Real-Time Speaker Identification Using Speaker Model Distance

Hossein Zeinali, Hossein Sameti, Hossein Hadian
Department of Computer Engineering
Sharif University of Technology
Tehran, Iran
zeinali@ce.sharif.edu, sameti@sharif.edu, hadian@ce.sharif.edu

***Abstract*—In real-time speaker identification systems, sufficiently fast methods are required. As the number of registered speakers increases, the traditional methods become intractable due to high computational costs. In this paper, a two-step approach is proposed which prunes the speakers' search space in the first step using the distances between models and then finds the target speaker among the candidate speakers in the second step. This method proved empirically effective and increased the speed of the common GMM method more than 28-fold without any decrease in accuracy.**

Keywords—Real Time; Speaker Identification; Speaker Model Distance; GMM.

I. INTRODUCTION

Identifying individuals using biometrics is a common approach to authentication. Various biometrics such as fingerprint, eye (especially retina and iris), and speech are used to achieve this. Using voice for identification is more computationally expensive than fingerprints. Thus, to make it practical one needs to decrease the test stage computational costs.

Speaker identification is the process through which the identity of the speaker is detected. This is done by searching in the set of registered speakers which can be quite time-consuming if there are too many speakers registered in the system. This is necessary for a practical system to be real-time. But running a full search among the speakers leads to high hardware costs. So the need for methods to reduce the computational costs is inevitable.

So far, various methods have been proposed for speed-up in speaker identification systems which can be grouped into two categories in general. The first are the methods that try to reduce computational complexity for each speaker thereby reducing the overall computational complexity in test stage [1, 2]. The second is the set of methods that perform the main computations on fewer speakers by limiting the search space in the test stage. This is usually done in a multi-step manner; where in each step the search space becomes smaller until the target speaker is identified [3-9].

A new two-step method for speed-up is proposed in this article. In the first step, the search space is limited using the speaker model distances. Then the target speaker is selected in the next step. This method proved effective in limiting the search space and so increased the speed of speaker identification system substantially.

In the following Section certain methods which are to be compared are described. In Section 3 the proposed method is explained and the results are presented in Section 4. Finally the conclusions are given in Section 5.

II. BASELINE METHODS

A. GMM baseline method

The GMM (Gaussian Mixture Model) based speaker identification method is implemented as the baseline method in this work. In this method for each speaker a single GMM is used to model the corresponding utterances in the training stage. A GMM is a weighted sum of M components of a multidimensional Gaussian probability distribution given as follows:

$$p(x|\lambda) = \sum_{i=1}^{M} w_i N(x|\mu_i, \Sigma_i)\,, \tag{1}$$

where $\boldsymbol{x}$ is a d-dimensional continuous-valued data vector, $w_i,\ i = 1,\dots,M,$ are the mixture scalar weights and $N(x|\mu_i,\Sigma_i),\ \mathrm{i} = 1,\dots,\mathrm{M},$ are the component Gaussian PDFs. The weights are constrained to $\sum_{i=1}^{M} w_i = 1$.

Given a sequence of T vectors as $X = \{x_1, x_2, \dots, x_T\}$ and assuming the vectors are independent, the GMM likelihood is:

$$p(X|\lambda) = \prod_{i=1}^{T} p(x_i|\,\lambda), \tag{2}$$

and in log scale:

$$log\big(p(X|\lambda)\big) = \sum_{i=1}^{T} log\big(p(x_i|\lambda)\big). \tag{3}$$

In the test stage, T feature vectors $x_i^{test}, i = 1, \dots, T$ are extracted from a test utterance and likelihoods are calculated for all S speaker models using (3). The most likely speaker is then found according to:

$$s = \underbrace{argmax}_{1 \le s \le S} \sum_{i=1}^{T} log\big(p(x_i^{test} | \lambda)\big). \quad (4)$$

Please refer to [8] for more details.

