Text-Independent Speaker Identification Using the Histogram Transform Model

Ma, Zhanyu; Yu, Hong; Tan, Zheng-Hua; Guo, Jun

Published in:
IEEE Access

DOI (link to publication from Publisher):
10.1109/ACCESS.2016.2646458

Publication date:
2016

Document Version
Publisher's PDF, also known as Version of record

Link to publication from Aalborg University

Citation for published version (APA):
Ma, Z., Yu, H., Tan, Z-H., & Guo, J. (2016). Text-Independent Speaker Identification Using the Histogram Transform Model. *IEEE Access*, *4*, 9733-9739. [7803586]. https://doi.org/10.1109/ACCESS.2016.2646458

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 10, 2016, accepted December 26, 2016, date of publication January 2, 2017, date of current version January 27, 2017.

Digital Object Identifier 10.1109/ACCESS.2016.2646458

Text-Independent Speaker Identification Using the Histogram Transform Model

ZHANYU MA[1], (Member, IEEE), HONG YU[1], ZHENG-HUA TAN[2], (Senior Member, IEEE), AND JUN GUO[1]

[1]Pattern Recognition and Intelligent System Laboratory, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]Department of Electronic Systems, Aalborg University, 9100 Aalborg, Denmark

Corresponding author: H. Yu (hongyu@bupt.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61402047 and Grant 61511130081, in part by the Beijing Nova Program under Grant 2017045, in part by the Beijing Natural Science Foundation under Grant 4162044, in part by the Scientific Research Foundation for Returned Scholars, Ministry of Education of China, Chinese 111 program of Advanced Intelligence, and in part by the Network Service under Grant B08004.

ABSTRACT In this paper, we propose a novel probabilistic method for the task of text-independent speaker identification (SI). In order to capture the dynamic information during SI, we design super-mel-frequency cepstral coefficients (MFCCs) features by cascading three neighboring MFCCs frames together. These super-MFCC vectors are utilized for probabilistic model training such that the speaker's characteristics can be sufficiently captured. The probability density function (PDF) of the aforementioned super-MFCCs features is estimated by the recently proposed histogram transform (HT) method. To recede the commonly occurred discontinuity problem in multivariate histograms computing, more training data are generated by the HT method. Using these generated data, a smooth PDF of the super-MFCCs vectors is obtained. Compared with the typical PDF estimation methods, such as Gaussian mixture model, promising improvements have been obtained by employing the HT-based model in SI.

INDEX TERMS Speaker identification, mel-frequency cepstral coefficients, histogram transform model, Gaussian mixture model.

I. INTRODUCTION

Speaker identification is a biometric task that has been intensively studied in the past decades [1]–[4]. Given an input speech, the task of SI is to determine the unknown speaker's identity by selecting one speaker from the whole set of speakers registered in the system [4].

Generally speaking, a typical automatic speaker identification system includes three steps as shown in Fig. 1. The first step is feature extraction. In this part the original speech signals are transformed into feature vectors which can represent speaker-specific characteristics. To this end, a lot of features have been considered, *e.g.*, the Mel-frequency Cepstral coefficients (MFCCs) [5], and the line spectral frequencies (LSFs) [2]. Among them, MFCCs are widely utilized in speech processing tasks, *e.g.*, language identification [6], speech emotion classification [7], and speaker identification [8]. In general, these static features are supplemented by their corresponding velocity and acceleration coefficients such that the dynamic information can be partially preserved. For the purpose of preserving the "full" information, some researchers tend to use the static features and their corresponding neighbors directly, rather than the dynamic velocity and acceleration coefficients, as the features to build a SI system. In [2] and [3], LSFs are directly used in super-Dirichlet mixture models and in [9], static MFCCs are used in the deep learning model.

In this paper, we propose a so-called super MFCCs feature, which are generated by the static MFCCs and the corresponding neighbors. The super MFCCs are created by grouping several neighboring static frames together for the purpose of capturing the dynamic information.

The second step is model training. As the extracted features can describe the unique characteristic of an individual speaker, this allows us to classify each speaker by their voices using a probabilistic models [10]. In describing the statistical properties of the extracted features, we train one model for each speaker.

The third step is identification. In this stage, the feature vectors extracted from the unknown person's speech are compared against the models trained in the second step to

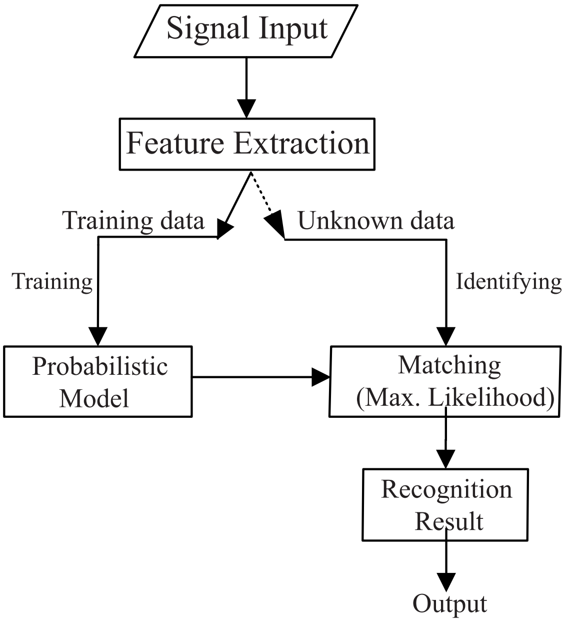


FIGURE 1. A typical speaker identification system.

make the final decision by using the maximum likelihood method.

