A Fast Speaker Identification Method Using Nearest Neighbor Distance

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

Abstract— **By increasing the number of registered speakers in Speaker Identification (SI) systems, computation time for identifying an unknown speaker is significantly increased. This problem arises from the simple design of conventional methods. Due to this limitation, we cannot use conventional SI methods in real time applications. In this paper, we propose a two-step method to overcome this limitation. We use different identification methods for each step. In the first step we reduce the search space using Nearest Neighbor method. In the second step we identify the target speaker using the conventional GMM-based SI method. The experimental results show 3.4× speed-ups without any accuracy loss using the proposed method. It is also shown that if the number of selected speakers in first step is furthermore reduced, the identification speed will increase, at the cost of a slight decrease of accuracy. So, there is a trade-off between accuracy and speed-up that can be controlled by a threshold.**

Keywords- Speaker Identification; Real time Application; Nearest Neighbor; GMM

I. INTRODUCTION

In recent years with advances in technology, several methods have been devised for improving security. From these methods we can refer to fingerprint recognition that has many applications in identifying individuals. Another identification method is using human speech to specify speakers. Using speech for identifying speakers has several advantages such as ease of use and the capability to be used in telephony applications. However this method has several limitations. The major disadvantages are reducing accuracy in noisy environments and considerable computation costs for large population applications.

Gaussian Mixture Model (GMM) is the conventional method that is usually used to model speaker features in training stage [1]. This method has good accuracy in large population applications. When the number of speakers that register in the system is in the order of ten thousands or more, use of conventional methods will be impossible, due to extreme time consuming process.

Recently, several studies have been reported on reducing speaker identification processing.

Beigi et al. [2] proposed a hierarchical approach to speed up large-scale speaker recognition. They used a binary tree of trained speaker models that was created in training stage to reduce speaker verification time.

In [3] another hierarchical approach was proposed. The authors grouped together feature vectors of different speakers having similar acoustic characteristics and trained a model for this group of feature vectors. We can think of each group as a cluster of similar speakers. In test stage, they first selected k best clusters and searched target speaker in speakers just from those clusters. They reached 3.3× speedups using this method.

In [4] we showed that if we have a fast method without necessarily very good identification accuracy, we can use a two level method to compensate any loss in the accuracy. Using this method, the speaker identification test stage was very fast. We used sparse representation in the first level to decrease the search space. By selecting n-best speakers from sparse representation results and search target speaker on these, we achieved 18× speedups without any loss of accuracy.

Vijendra et al. [5] proposed a GMM-based speaker model clustering method. In training stage the *k*-means algorithm is used to cluster speaker models. During the test stage the best clusters are selected first, and the search is carried just between speaker models in these clusters. They showed that this method gains 4.4× speed-ups with little to no loss in identification accuracy.

Reynolds et al. [6] proposed a GMM-UBM method that uses top C-best mixtures for likelihood computation. During the test stage, test utterance is first scored against UBM to find the best scoring mixture components. Then, for every frame, indices of the dominant Gaussian components are used for likelihood computation. Using this method the SI task will be about 5 times faster. Although there is a tradeoff between the accuracy and speed-up that can be adjusted by the number of selected best mixtures (C-best).

Several other studies have been reported to speed-up the test stage of speaker identification. Pre-Quantization (PQ) technique proposed in [7] uses a subset of the test utterance feature vectors for likelihood computation. The authors describe several different approaches to select this subset. Speaker Pruning [8] is another technique which is a step by step method to identify the speaker. In each step, a small portion of the test utterance is used to further prune the candidate speakers. Finally, at the last step the target speaker is identified.

In this paper our main concern is to achieve significant speed-up in the test stage of speaker identification. Also, because in SI systems the accuracy is very important, our peripheral goal was to avoid remarkable loss of accuracy. We

propose a two-step method for speaker identification. In the first step the Nearest Neighbor (NN) distance method [9] has been used to score the speakers. By selecting *n*-best speakers with highest scores, and searching target speaker among them, we achieved 3.4× speed-ups without any loss of accuracy.

The paper is organized as follows: In Section II, we describe two SI methods which we have compared our method with. Our proposed method which is a combination of NN technique and conventional GMM-based is described in Section III. In Section IV, experiments and results on TIMIT corpus are presented. We conclude the paper and mention future work in Section VI.

