

PAPER • OPEN ACCESS

A Group Based Decision-making Mechanism for Speaker Verification

To cite this article: Xiaoyu Yan and Wang Lei 2019 *IOP Conf. Ser.: Earth Environ. Sci.* **252** 042094

View the [article online](#) for updates and enhancements.

A Group Based Decision-making Mechanism for Speaker Verification

Xiaoyu Yan^a, Lei. Wang^b

School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China

^ayanxiaoyu@bupt.edu.cn, ^bwangle_elf@bupt.edu.cn

Abstract. Speaker recognition has been a popular research topic in pattern recognition field. Speaker verification is an important procedure in speaker recognition. It is a process to accept or reject the identity claim of a speaker. How to make a decision is a critical problem in speaker verification system. This paper introduces a group-based mechanism which performs speaker verification system with Joint Factor Analysis (JFA). In this method, we first finds a similar speaker group for each test utterance, and then implements speaker verification. This paper also proposes a new efficient threshold setting method on the use of silhouettes. We carry out our comparison experiment on the TIMIT corpus and the experimental results show that compared with the traditional method, the overall recognition performance of the open-set speaker recognition system based on the new decision mechanism is improved.

1. Introduction

Speaker recognition includes speaker identification and verification [1]. In speaker verification system, we must have a priori threshold during decision-making in real applications. Essentially, a speaker verification system could make two types of mistakes, one is false acceptance (FA) which causes an impostor to be accepted, and the other is the false rejection (FR) which causes a genuine speaker's identity claim to be denied. So our main goal is to obtain an optimal threshold to minimize both FA and FR errors during decision making. Recent years, speaker verification has been increasingly demanded for security in miscellaneous information systems. Therefore, the development of effective speaker verification technologies for use in reality is of utmost importance.

With the previous studies on decision-making, there have been some approaches proposed for this problem. The basic idea of the threshold setting in speaker verification is to obtain an appropriate threshold based on a training data set, and the threshold will be applied to make a decision after identification. Some methods of threshold setting are introduced in [2] based on the score distribution and the speaker dependent threshold is a linear combination of the mean and deviation from clients and impostors scores. And in [2], it also proposes a Scoring Pruning (SP) technique which suppresses the effect of non-representative scores and removes them contributing to a better estimation of means and variances in order to set a suitable threshold. In [3], a nonlinear method based on weighting score distribution using a sigmoid function is proposed. Besides, an adaptive method combined with confidence measurement is proposed for online re-estimation of speaker dependent thresholds in [4].



On the basis of statistics theory, some normalization approaches also have been proposed to help the threshold setting methods against mismatch. Two methods that have shown significant improvement for speaker verification are Within Class Covariance Normalization (WCCN) [5] and Eigen Factor Radial (EFR) [6] which includes the length normalization proposed in [7]. In this work, we used the cohort model approach which finds a set of speakers whose characteristics in speech are similar to a specific speaker so that a cohort model can be built to model cohort speakers' ensemble characteristics [8], while the world model approach is to build a universal speaker model on a pool of speech utterance produced by various speakers [9].

The reminder of this paper is organized as follows. In Section II, the theory used in speaker recognition base system is introduced, including the theory of the joint factor analysis and the usual methods used in threshold setting. In Section III, the proposed algorithms, including the construction of silhouettes, the confirm threshold approach using silhouettes and the group based decision-making mechanism, are described in detail. In Section IV, corresponding experimental results on the TIMIT corpus are present. In section V, discussions and conclusions are drawn.

2. Theory of Speaker Verification Threshold Setting

2.1. Speaker Recognition System Based On Joint Factor Analysis

In this work, we implement the speaker recognition system based on JFA [10]. JFA speaker modeling technique has achieved great success and has become the state of the art method in speaker recognition system.

Recall that the general JFA model assumes that, a speaker utterance is represented by a supervector M that consists of additive components from a speaker and a channel/session subspace. Specifically, the speaker-dependent supervector is defined as:

$$M = m + Vy + Ux + Dz$$

Where m is a speaker and session independent supervector (generally from a Universal Background Model (UBM)), V and D define a speaker subspace (eigenvoice matrix and diagonal residual, respectively), and U defines a session subspace (eigenchannel matrix). The vectors y , z and x are the speaker and session dependent factors in the respective subspaces and each is assumed to be a random variable with a normal distribution $N(0, I)$. To apply JFA to speaker recognition consists of first estimating the subspaces (i.e., V, D, U) from appropriately labelled development corpora and then estimating the speaker-dependent factors (i.e., y, z) for each speaker and session. Scoring is done by computing the likelihood or distance of the test utterance feature vectors against to speaker model. The comparison among several JFA scoring methods is present in [11].

