techniques for speaker identification in noisy environments," in *Proc. 9th IEEE Int. Symp. Multimedia Workshops (ISMW)*, Beijing, China, Dec. 2007, pp. 235–239.

[9] B. Bharathi, S. Kavitha, and K. M. Priya, "Speaker verification in a noisy environment by enhancing the speech signal using various approaches of spectral subtraction," in *Proc. 10th Int. Conf. Intell. Syst. Control (ISCO)*, Coimbatore, India, Jan. 2016, pp. 1–5.

[10] O. Plchot, L. Burget, H. Aronowitz, and P. Matějka, "Audio enhancing with DNN autoencoder for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 5090–5094.

[11] M. Kolbœk, Z.-H. Tan, and J. Jensen, "Speech enhancement using long short-term memory based recurrent neural networks for noise robust speaker verification," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, San Diego, CA, USA, Dec. 2016, pp. 305–311.

[12] Michelsanti, Daniel, and Z. H. Tan. (2017). "Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification." [Online]. Available: https://arxiv.org/abs/1709.01703

[13] S. Du, X. Xiao, and E. S. Chng, "DNN feature compensation for noise robust speaker verification," in *Proc. IEEE China Summit Int. Conf. Signal Inf. Process. (ChinaSIP)*, Chengdu, China, Jul. 2015, pp. 871–875.

[14] D. Matrouf, W. B. Kheder, P.-M. Bousquet, M. Ajili, and J.-F. Bonastre, "Dealing with additive noise in speaker recognition systems based on i-vector approach," in *Proc. 23rd Eur. Signal Process. Conf. (EUSIPCO)*, Nice, France, Aug./Sep. 2015, pp. 2092–2096.

[15] H. Yu, T. Hu, Z. Ma, Z.-H. Tan, and J. Guo, "Multi-task adversarial network bottleneck features for noise-robust speaker verification," in *Proc. Int. Conf. Netw. Infrastruct. Digit. Content (IC-NIDC)*, Guiyang, China, Aug. 2018, pp. 165–169.

[16] L. P. Wong and M. Russell, "Text-dependent speaker verification under noisy conditions using parallel model combination," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Salt Lake City, UT, USA, vol. 1, May 2001, pp. 457–460 .

[17] M.-W. Mak, X. Pang, and J.-T. Chien, "Mixture of PLDA for noise robust i-vector speaker verification," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 1, pp. 130–142, Jan. 2016.

[18] N. Li and M.-W. Mak, "SNR-invariant PLDA modeling in nonparametric subspace for robust speaker verification," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 10, pp. 1648–1659, Oct. 2015.

[19] D. Garcia-Romero, X. Zhou, and C. Y. Espy-Wilson, "Multicondition training of Gaussian PLDA models in i-vector space for noise and reverberation robust speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Kyoto, Japan, Mar. 2012, pp. 4257–4260.

[20] A. Venturini, L. Zao, and R. Coelho, "On speech features fusion, α-integration Gaussian modeling and multi-style training for noise robust speaker classification," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 12, pp. 1951–1964, Dec. 2014.

[21] L. Besacier and J. F. Bonastre, "Frame pruning for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process., (ICASSP)*, Seattle, WA, USA, vol. 2, May 1998, pp. 765–768.

[22] M. McLaren, M. Graciarena, and Y. Lei, "Softsad: Integrated frame-based speech confidence for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brisbane, QLD, Australia, Apr. 2015, pp. 4694–4698.

[23] K. P. Markov and S. Nakagawa, "Text-independent speaker recognition using non-linear frame likelihood transformation," *Speech Commun.*, vol. 24, no. 3, pp. 193–209, Jun. 1998.

[24] S. Song *et al.* (2018). "Noise invariant frame selection: A simple method to address the background noise problem for text-independent speaker verification." [Online]. Available: https://arxiv.org/abs/1805.01259

[25] H. Yu, Z.-H. Tan, Z. Ma, R. Martin, and J. Guo, "Spoofing detection in automatic speaker verification systems using DNN classifiers and dynamic acoustic features," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 10, pp. 4633–4644, Oct. 2018.

[26] Z. Ma, H. Yu, Z.-H. Tan, and J. Guo, "Text-independent speaker identification using the histogram transform model," *IEEE Access*, vol. 4, pp. 9733–9739, 2016.