The effectiveness of a speaker identification system is mainly decided by the design of the statistical model in the second part. The mixture model based methods are widely employed, *e.g.*, Dirichlet mixture model (DMM) [2], [11], [12], beta mixture model (BMM) [13], von-Mises Fisher mixture model [14], [15], and Gaussian mixture model (GMM) [16]–[19]. All these models belong to parametric methods, where the aim of training is to optimize the parameters of the models.

In addition to the mixture model-based methods, nonparametric approaches are also widely applied [20]–[25]. One of the most popular non-parametric approaches is the histogram probability estimation. The training feature space is divided into discrete intervals (*i.e.*, bins), by counting the number of training data that fall into each bin. The distributions of the training data over the whole feature space can be estimated as the probabilities of the bins. Using sufficient large amount of training data and selecting an appropriate bin size, we can get a good estimation performance [26]. However, the probability estimated by the histogram model has large discontinuities [27], especially in high-dimensional space when the multivariate histograms-based method is applied. With the increasing of the feature's dimension, the bin number will raise up at a geometrical level. When the feature dimension is high, we cannot get sufficient training data to ensure most of the bins have data fall in it. Therefore, the empty bins in the feature space have negative effect on the estimation performance.

Recently, a histogram transform (HT) model was proposed to overcome such problem [27]. The HT model can alleviate the discontinuity problem by averaging multiple multivariate histograms. This method has been successfully applied in several applications, such as image segmentation [27], speaker identification [28]. A speaker identification models based on HT model will be introduced in this paper. In the experimental part, we compare the SI performances by using HT and GMM model, respectively. Experimental results show that the HT model perform better than the classical GMM model. This paper is organized as follows: in Section II, we introduce the principles about how to generate the acoustic feature used for speaker model training. Details of HT model construction is described the in Section III. In Section IV, we analysis the experimental results and some conclusions and further work are given in Section V.

II. FEATURE EXTRACTION

As a representation of the short-term power spectrum of a speech signal, the Mel-frequency Cepstral coefficients (MFCCs), which is generated by a linear cosine transform of a log power spectrum on a nonlinear mel-scale of frequency [5], have been widely utilized in many morden speech processing task. The training speeches are segmented in to short frames and from the frame at time t, a D-dimensional MFCC vector can be extracted as

$$\mathbf{x}(t) = [x_1(t), \ldots, x_D(t)]^{\mathrm{T}}. \tag{1}$$

For exploiting the dynamic information, the traditional method constructs a super feature vector that contains the first- and second-order frame-to-frame difference coefficients of the MFCCs (*i.e.*, $\Delta x(t)$, the velocity of MFCCs and $\Delta\Delta x(t)$, the accelaration of MFCCs) [29]. To this end, the super frame is defined as

$$\Delta MFCC_{\text{sup}}(t) = [x(t)^{\mathrm{T}}, \Delta x(t)^{\mathrm{T}}, \Delta\Delta x(t)^{\mathrm{T}}]^{\mathrm{T}}. \tag{2}$$

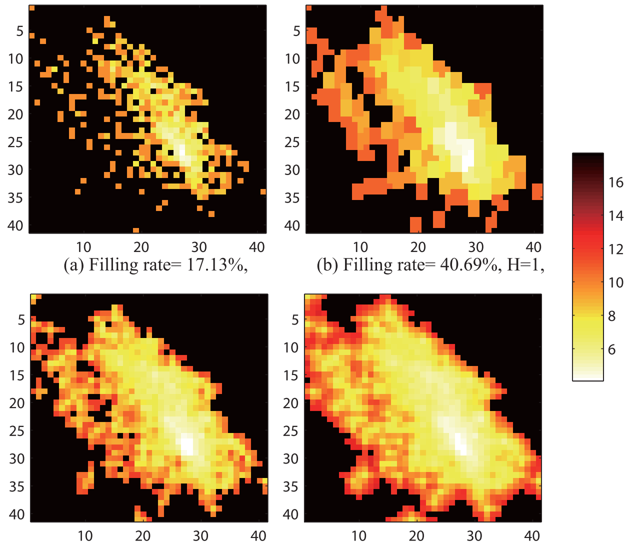
Inspired by the methods introduced in [2], [9], [14], and [15], we propose a novel super frame by grouping two neighbors of the current frame. Set the time interval between two adjacent frames as τ, the super MFCCs frame $x_{\text{sup}}(t)$, which is defined by cascading the current frame and its neighbors, can be obtained as [2]