B. Speaker model clustering method

In the baseline method the test utterance likelihood should be computed for all the speakers in the system using (3) and then the speaker with maximum likelihood is chosen as the target speaker. Since the likelihood calculation for all speakers is time-consuming, it would be better to select a subset of the speakers and calculate (3) only for them.

In [8] a method based on clustering of the speakers is proposed and we implement it in this research to compare it with our proposed method. According to this method, in the training stage the speaker models are clustered. Then in the test stage, first the most likely clusters are found and then the search is done only among the speakers which are in these clusters. Certain different methods have been proposed for clustering of speakers all of which are based on the k-means algorithm. The best of them is described here briefly.

In this method which is based on GMM and Kullback-Leibler distance, first a GMM is trained for each speaker. Then each speaker's model is represented as the weighted average vector of its mixtures as follows:

$$\bar{\mu} = \sum_{i=1}^{M} w_i \mu_i \quad (5)$$

After showing each speaker by its average vector, the speakers are clustered. The centroid of each cluster is the speaker whose mean vector has the minimum Euclidean distance to the center of all speakers. The distance measure used for speaker clustering is an approximation of the Kullback-Leibler distance as below:

$$d\left(\lambda_s, \lambda_n^{CR}\right) \approx \frac{1}{M} \sum_{m=1}^{M} \left(\log p\left(X_{s,m}^{train} | \lambda_s\right) - \log p\left(X_{s,m}^{train} \left| \lambda_{\mathrm{n}}^{\mathrm{CR}} \right.\right)\right), \quad (6)$$

where, M is the number of train vectors for s[th] speaker, λ_s is the s[th] speaker model, $\lambda_{\mathrm{n}}^{CR}$ shows the n[th] cluster and $X_{s,m}^{train}$ is the m[th] vector from the training data of the s[th] speaker.

After clustering the speakers' models in the training stage, the clusters which are to be searched are chosen according to (7) in the test stage and the likelihoods are only calculated for the speakers in these clusters.

$$C_n = \underbrace{\arg max}_{1 \le n \le N} \sum_{m=1}^{M'} \log p\left(X_{s,m}^{test} \left| \lambda_n^{CR} \right.\right), \quad (7)$$

In the above relation, λ_n^{CR} is the model representative of the n[th] cluster and $X_{s,m}^{test}$ is the m[th] vector from the test data of the s[th] speaker.

III. The Proposed Method

Distance can often be used as a measure of similarity; therefore it has been used in certain applications for speaker identification. If the number of distance calculations in test stage is low, this method yields acceptable speed. In the method proposed here, 8 distances are calculated for each speaker which needs remarkably fewer computations than the GMM likelihood calculation. Due to the high speed of this method compared to GMM, we propose a two-step method which uses distance measures in the first step to limit the speaker search space. This is done by finding the n-best speakers using the distance measures and then choosing the best speaker among them using one of the common speaker identification methods.

To use distance measures, first we need to assume a model that can be compared using distance measures for each speaker. Therefore each speaker is represented as a set of GMMs. To be able to compare two GMMs effectively and measure their distance, we should use a Universal Background Model (UBM) to obtain the models for each speaker. This is because the indexes of the mixtures in models obtained using UBM is determined completely and thus we can compare the corresponding mixtures easily.

It's better to use probability distribution distance measures since each speaker is represented as a set of GMM models which are in turn composed of several probability distributions. Certain distance measures have been proposed for distributions which are reviewed here for the Gaussian distributions. Assuming P and Q are two one-dimensional Normal distributions:

$$P \sim N(\mu_1, \sigma_1^2), \;\; Q \sim N(\mu_2, \sigma_2^2),$$

we will have the distance measures as follows:

Hellinger:

$$H^2(P,Q) = 1 - \sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2}}\, e^{-\frac{1}{4}\frac{(\mu_1-\mu_2)^2}{\sigma_1^2+\sigma_2^2}} \quad (8)$$