II. COMPARED METHODS

A. Baseline system

In this paper we implement GMM-based Speaker Identification method [1] as the baseline system. In this method each speaker's utterance features are modeled using a single GMM in the training stage. A GMM is a weighted sum of M component of a multidimensional Gaussian density function as given by the below equation,

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

where $\boldsymbol{x}$ is a D-dimensional continuous-valued data vector, w_i, i $= 1,\ldots,$M, are the mixture scalar weights, and $N(x|\lambda_i, \Sigma_i)$, i $= 1,\ldots,$M, are the component Gaussian densities. The mixture weights must satisfy the constraint that $\sum_{i=1}^{M} w_i = 1$.

For a sequence of T vectors X = $\{x_1,\ldots,x_T\}$, the GMM likelihood with independence assumption between the vectors, can be written as

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

And in log domain

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

In the testing stage, T feature vectors $x_i^{Test}, i = 1, \ldots, T$ are extracted out of a test utterance. Then the probability is calculated for all *S* speaker models using a log-likelihood calculation as in Equation (3) and the most likely speaker identity s decided according to

$$s = \underbrace{\arg\max}_{1\le s\le S} \sum_{i=1}^{T} \log\left(p(x_i^{Test}|\lambda_s)\right). \quad (4)$$

B. Speaker Model Clustering Method

In conventional methods such as the implemented GMM-based method, we need to calculate the probability of test feature vectors for each registered speaker. This probability is calculated using Equation (3). This computation is very time consuming. In [5], Speaker Model Clustering method was proposed to overcome this difficulty. At the training stage, all speakers are grouped into *N* clusters. The test stage consists of two steps. First, using the test feature vectors, *n*-best clusters containing the most probable speakers are selected. Then search for target speaker takes place over those clusters. By this two step search, the search space is effectively reduced. In this paper we just implement the best clustering method proposed in [5] for comparison. This method is called Kullback–Leibler, GMM-Based Clustering.

III. PROPOSED METHOD

When the number of registered speakers in system is in the order of 10.000, calculations required to identify a speech at the test stage is very high. If we use a conventional method, we will have to find target speaker in one step. This is a linear search in all registered speakers that is very time consuming and not usable in real time applications or systems that have too many registered speakers. Instead of using one step identification method we can do it in several stages and prune some of speakers in every stage. In this paper we propose a two level method that is a combination of a fast Nearest Neighbor (NN) distance method and conventional GMM-based method to identify the speakers. This method increases the speed of identification while preserving the identification accuracy.

In [9] a NN distance method has been used as the main method for formant based speaker identification. We used the same NN method here as the first step of method. In NN distance method we represented each one of the speakers by a set of mean-vectors. First we trained a UBM from training utterances of all speakers. Then we adapted four GMM-UBMs for each speaker. Finally we got the mixture mean-vectors of each speaker's GMMs and created a reference set of mean-vectors. By doing this, each speaker is represented by a set of mean-vectors. Let r_i be the mean-vector of speaker reference set and R be the correspondening reference set of $\{r_i\}$, $i = 1,\ldots,M$.

During test stage, first we adapted a GMM-UBM using the test utterance. Then we created a test set using the mixture mean-vectors of this GMM. Let u_j be the mean-vector of test set and U be the correspondening test set of $\{u_j\}$, $j = 1,\ldots,K$. By doing this, we reduced the number of feature vectors. This is similar to Vector Quantization method.

We used Euclidean distance measure to calculate the distance between separate vectors. In this method first we calculated NN distance from each mean-vector of the test set to all reference sets. Let $d_{NN}(u_i, R)$ be the NN distance between test vector u_i and reference set R. It is also needed to calculate $r_{NN}(u_i, R)$ that is a mean-vector in reference set R that is closest to u_i (nearest neighbor).

$$d_{NN}(u_i, R) = \underbrace{\min}_{r_j \in R} |u_i - r_j| \quad (5)$$

$$r_{NN}(u_i, R) = \underbrace{\arg\min}_{r_j \in R} |u_i - r_j| \quad (6)$$

In the same manner, we calculated $d_{NN}(r_j, U)$ that is NN distance between reference vector r_j and test set U. Likewise we calculated $u_{NN}(r_j, U)$ which is the nearest neighbor vector to r_j.

$$d_{NN}(r_j, U) = \underbrace{\min}_{u_i \in U} |r_j - u_i| \quad (7)$$

$$u_{NN}(r_j, U) = \underbrace{\arg\min}_{u_i \in U} |r_j - u_i| \quad (8)$$

After calculating NN distances, the distance between two sets R and U is calculated using symmetric distance formula as below,

$$D(U,R) = \frac{1}{M}\sum_{i=1}^{M}\left(d_{NN}(u_i,R) - d_{NN}(r_{NN}(u_i,R), R - r_{NN}(u_i,R))\right)^2 + \frac{1}{K}\sum_{j=1}^{K}\left(d_{NN}(r_j,U) - d_{NN}\left(u_{NN}(r_j,U), U - u_{NN}(r_j,U)\right)\right)^2, \quad (9)$$

where M is the number of mean-vectors in the reference set R and K is the number of mean-vectors in the test set U. Note that if difference in parentheses became negative it is changed to zero. This corresponds to the situation when a vector in one set is closer to another set compared to its own set.