$$x_{\text{sup}}(t) = [x(t-\tau)^{\mathrm{T}}, x(t)^{\mathrm{T}}, x(t+\tau)^{\mathrm{T}}]^{\mathrm{T}}, \tag{3}$$

where τ is chosen as an integer (*e.g.*, $\tau = 1, 2, 3$). Comparing with the $\Delta\text{MFCC}_{\text{sup}}(t)$ feature, the super MFCCs $x_{\text{sup}}(t)$ contains all the "raw" information among $x(t-\tau)$, $x(t)$, $x(t+\tau)$, which indicates that it can also capture the dynamic information represented in $\Delta\text{MFCC}_{\text{sup}}(t)$ [2]. Moreover, extra information can also be represented by $x_{\text{sup}}(t)$, by using the neighbor frames directly.

III. TRAINING OF THE HT MODELS

In principle, the non-parametric probabilistic models, such as histogram-based models, are driven by training data directly and can simulate any complicated probability density function (PDF). However, the multivariate histograms-based method, which is one of the histogram-based method, is rarely applied, due to the fact that the learned PDF has large discontinuities over the boundaries of the bins.



(a) Filling rate= 17.13%,
(b) Filling rate= 40.69%, H=1,
(c) Filling rate= 43.44%, H=10,
(d) Filling rate= 59.81%, H=50.

FIGURE 2. The original histogram and the transformed ones via HT. (a) The original one. (b) The histogram with one HT. (c) The average histogram with 10 HTs. (d) The histogram with 50 HTs. Filling rate is the ratio of the number of non-zero bins to the total number of bins. H denotes the number of transformations. The values of the negative logarithm of PDF are plotted. The black color denotes zero density and white color presents the highest density.

Figure 2(a) illustrates the negative logarithm of PDF estimated for two randomly selected dimensions of 48-dimensional x_{sup} features using the original histogram method. The 16-dimensional MFCCs vectors $\mathbf{x}(t)$ are extracted from wide-band speech in the TIMIT dataset [30]. The feature space is segmented into 40×40 bins and only 17.13% of the 1600 bins have been filled while the rest yield zero (black color).

In order to get a smooth PDF, parametric probabilistic models, such as mixture models, are usually employed [31]–[33]. In these models, the combination of some simple smooth functions are recommended to estimate the actual PDF. If the function form and the number of mixture components are chosen appropriately, the mixture models can fit the real probability distribution well. However, when the actual PDF is overcomplex, the combination of several simple functions can not represent the true PDF properly.

Recently, an HT model was proposed in [27]. In this model, by H random affine transformations, one training data set can be converted into H data sets. By computing the average histogram of H transformed datasets, we can estimate a smooth PDF more precisely. As shown in Fig. 2(b), after one transform, some points fall in the bins where the original histogram has zero density and 40.69% bins have been filled (comparing with 17.13% in the original one). The PDFs estimated by the average histogram of 10 and 50 transformations are shown in Fig. 2(c) and 2(d), respectively. It is observed that the PDFs have been more smoother than the original one and the filling rates increase to 43.44% and 59.81%, respectively. Hence, the discontinuity in the original histogram has then been overcome.

The HT model is based on histogram methods, and it has advantage of strong adaptability. Meanwhile, the transformation can overcome the disadvantage of discontinuity. A parametric probability density function is adopted in this model as prior, therefore, some merits of parametric models are also preserved in this method.

The random affine function is defined as [27]

$$AF(x; A, b) = Ax + b, \tag{4}$$

where x is a training sample vector of size $D \times 1$, A is a $D \times D$ matrix, and b is a $D \times 1$ vector. After H times randomizing transforms, one training dataset $X = [x_1, \ldots, x_N]$ with N samples is mapped to H training datasets (with $H \times N$ samples in total). By using the averaged histogram of these datasets, we learn a more smoother PDF, where the discontinuous problem [27] can be partly solved. The probability of an input feature x_{in} in the HT method is calculated as

$$\text{HT}(x_{\text{in}}; X) = \pi_0 \text{P}(x_{\text{in}}|X)_0 + \frac{1-\pi_0}{H} \sum_{i=1}^{H} \text{P}(x_{\text{in}}|A_i, b_i, X). \tag{5}$$

The first item of (5) is a priori probability of finding a test sample in a zone where the histograms yield zero density, π_0 is defined as $\pi_0 = (N+1)^{-1}$ and $\text{P}(x_{\text{in}}|X)_0$ is defined as a multivariate Gaussian distribution,

$$\text{P}(x_{\text{in}}|X)_0 = \mathcal{N}(x_{\text{in}}; \mu, C), \tag{6}$$

where

$$\mu = \frac{1}{N} \sum_{j=1}^{N} x_j,\ C = \frac{1}{N-1} \sum_{j=1}^{N} (x_j - \mu)(x_j - \mu)^{\text{T}}. \tag{7}$$