Kullback-Leibler (KL):

$$D_{KL}(P,Q) = \frac{1}{2}\left[log\left(\frac{\sigma_2^2}{\sigma_1^2}\right) + \frac{(\mu_1-\mu_2)^2}{\sigma_2^2} + \left(\frac{\sigma_1^2}{\sigma_2^2} - 1\right)\right] \quad (9)$$

Euclidean:

$$D(P,Q) = |\mu_1 - \mu_2| \quad (10)$$

In general, for the case of more than one dimension we can use the multi-dimensional form of these measures. However, we used the assumption of independence of the cepstral coefficients instead. Therefore to calculate the multi-dimensional distance, the distances are calculated

independently in each dimension and then are summed up to give the multi-dimensional distance.

So far we introduced distance measures for distributions. But we need to compute the distance between the GMMs. To do this, we can use a weighted sum of the distributions which have the same index as follows:

$$D(G_1, G_2) = \sum_{i=1}^{M} w_i D\left(\lambda_1^i, \lambda_2^i\right), \quad (11)$$

where M is the number of Gaussian distributions in a GMM. It can be seen that in (11) we assume the weights of the corresponding distributions in the two models are identical. Thus we should keep the weights fixed in obtaining the speakers model using the UBM. In the Euclidian distance the weights of the distributions as well as their variances should remain unchanged.

In the test stage, first a GMM-UBM model is built using the test data. Then its average distance to each speaker model is calculated and n-best speakers with minimum distances are selected. In the next step, the target speaker is found using the baseline method.

IV. Evaluations and Results

The TIMIT [10] dataset was used for all experiments. This dataset consists of 630 speakers each of which has 10 utterances. The text of two utterances from these 10, are the same across all speakers which are indicated by 'SA' in TIMIT. For each speaker, we used these 2 utterances as the test data, and the rest 8 utterances as the training data. In all experiments, MFCCs extracted from 25ms frames with 15ms overlapping were used as feature vectors. The silence segments were removed using an energy-based VAD and the speech signal was filtered using a pre-emphasis filter (with a 0.975 factor) prior to feature extraction. As in [8] 29 cepstral coefficients along with C0 were used for the baseline and the clustering method and 12 cepstral coefficients along with C0 were used for the proposed method. In both the baseline method and the clustering method a GMM with M=32 distributions was trained for each speaker. The HTK toolkit [11] was used for all experiments.

To show the effect of amount of test data on the results, two different experiments were performed. In the first experiment, the two test utterances were considered separately which means two identification tests were done for each speaker. In this case, the average of test speech was about 3 seconds. In the second experiment, the two test utterances were considered together and therefore one identification test was done for each speaker with a test speech of 6 seconds. We call the first experiment 'One-Utterance' and the second 'Two-Utterance'.

Before training the models for the distance method (our proposed method), the variances in all dimensions were normalized in order to reduce the effect of variances in different dimensions. Otherwise the effect of dimensions with high variances excessively increases while the effect of other dimensions decreases which results in poor Euclidean distance measures. In the KL and Hellinger methods, distance is proportional to the reciprocal of variance hence normalization is not necessary for these distance measures.

In the training stage for the distance method, first a UBM with 64 components was trained using the training data from all speakers. Then for each utterance of each speaker a separate model was trained by updating the mean of the mixtures using the MAP method. Therefore we have 8 GMMs per speaker which forms an 8-tuple reference set for each speaker. For the speaker identification in the second step the 32-component GMMs of the baseline method were used.

A. One-Utterance Experiments

In the test stage, first using the test utterance and the UBM a GMM with 64 components was created by MAP adaptation. Then using the aforementioned distance measures the average distance of this model to all the models in the reference sets of each speaker was calculated. Then n-best speakers with minimum distances were chosen. In the second step the likelihoods for these speakers were computed using the corresponding 32-component GMMs. The target speaker was identified as the speaker with maximum likelihood. Fig. 1 shows the speaker identification accuracy as a function of n (number of best speakers).