[27] I. D. Gebru, X. Alameda-Pineda, F. Forbes, and R. Horaud, "EM algorithms for weighted-data clustering with application to audio-visual scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 12, pp. 2402–2415, Dec. 2016.

[28] S. O. Sadjadi, M. Slaney, and L. Heck, "MSR Identity Toolbox v1.0: A MATLAB toolbox for speaker recognition research," *Speech Lang. Process. Tech. Committee Newslett.*, vol. 1, no. 4, pp. 1–32, Nov. 2013.

[29] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc., B (Methodol.)*, vol. 39, no. 1, pp. 1–22, 1977.

[30] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 3, pp. 345–354, May 2005.

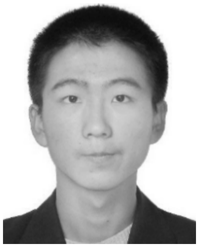
[31] M. McLaren, L. Ferrer, D. Castan, and A. Lawson, "The speakers in the wild (SITW) speaker recognition database," in *Proc. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, San Francisco, CA, USA, Mar. 2016, pp. 818–822.

[32] L. F. Lamel, R. H. Kassel, and S. Seneff, "Speech database development: Design and analysis of the acoustic-phonetic corpus," in *Proc. DARPA Speech Recognit. Workshop*, 1989.

[33] A. Nagrani, J. S. Chung, and A. Zisserman. (2017). "VoxCeleb: A large-scale speaker identification dataset." [Online]. Available: https://arxiv.org/abs/1706.08612

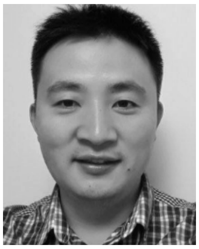
[34] J. S. Chung, A. Nagrani, and A. Zisserman. (2018). "VoxCeleb2: Deep speaker recognition." [Online]. Available: https://arxiv.org/abs/1806.05622

[35] *The REVERB Challenge*. Accessed: 2014. [Online]. Available: https://reverb2014.dereverberation.com



XINGYU ZHANG received the B.S. degree from the Department of Electronic Information Engineering, Hebei University of Technology, Tianjin, China, in 2017. He is currently pursuing the master's degree with the Laboratory of Intelligent Information Processing, Army Engineering University, Nanjing, China. His research interests include speech signal processing, speaker recognition, and machine learning.

XIA ZOU received the Ph.D. degree in multimedia signal processing from the PLA University of Science and Technology, Nanjing, China. He is currently an Associate Professor with Army Engineering University. His research interest includes speech signal processing.



MENG SUN received the Ph.D. degree from the Department of Electrical Engineering, Katholieke University Leuven. He is currently an Associate Professor with Army Engineering University, Nanjing, China. His research interests include speech processing, unsupervised/semi-supervised machine learning, and sequential pattern recognition.

THOMAS FANG ZHENG (M'99–SM'06) received the Ph.D. degree in computer science and technology from Tsinghua University, Beijing, China, in 1997. He is currently a Research Professor and the Director of the Center for Speech and Language Technologies, Tsinghua University. His research focuses on speech and language processing. He has published over 230 papers. He plays active roles in a number of communities, including the Chinese Corpus Consortium (Council Chair), the Standing Committee of the China's National Conference on Man-Machine Speech Communication (Chair), the Subcommittee 2 on Human Biometrics Application of Technical Committee 100 on Security Protection Alarm Systems of Standardization Administration of China (Deputy Director), the Asia-Pacific Signal and Information Processing Association (APSIPA) (Vice-President and Distinguished Lecturer) (2012–2013), the Chinese Information Processing Society of China (Council Member and the Speech Information Subcommittee Chair), the Acoustical Society of China (Council Member), and the Phonetic Association of China (Council Member). He is an Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING and the *APSIPA Transactions on Signal and Information Processing*. He is on the editorial board of the *Speech Communication*, the *Journal of Signal and Information Processing*, the *Springer Briefs in Signal Processing*, and the *Journal of Chinese Information Processing*.

CHONG JIA received the Ph.D. degree in communication engineering from the PLA University of Science and Technology. He is currently an Associate Professor with Army Engineering University, Nanjing, China. His research interests are speech and image signal processing.

YIMIN WANG is currently a Researcher with Army Engineering University, Nanjing, China. Her research interests are in quantum algorithm in many body systems and quantum artificial intelligence.