The second item in (5) describes the average histogram probability and $\text{P}(x_{\text{in}}|A_i, b_i, X)$ is the histogram probability of the input data in the i-th transform. Following the suggestion in [27], by adjusting the scale factor of A, the bin width h^* on the transformed space can be chosen as $h^* = 1$. Set

$$y_{i,\text{in}} = \text{round}\left(\text{AF}(x_{\text{in}}; A_i, b_i)\right), \tag{8}$$

$$y_{ij} = \text{round}\left(\text{AF}(x_j; A_i, b_i)\right), \tag{9}$$

where round function means changing the components of the transformed vector to the nearest integer, the histogram probability of input data x_{in} in the i-th transform is defined as

$$\text{P}(x_{\text{in}}|A_i, b_i, X) = \frac{1}{N v_i} \sum_{j=1}^{N} \text{II}(y_{i,\text{in}}, y_{ij}). \tag{10}$$

In (10), v_i is the D-dimensional volume of the histogram bins in the input space, which is defined as

$$v_i = |A_i|^{-1}. \tag{11}$$

And II stands for the indicator function, which is

$$\text{II}(x,y)=\begin{cases}1, & x=y\\ 0, & x\neq y.\end{cases}\tag{12}$$

The transform parameters A and b should take the following rules. Since the bin width on the transformed space is $h^*=1$, we draw b from the uniform distribution over the hypercube $[0,1]^D$.

In the above equations, A can be expressed as the product of a unit rotation matrix U and a diagonal scaling matrix Λ. The random unit rotation matrix U is usually generated by making QR decomposition on a standard normal random matrix [34]. The diagonal elements of Λ, λ_k, can be generated using Jeffreyąŕs prior for the scale parameters [35]. To this end, $\log(\lambda_k)$ should be drawn from the uniform distribution over certain interval of real numbers $[\log(\lambda_{\min}), \log(\lambda_{max})]$, where

$$\log(\lambda_{\min})=\theta_{\min}+\log(\hat{\lambda}),\tag{13}$$
$$\log(\lambda_{\max})=\theta_{\max}+\log(\hat{\lambda}),\tag{14}$$

where $\theta_{\min}$ and $\theta_{\max}$ are tunable parameters. In this paper we empirically choose $\theta_{\min}=0$ and $\theta_{\max}=2$. To make the bin width on the transformed space equal to 1, according to the multivariate histograms theory [36], $\hat{\lambda}$ should be calculated as

$$\hat{\lambda}=\frac{N^{\frac{1}{2+D}}}{3.5}\sqrt{\frac{D}{\text{Tr}(C)}}.\tag{15}$$

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

We evaluated the HT based SI performance on the TIMIT database [30]. The TIMIT database contains 630 male and female speakers coming from 8 different regions and each speaker has 10 utterance recordings. During each round of evaluation, we randomly selected 100 speakers from the database.

The speech was segmented into frames with a 25 ms window size and a 10 ms step size. The silence frames were removed. For each frame, a Hann window was used to reduce the high frequency components. The dimension of the static MFCCs is set as 16.

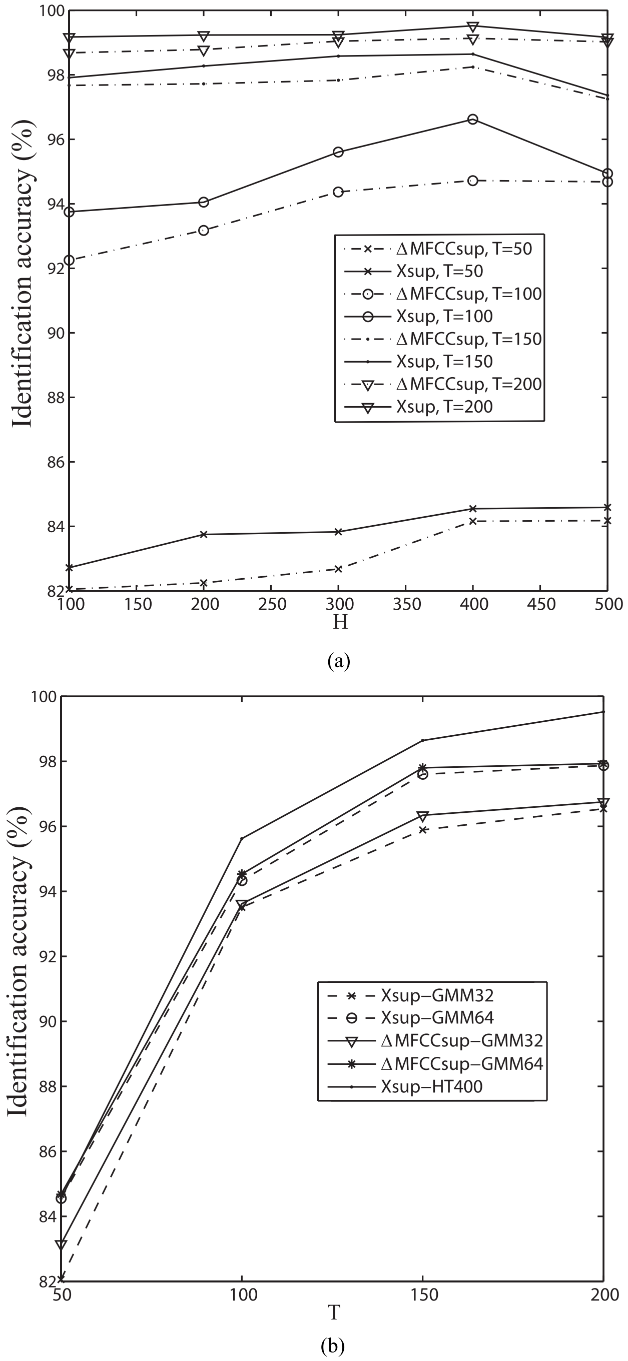
In order to compare the traditional $\Delta\text{MFCC}_{\text{sup}}$ and the super frame x_{sup} proposed in this paper, $\Delta x(t)$ and $\Delta\Delta x(t)$) were also calculated according to the methods described in Section II. Finally, two sets of 48-dimensional super frames were used for speaker model training, respectively.