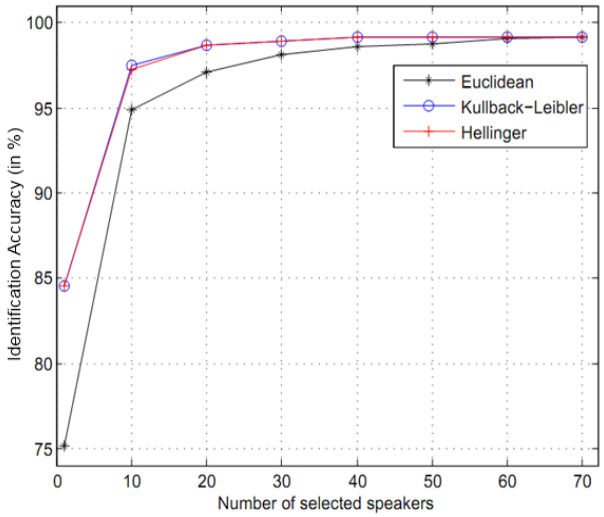

Fig. 1. Speaker identification accuracy as a function of n for One-Utterance case

It can be seen in Fig. 1 that the results of the KL and Hellinger methods are very close and better that the results of the Euclidean measure. Since these two distance measures have been proposed for distributions and consider the effect of variance they were expected to perform better than the Euclidean distance and this fact is apparent in this figure.

In TABLE I the results from baseline methods and the proposed method for different distance measures and for different values of N are presented.

TABLE I. Speed-up results for One-Utterance Experiments

Method	Speed-Up	Accuracy
GMM Base line	1×	99.12%
Clustering	2.8×	97.23%

Euclidean	n = 1	357×	75.16%
	n = 10	48.4×	94.92%
	n = 30	19×	98.09%
	n = 50	11.9×	98.73%
	n = 70	8.6×	99.12%
Kullback-Leibler	n = 1	353×	84.52%
	n = 20	28.6×	98.65%
	n = 40	**15.1×**	**99.12%**
Hellinger	n = 1	232×	84.52%
	n = 20	27.4×	98.65%
	n = 40	**14.6×**	**99.12%**

TABLE I shows that the proposed method is remarkably more efficient than the clustering method. In addition the KL measure has led to better results in comparison with other distance measures and in the best case it has given a 15.1-fold speed-up without degrading the accuracy. This speed-up is several times the speedup achieved by the clustering method.

B. Two-Utterance Experiments

In this case, a model was created for each utterance as in the previous case. Then the average distance of each model to the reference sets of each speaker was calculated and the overall distance was computed as the summation of these two distances. The rest is the same as the 'One-Utterance' experiments. The accuracy as a function of n is shown in Fig. 2.

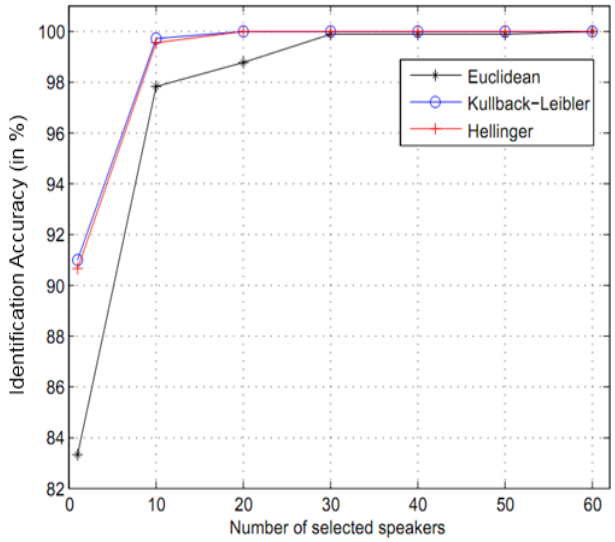

Fig. 2. Speaker identification accuracy as a function of n for Two-Utterance case

As it can be seen in Fig. 2, the best result is achieved for the two distance measures of KL and Hellinger by setting N=20 (best speakers) and for the Euclidean distance by setting N=60.