2.2. Theoretical Approaches of Threshold Setting

Most of the methods of the threshold setting are based on statistics, mainly the mean and deviation of the scores distribution from clients and impostors. Besides, they also include some other parameters which affect the recognition results and must be empirically determined. There are two main methods. The first method is called 3σ method [2]. Given a set of data associated with speakers and impostors, intra-speaker's and inter-speaker's scores are achieved. The simplest way is to assume that these scores are subject to Gaussian distribution. In addition, there is a statistical principle especially for a Gaussian distribution $G(x|\mu, \sigma)$; that is, 99.7% out of all the samples drawn from this distribution should be located with only the interval $[\mu - 3\sigma, \mu + 3\sigma]$. So the threshold can be written as:

$$T_s = \begin{cases} \mu - 3\sigma & \text{if } \mu - 3\sigma > \bar{\mu} + 3\bar{\sigma} \\ (\mu\bar{\sigma} + \bar{\mu}\sigma) / (\sigma + \bar{\sigma}) & \text{otherwise} \end{cases}$$

Motivated by Furui's idea, Lindberg et al. propose a threshold setting method [12]. This approach uses a linear combination of the estimates of means on intra-speaker's score and inter-speaker's score to set the threshold. Thus, the threshold is in the following form:

$$T_s = \gamma\mu + (1-\gamma)\bar{\mu}$$

Where γ is a speaker-independent parameter and is empirically determined. This method is simple and does not need to estimate the standard deviation. Besides, even if the data is not sufficient, it could still keep a certain degree of stability.

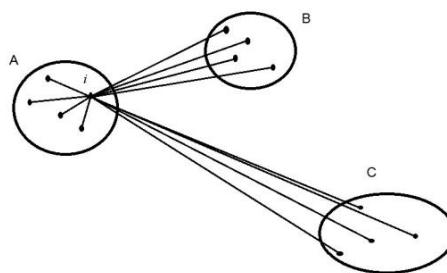


Figure 1. An illustration of the elements involved in the computation of $S(i)$ where the object i belongs to cluster A.

In the real applications, the ordinary threshold setting methods often use two methods introduced above. In this work, we propose a new confirm threshold setting approach and a group-based mechanism for speaker verification.

3. The proposed Algorithm

3.1. Construction of Silhouettes

The silhouette value for each point is a measure of how similar that point is to points in its own cluster compared to points in other clusters, and ranges from -1 to +1. It is defined as (For a concentrate illustration, see Fig. 1):

$$S(i) = \frac{\min((b(i,:), 2) - a(i))}{\max(a(i), \min(b(i,:)))}$$

$a(i)$ is the average of dissimilarity of i to all other objects of A .

$b(i, j)$ is the average of dissimilarity of i to all objects of $k(k \neq A)$.

The larger the value of $S(i)$ is, the point of i is better classified. In this work, each utterance of each speaker is regarded as a point. Each speaker is regarded as a class and the number of the speakers if equal to the number of the class. $S(i)$ evaluates the degree of which each utterance belongs to its reference speaker.

3.2. Confirm Threshold Setting Approach Using Silhouettes

In this paper, we propose a new method using silhouettes to adjust speaker confirm threshold. The main idea of this method is to remove the utterance which has low silhouette value when estimating the threshold with true scores and false scores.

In the real application, the system sometimes accept the impostors when the confirm threshold is not suitable. It may be caused by the bad utterances of the register which leads to a low confirm threshold. So, we combine the new method using silhouettes with the traditional threshold setting

method to develop a new algorithm for this situation. Here, we call the new method as Silh_TH method. The specific steps are shown as follows:

(1) For the utterance i of the enrollment speaker k , the impostor score is $s(s_1, \dots, s_n)$, calculate the mean of the impostor score:

$$\bar{\mu} = \frac{1}{n} \sum_{i=1}^n s_i$$

(2) Calculate the silhouette value $S(i)$ with the equation (4).

(3) Get new true scores which are nearest to the true speaker score by removing the true scores where $S(i) < s_thd$. And estimate the mean of the new true scores, μ_{silh}

(4) Estimate the speaker's verification threshold using the following formulas:

$$silh_thd = \gamma \mu_{silh} + (1 - \gamma) \bar{\mu}$$

$$confirm_thd = (1 - \alpha) silh_thd + \alpha \mu_{silh}$$

Where α is a confirm parameter which is empirically determined.