In this two-step method, first we calculated the distance between U and all reference sets R_i, $i = 1,...,S$. Then we sorted speakers based on these results and we found that in most cases the target speaker is among the first speakers. Therefore we selected the n-best speakers from the results generated by this step. We selected the target speaker from the n-best using conventional GMM-based method at the second step. By this two-step method we achieved a significant speed-up without any loss of accuracy.

IV. EXPERIMENTS AND RESULTS

To compare speed-up of our method, we implemented two other existing methods. We implemented conventional GMM-based method as the baseline [1] and Speaker Model Clustering method [5] as the second compared method.

Our speaker identification experiments were done on TIMIT corpus [10]. In this paper we used all 630-speakers. In TIMIT each speaker has 10 utterances and each utterance contains a single sentence. These 10 sentence-texts consist of 2 "SA", 5 "SX", and 3 "SI" sentences. In all tests we used the two SA utterances as the test data. The other 8 utterances were used as the training data.

Since Mel Frequency Cepstral Coefficients (MFCCs) have good performance in SI tasks, we used these features in our implementations. Before extracting features, we used an energy based voice activity detector to remove silence from the utterances. Then for our baseline method, we used 29 MFCCs including C_0 as feature vectors. We modeled every speaker by a 32-mixture GMM. To train GMM of each speaker and to extract features from utterances we used HTK [11] in all experiments. We used the same trained GMMs of speakers for speaker model clustering method.

In order to create a reference set for each speaker that can be used in NN distance method we used the following approach. We used 13 MFCCs including C_0 feature vectors and trained a 32-mixture Universal Background Model (UBM) with the training utterances of all speakers. After that we used map-adaptation to update means of UBM mixture to create a single model for each set of two utterances. Since each speaker has 8 training utterances, 4 training models will be created for each speaker. By grouping the mean vectors of these 4 models, the reference set for each speaker is created. In the first step of our method we used NN distance to select n-best speakers. For the second step, we used the same 32 mixtures GMM that was used for the baseline method.

In the test stage, first we adapted a 32-mixture GMM for the test utterance using the UBM. Then the test set was created from the mean vectors of this model. After that we selected n-best speakers out of all registered speakers with the lowest NN distances. Then, we did likelihood computation only for the n selected speakers and the target speaker was selected.

In the rest of this section we present the experimental results of our method to examine its performance.

A. *Effect of the number of speakers on identification time*

In order to achieve intuition about relation between the number of speakers and time required for SI task, we designed the following test. We ran the test on a computer with Dual-Core CPU with 2.6GHz clock rate. We tested only the baseline method for demonstration.

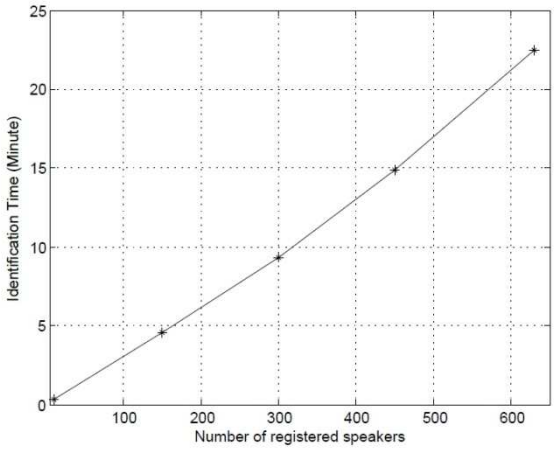

Figure 1. Identification Time as a function of registered speakers

According the results, we can see a linear relation between the required time and the number of speakers. The reason behind this is the fact that the method does a linear search among all speakers. As it is shown in Fig. I, the required time to identify a speaker among 630 speakers is about 22 minutes. This is obviously far from an acceptable time for real time applications. Furthermore if we increase the number of speakers to 10000, we can easily compute that the required time for a single SI task will be about 6 hours. The result of this test supports the need for quicker methods for Speaker Identification.

B. *Evaluation of Proposed Method*

In order to evaluate the proposed method, SI accuracy as a function of the number of selected speakers has been depicted in Fig. 2.