The comparisons of speed-up and accuracy for this case are presented in TABEL II.

TABLE II. SPEED-UP RESULTS FOR TWO-UTTERANCE EXPERIMENTS

Method		**Speed-Up**	**Accuracy**
GMM Base line		1×	100%
Clustering		2.8×	99.84%
Euclidean	n = 1	357×	83.33%
	n = 10	53×	97.78%
	n = 30	29.2×	98.73%
	n = 50	19.8×	99.84%
	n = 70	9.9×	100%
Kullback-Leibler	n = 1	353×	90.95%
	n = 20	52×	99.68%
	n = 40	**28.8×**	**100%**
Hellinger	n = 1	232×	90.63%
	n = 20	49.6×	99.52%
	n = 40	**27.8×**	**100%**

This table shows that in this case the KL distance measure gives the best results too. In the best case, the proposed method based on the KL distance measure, speeds up the identification process 28.8-fold without any decrease in accuracy.

Comparing tables I and II reveals that the efficiency of the proposed method (in terms of speed-up) increases by the test speech duration. This is for two reasons. Firstly the precision of the distance measures increases as the duration of test utterances increases which lets us choose fewer speakers in the first step (smaller n). The second reason is that the increase in the test speech duration has a negligible effect on the computations (and runtime) of the first step while this increase in duration, decreases the speed of the baseline method linearly.

V. CONCLUTIONS

In this paper, a two-step method based on the speaker model distances was proposed in order to increase the speed of speaker identification. The proposed method remarkably increased the speed of the system in comparison to the clustering method. Specifically, this method could increase the speed 28.8-fold without any decrease in accuracy, and 52-fold with a less that 0.5 percent decrease in accuracy.

REFERENCES

[1] J. McLaughlin, D. A. Reynolds, and T. P. Gleason, "A study of computation speed-UPS of the GMM-UBM speaker recognition system," in *EUROSPEECH*, 1999, pp. 1215-1218.

[2] T. Kinnunen, E. Karpov, and P. Franti, "Real-time speaker identification and verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, pp. 277-288, 2006.

[3] H. S. Beigi, S. H. Maes, U. V. Chaudhari, and J. S. Sorensen, "A hierarchical approach to large-scale speaker recognition," in *EUROSPEECH*, 1999.

[4] H. Zeinali, H. Sameti, and B. Babaali, "A fast Speaker Identification method using nearest neighbor distance," in *Signal Processing (ICSP), 2012 IEEE 11th International Conference on*, 2012, pp. 2159-2162.

[5] T. Kinnunen, E. Karpov, and P. Fränti, "A speaker pruning algorithm for real-time speaker identification," in *Audio-and Video-Based Biometric Person Authentication*, 2003, pp. 639-646.

[6] A. Sarkar, S. Rath, and S. Umesh, "Fast approach to speaker identification for large population using MLLR and sufficient statistics," in *Communications (NCC), 2010 National Conference on*, 2010, pp. 1-5.

[7] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital signal processing*, vol. 10, pp. 19-41, 2000.

[8] V. R. Apsingekar and P. L. De Leon, "Speaker model clustering for efficient speaker identification in large population applications," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, pp. 848-853, 2009.

[9] H. Zeinali, H. Sameti, H. Khaki, and B. BabaAli, "A fast two-level Speaker Identification method employing sparse representation and GMM-based methods," in *Information Science, Signal Processing and their Applications (ISSPA), 2012 11th International Conference on*, 2012, pp. 45-48.

[10] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, p. 27403, 1993.

[11] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, *et al.*, *The HTK book* vol. 2: Entropic Cambridge Research Laboratory Cambridge, 1997.