In the training phase, seven utterances were randomly selected from each speaker as the training data and the remained three utterances were used for testing. In each test utterances we randomly intercepted 10 segments, each including T consecutive frames, as test sets. Hence, there were $3\times 10\times 100=3000$ test sets in total. For MFCC_{sup} and x_{sup}, we trained 100 HT models, respectively.

The log-likelihood of test segment was then calculated as

$$\text{L}_j(\tilde{X})=\sum_{i=1}^{T}\log\left(\text{HT}(x_i;X_j)\right),\tag{16}$$

where $\tilde{X}$ is the test segment set with T feature frames, x_i denotes the i-th frame and X_j stands for the training set of the j-th person. The trained model that yielded the largest log-likelihood value was considered to have the same statistical



(a)

(b)

FIGURE 3. **Comparison of identification accuracies. (a) SI performance of $\Delta\text{MFCC}_{\text{sup}}$ and x_{sup} with the HT model. (b) SI performance of the HT and GMM.**

property as the test feature set, and, therefore, we assigned the test segment with the identity of this trained model.

During evaluations, we set the number of transforms H as $\{100, 200, 300, 400, 500\}$ and the frame interval $\tau = 1$. The frame number T in each test set was chosen as $\{50, 100, 150, 200\}$, which means the durations of each test utterance are $\{0.5, 1, 1.5, 2\}$ seconds, respectively. The identification score is calculated by the number of correctly identified test sets divided by the total number of test sets. We ran 10 rounds of evaluations and the averaged scores in different parameter and methods were reported in Fig. 3.

The performance comparisons of using $\Delta\mathrm{MFCC}_{\mathrm{sup}}$ and x_{sup} in HT model is shown in Fig. 3(a). The HT model trained by x_{sup} reaches higher identification accuracies, for a wide range of numbers of transforms. This indicates that the proposed x_{sup} features are more suitable for the HT model than the conventional $\Delta\mathrm{MFCC}_{\mathrm{sup}}$, for the task of SI. As introduced in Section III, the affine transform matrix A is generated according to a single parameter $\hat{\lambda}$, so the feature x_{sup} in which all components have similar attribute fits the HT model better.

The results also show that the number of transforms H affects the final score. Increasing H improves the identification accuracy, but when H is higher than 400, the accuracy decreases instead. This indicates that too many transformations will make the estimated PDF over-smooth and will reduce speaker specific information. For example, when the speech duration is longer, *e.g.*, more than 0.5s, we have sufficient amount of feature frames to describe the speaker's characteristics, and less error caused by one frame can be compensated by the average of other frames. Hence, we want to increase the "specificity" of each frame, which means we want a "cliffy" PDF curve. Therefore, smaller H is required in this case. However, when less amount of feature frames are presented, the requirement of smoothness get higher. Thus, larger H should be employed to obtain a smooth PDF curve.

We also investigated the performance of GMM models with mixture numbers $\{32, 64\}$ and compared it with HT model. The results are illustrated in Fig. 3(b). Comparing with x_{sup}, the $\Delta\mathrm{MFCC}_{\mathrm{sup}}$ features perform better when using the GMM model. This means that the $\Delta\mathrm{MFCC}_{\mathrm{sup}}$ features are more suitable for the GMM model, which verifies the well-known strategy in SI tasks. Moreover, when the number of test segments is relatively larger (*e.g.*, more than 50 frames) the $x_{\mathrm{sup}} + \mathrm{HT}$ methods can get lower error rates than the $\Delta\mathrm{MFCC}_{\mathrm{sup}} + \mathrm{GMM}$ method.

The boxplots in Fig. 4 compare the precision and stability between the $x_{\mathrm{sup}} + \mathrm{HT}$ method (with $H = 400$) and GMM+$\Delta\mathrm{MFCC}_{\mathrm{sup}}$ method (setting the number of components as 64). We can observe that, when $T = 50$, the HT model's identification accuracy is lower (but more compact) than the traditional GMM model, when the durations of the test utterance data are longer (*e.g.*, $T = 100, 150, 200$), the $x_{\mathrm{sup}} + \mathrm{HT}$ method obtains more accurate and stable results.