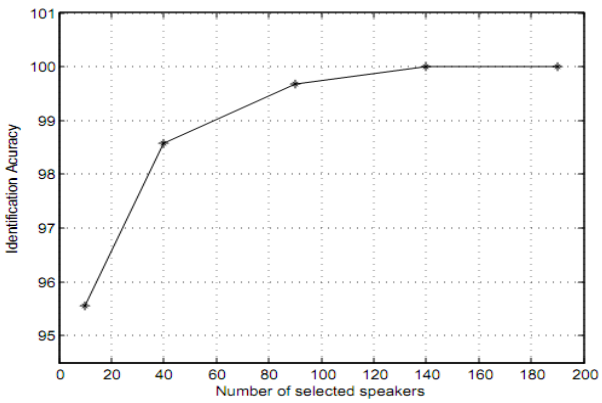

Figure 2. Identification accuracy as a function of number of selected speakers

We can easily see that with selecting only the n = 140 best speakers, we can achieve perfect SI accuracy.

C. *Speed-up Results*

We present the results of our method and the other two implemented methods. We consider the GMM-based method as the baseline method. We report results of our method for 3 different setups. The details of these comparisons are presented in TABLE I. In this table we report both the speed-up factor and the accuracy of each method.

TABLE I. AVERAGE SPEED-UP FACTORS AND SI ACCURACY

Testing Method	Speed-up	Identification Accuracy
GMM-based	1×	100%
Clustering	2.8×	99.84%
Two-Step (n = 40)	7.3×	98.57%
Two-Step (n = 90)	4.6×	99.68%
Two-Step (n = 140)	3.4×	100%

For speaker model clustering method we report the results where the search space was 20% of all clusters. The proposed method achieved 3.4× speed-ups without any loss of accuracy. Generally, there is a trade-off between accuracy and speed-up.

V. CONCLUSIONS AND FUTURE WORK

A new two-step speaker identification algorithm suitable for real time applications and systems with huge number of registered speakers (>10000) was presented in this paper. This research was motivated by our studies on combining different SI methods and designing multi step methods for Speaker Identification task. Our experiments show that by using a two-step method we can identify speakers faster than conventional one-step methods without any loss in identification accuracy.

In this paper we used NN distance as a fast identification method for the first step of our method. In the future, we plan to combine other methods and devise a multi step technique; especially we intend to replace NN distance with some other fast SI method.

ACKNOWLEDGMENT

The authors would like to thank Ali Mahjoob for his help in this work.

REFERENCES

[1] D. A. Reynolds, "Speaker Identification and Verification using Gaussian Mixture Speaker Models," Speech Communication, vol. 17, pp. 91–108, Aug1995.

[2] H. S. M. Beigi, S. H. Maes, J. S. Sorensen, and U. V. Chaudhari, "A hierarchical approach to large-scale speaker recognition," in Proc. Eur. Conf. Speech Commun. Technol. (Eurospeech), 1999, pp. 2203–2206.

[3] B. Sun, W. Liu, and Q. Zhong, "Hierarchical speaker identification using speaker clustering," in Int. Cont. Natural Lang. Proc. Knowledge Engineering, 2003, pp. 299–304.

[4] H. Zeinali, H. Sameti, H. Khaki, B. Babaali, "A Fast Two-Level Speaker Identification Method Employing Sparse Representation and GMM-Based Methods," in 11th International Conference on Information Science, Signal Processing and their Applications, 2012, pp. 80-83.

[5] V. R. Apsingekar, and P. L. D. Leon, "Speaker Model Custering for Efficient Speaker Identification in Large Population Application," IEEE Tran. on Audio, Speech and Language Processing, vol. 17, No.4, May2009.

[6] D. Reynolds, T. Quateri, and R. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," Digital Signal Processing, vol. 10, pp. 19–41, Jan2000.

[7] J. McLaughlin, D. A. Reynolds, and T. Gleason, "A Study of Computation Speed-ups of the GMM-UBM Speaker Recognition System," in Proc. Eur. Conf. Speech Commun. Technol. (Eurospeech), pp. 1215–1218, 1999.

[8] T. Kinnunen, E. Karpov, and P. Franti, "A Speaker Pruning Algorithm for Real-Time Speaker Identification," in Proc. Audio- and Video-Based Biometric Authentication, Guildford, U.K., pp. 639–646, 2003.

[9] P. V. Labutin, S. L. Koval, and A. N. Raev, "Automatic speaker recognition system using telephone speech," unpublished.

[10] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM," LDC catalog number LDC93S1, 1993.

[11] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, "Hidden Markov model toolkit (HTK) version 3.4 user guide," 2002.