In order to check the statistical significance of the improvement, we analyzed the statistical independence of these two models by student's t-test. We assumed the identification results from these two models obey independent random normal distributions with equal means and equal but unknown

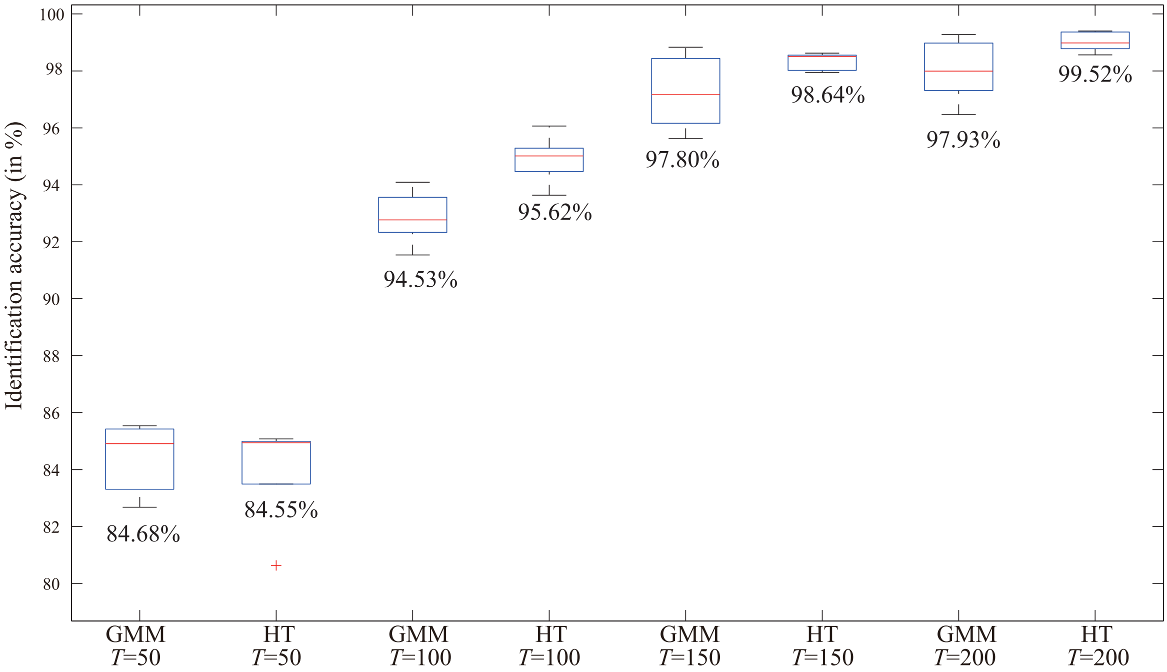


FIGURE 4. Comparisons of the identification accuracies between GMM with 64 components using $\Delta\mathrm{MFCC}_{\mathrm{sup}}$ features and HT model with $H = 400$ using $\Delta\mathrm{MFCC}_{\mathrm{sup}}$ features in different duration T. The central red mark is the median,the edges of the box are the 25^{th} and 75^{th} percentiles. The outliers are marked with red crosses and the mean values are listed below each box.

TABLE 1. Student's t-test for the null-hypothesis that $x_{\mathrm{sup}} + \mathrm{HT}$ and $\Delta\mathrm{MFCC}_{\mathrm{sup}} + \mathrm{GMM}$ are similar.

T	50	100	150	200
p-value	0.1748	0.0030	0.0158	0.0193

variances. The p-values in different T are shown in Table 1. It can be observed that when $T = 50$, p-value is larger than 0.05, which means the null-hypothesis that $x_{\mathrm{sup}} + \mathrm{HT}$ and $\Delta\mathrm{MFCC}_{\mathrm{sup}} + \mathrm{GMM}$ are similar cannot be rejected. Hence, when $T = 50$, GMM model and HT model have similar identification performance, although the GMM model achieves higher average identification accuracy. When T is larger than 50, the p-values are smaller than 0.05, which indicates the improvement obtained by the HT model over the GMM model is statistically significant.

Through the above experiments, we can conclude that the $x_{\mathrm{sup}} + \mathrm{HT}$ model performs better, in general, than the $\Delta\mathrm{MFCC}_{\mathrm{sup}} + \mathrm{GMM}$ model, which is mainly due to the fact that the HT model can fit the complicated probability distribution better. This also encourages us to apply the HT model to improve the performance of other GMM based speech processing system, *e.g.*, speech recognition system based on the GMM+HMM model.

V. CONCLUSIONS AND FURTHER WORK

In this paper histogram transform (HT) model is applied to the speaker identification (SI) task. A novel dynamic acoustic feature, which is produced by grouping adjacent static mel-frequency cepstral coefficients (MFCCs) frames was used for model training. The identification accuracies were improved by using synthesized features generated through the random transform method. By selecting a reasonable number of transforms, more train features were generated to estimate the histogram. The experimental results show that, comparing with the traditional GMM model, the HT model make promising improvement for SI tasks.

In the future we can consider to use some other features, *e.g.*, the line spectral frequencies (LSFs) in the HT model. Some other distributions, *e.g.*, Dirichlet distribution or beta distribution, can also be used to replace the Gaussian distribution as the prior distribution to estimate the probability of the zero zones of the histogram. Recently, some researches showed that fusion of several different systems effectively improves SI performance [37]. Therefore, it is also worth to consider fusion of the HT model with the state-of-the-art i-vector based method.

REFERENCES

[1] S. Nakagawa, L. Wang, and S. Ohtsuka, "Speaker identification and verification by combining MFCC and phase information," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1085–1095, May 2012.

[2] Z. Ma and A. Leijon, "Super-Dirichlet mixture models using differential line spectral frequencies for text-independent speaker identification," in *Proc. INTERSPEECH*, Aug. 2011, pp. 2360–2363.

[3] Z. Ma, A. Leijon, and W. B. Kleijn, "Vector quantization of LSF parameters with a mixture of Dirichlet distributions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 9, pp. 1777–1790, Sep. 2013.

[4] Y. Hu, D. Wu, and A. Nucci, "Fuzzy-clustering-based decision tree approach for large population speaker identification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 4, pp. 762–774, Apr. 2013.

[5] M. Sahidullah and G. Saha, "Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition," *Speech Commun.*, vol. 54, no. 4, pp. 543–565, Mar. 2012.

[6] U. Bhattacharjee and K. Sarmah, "Language identification system using MFCC and prosodic features," in *Proc. IEEE Int. Conf. Intell. Syst. Signal Process. (ISSP)*, Sep. 2013, pp. 194–197.

[7] Z. M. Dan and F. S. Monica, "A study about MFCC relevance in emotion classification for SRoL database," in *Proc. IEEE Int. Symp. Elect. Electron. Eng. (ISEEE)*, May 2013, pp. 1–4.

[8] T. Ganchev, N. Fakotakis, and G. Kokkinakis, "Comparative evaluation of various MFCC implementations on the speaker verification task," in *Proc. SPECOM*, vol. 1. 2005, pp. 191–194.

[9] P. Zhou, L. Dai, Q. Liu, and H. Jiang, "Combining information from multi-stream features using deep neural network in speech recognition," in *Proc. IEEE 11th Int. Conf. Signal Process. (ICSP)*, vol. 1. Oct. 2012, pp. 557–561.

[10] M. A. Pathak and B. Raj, "Privacy-preserving speaker verification and identification using Gaussian mixture models," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 2, pp. 397–406, Feb. 2013.

[11] Z. Ma, S. Chatterjee, W. B. Kleijn, and J. Guo, "Dirichlet mixture modeling to estimate an empirical lower bound for LSF quantization," *Signal Process.*, vol. 104, pp. 291–295, Sep. 2014.

[12] Z. Ma, P. K. Rana, J. Taghia, M. Flierl, and A. Leijon, "Bayesian estimation of Dirichlet mixture model with variational inference," *Pattern Recognit.*, vol. 47, no. 9, pp. 3143–3157, Jan. 2014.

[13] Z. Ma and A. Leijon, "Bayesian estimation of beta mixture models with variational inference," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2160–2173, Nov. 2011.

[14] J. Taghia, Z. Ma, and A. Leijon, "On von-Mises Fisher mixture model in text-independent speaker identification," in *Proc. INTERSPEECH*, 2013, pp. 2499–2503.

[15] J. Taghia, Z. Ma, and A. Leijon, "Bayesian estimation of the von-Mises Fisher mixture model with variational inference," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 9, pp. 1701–1715, Sep. 2014.

[16] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Commun.*, vol. 17, no. 1, pp. 91–108, Jan. 1995.

[17] S. Nakagawa, W. Zhang, and M. Takahashi, "Text-independent speaker recognition by combining speaker-specific GMM with speaker adapted syllable-based HMM," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 1. Aug. 2004, pp. 1–81.

[18] X. Cao, Q. Zhao, D. Meng, Y. Chen, and Z. Xu, "Robust low-rank matrix factorization under general mixture noise distributions," *IEEE Trans. Image Process.*, vol. 25, no. 10, pp. 4677–4690, Oct. 2016.

[19] D. Meng and F. de la Torre, "Robust matrix factorization with unknown noise," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1337–1344.

[20] J.-N. Hwang, S.-R. Lay, and A. Lippman, "Nonparametric multivariate density estimation: A comparative study," *IEEE Trans. Signal Process.*, vol. 42, no. 10, pp. 2795–2810, Oct. 1994.

[21] W. K. Härdle, M. Müller, S. Sperlich, and A. Werwatz, *Nonparametric Semiparametric Models*. Berlin, Germany: Springer, 2012.

[22] W. Lin, Y. Mi, W. Wang, J. Wu, J. Wang, and T. Mei, "A diffusion and clustering-based approach for finding coherent motions and understanding crowd scenes," *IEEE Trans. Image Process.*, vol. 25, no. 4, pp. 1674–1687, Apr. 2016.

[23] W. Lin *et al.*, "A tube-and-droplet-based approach for representing and analyzing motion trajectories," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PP, no. 99, page 1–1.

[24] W. Lin, Y. Zhang, J. Lu, B. Zhou, J. Wang, and Y. Zhou, "Summarizing surveillance videos with local-patch-learning-based abnormality detection, blob sequence optimization, and type-based synopsis," *Neurocomputing*, vol. 155, pp. 84–98, May 2015.

[25] J. Lei *et al.*, "A universal framework for salient object detection," *IEEE Trans. Multimedia*, vol. 18, no. 9, pp. 1783–1795, Sep. 2016.

[26] W. N. Venables and B. D. Ripley, *Modern Applied Statistics With S-PLUS*. New York, NY, USA: Springer-Verlag, 2013.

[27] E. Löpez-Rubio, "A histogram transform for probability density function estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 4, pp. 644–656, Apr. 2014.

[28] H. Yu, Z. Ma, M. Li, and J. Guo, "Histogram transform model using MFCC features for text-independent speaker identification," in *Proc. 48th Asilomar Conf. Signals, Syst. Comput.*, Nov. 2014, pp. 500–504.

[29] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer Handbook of Speech Processing*. Berlin, Germany: Springer, 2007.

[30] *Acoustic-Phonetic Continuous Speech Corpus, NIST Speech Disc 1.1-1*, DARPA-TIMIT, 1990.

[31] P. Kaewtrakulpong and R. Bowden, "An improved adaptive background mixture model for real-time tracking with shadow detection," in *Video-Based Surveillance Systems*. New York, NY, USA: Springer, 2002, pp. 135–144.

[32] Z. Zivkovic, "Improved adaptive Gaussian mixture model for background subtraction," in *Proc. 17th Int. Conf. Pattern Recognit. (ICPR)*, vol. 2. Aug. 2004, pp. 28–31.

[33] S. G. Walker, "Sampling the Dirichlet mixture model with slices," *Commun. Statist.-Simul. Comput.*, vol. 36, no. 1, pp. 45–54, 2007.

[34] F. Mezzadri. (Sep. 2006). "How to generate random matrices from the classical compact groups." [Online]. Available: https://arxiv.org/abs/math-ph/0609050

[35] H. Jeffreys, "An invariant form for the prior probability in estimation problems," *Proc. Roy. Soc. London. Ser. A. Math. Phys. Sci.*, vol. 186, no. 1007, pp. 453–461, 1946.

[36] D. W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*. Hoboken, NJ, USA: Wiley, 2015.

[37] O. Plchot *et al.*, "Developing a speaker identification system for the DARPA rats project," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2013, pp. 6768–6772.

ZHANYU MA received the Ph.D. degree in electrical engineering from Royal Institute of Technology (KTH), Sweden, in 2011. From 2012 to 2013, he was a Post-Doctoral Research Fellow with the School of Electrical Engineering, KTH. He has been an Associate Professor with the Beijing University of Posts and Telecommunications, Beijing, China, since 2014. He is also an adjunct Associate Professor with Aalborg University, Aalborg, Denmark, since 2015. His research interests include pattern recognition and machine learning fundamentals with a focus on applications in multimedia signal processing, data mining, biomedical signal processing, and bioinformatics.

HONG YU received the master degree in signal and information processing from Shandong Unviversity, Jinan, Chinan, in 2006. He is currently pursuing the Ph.D. degree with the Beijing University of Posts and Telecommunications, Beijing, China. He is also a Visiting Ph.D. Student with Aalborg University, Aalborg, Denmark, since 2015. From 2006 to 2013, he worked as a lecturer in Ludong Unviversity, Shandong, China. His research interests include pattern recognition and machine learning fundamentals with a focus on applications in image processing, speech processing, data mining, biomedical signal processing, and bioinformatics.

ZHENG-HUA TAN is currently an Associate Professor with the Department of Electronic Systems, Aalborg University, Aalborg, Denmark, since 2001. His research interests include speech and speaker recognition, noise-robust speech processing, multimedia signal, and information processing, human-robot interaction, and machine learning. He has served as an Editorial Board Member/Associate Editor of Elsevier *Computer Speech and Language*, Elsevier *Digital Signal Processing* and Elsevier *Computers and Electrical Engineering*. He was a Lead Guest Editor of the IEEE Journal of Selected Topics in Signal Processing.

JUN GUO received the B.E. and M.E. degrees from the Beijing University of Posts and Telecommunications (BUPT), China, in 1982 and 1985, respectively, the Ph.D. degree from Tohuku-Gakuin University, Japan, in 1993. He is currently a Professor and a Vice President of BUPT. His research interests include pattern recognition theory and application, information retrieval, content based information security, and bioinformatics. He has authored over 200 papers on the journals and conferences, including science, nature scientific reports, the IEEE Transactions on PAMI, pattern recognition, the AAAI, the CVPR, the ICCV, and the SIGIR.