Compared with the traditional method, we find that the new method using the algorithm above can remove the low true scores which are near to false scores. So that the confirm threshold is improved.

Table I shows the effect of discriminant change on silhouette threshold of speaker verification decision-making setting. It is apparent that when using silhouettes to adjust the confirm threshold, the Precision improves and FA reduces with the increase of silhouettes threshold. This results indicates that the proposed Silh_TH method gets a better performance during decision-making than traditional method.

3.3. Group Based Decision-Making Mechanism

In this work, we propose a group-based mechanism for speaker verification. Before speaker verification, we first find the similar speaker group for each utterance. And then we make a decision according to the value of the silhouettes. The difference between the new method and the traditional method is shown in Fig. 2. The top half of the figure is the traditional decision-making method and the second half is the proposed approach.

In the proposed method, firstly we compute s for each utterance i corresponding to each register j using the following formula:

Table 1. Comparison of Speaker Verification Performance

s_thd	Precision	FR	FA	HTER
Traditional Method	82.85%	11.72%	16.26%	13.99%
Silh_TH with s_thd=0	84.28%	13.10%	14.42%	13.76%
Silh_TH with s_thd=0.1	87.32%	14.14%	12.58%	13.36%
Silh_TH with s_thd=0.2	88.81%	14.48%	11.04%	12.76%

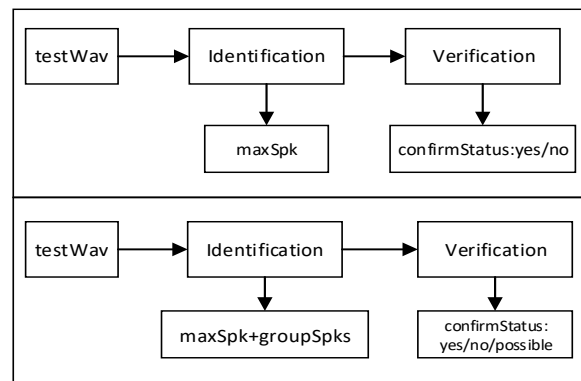


Figure 2. Difference of the process between the proposed mechanism and the traditional method.

$$s(i, j) = \frac{b(i, j) - a(i)}{\max(a(i), b(i, j))}$$

where $a(i)$ is the within distance of utterance i and $b(i, j)$ is the distance between utterance i and speaker j . $s(i, j)$ evaluates the similarity of utterance i and speaker j .

Secondly, we set a threshold $sthd$, and find all the speakers for each utterance where $s(i, j) < sthd$. Then we can get the speaker group for every register by putting all the similar speakers who belong to one register together. The $sthd$ is estimated by equation (1). Thus, each test utterance can be recognized in a group G which includes a max speaker and a similar speaker group for this max speaker.

Thirdly, compute the silhouette values $S_G(k)$ under the assumption of utterance i belongs to speaker $k, k \in G$ using the equation (4).

Finally, we make a decision on the base of the value of $S_G(k)$.

An example of the new mechanism can improve the performance is shown in table II. The traditional method scores in Probability Linear Discriminant Analysis (PLDA) mode with tnorm and logistic regression (LR). It can be seen that the system accepts the false speaker. But in the group based mechanism, the system rejects the false speaker because of the low silhouette value. It reduces the FA rate. And it can also accept the true speaker in the case of low PLDA score but high silhouette value which reduce the FR rate. The overall experiment results is shown in table III. The proposed group based mechanism gets a better performance compared to traditional method.

4. Experiment Results

The speaker recognition task is conducted on the TIMIT corpus to evaluate the performance of the proposed approach. The TIMIT corpus contains a total of 6300 utterances spoken

Table 2. An Example of a Test Segment Results

TestSeg: faem0_female/sa2 maxspk: fcke0_female		
Similar speakers	PLDAScore_tnorm_LR	Silhouette value
fcke0_female	0.987	0.169535
faem0_female	0.977	-0.169535
fdkn0_female	0.381	-0.618905
fawf0_female	0.0	-0.810142
Original confirm status: yes		New confirm status : no

Table 3. Comparison of Speaker Verification performance

Method	Precision	FR	FA	Correct rate
PLDA_tnorm_LR method	82.8%	11.7%	16.3%	96.67%
Group based method	83.3%	11.6%	16.1%	97.67%

by 630 American English speakers. 400 speaker are selected from TIMIT corpus to train the background model. 150 speakers are used to register and test, and 8 utterances of each speaker are used to register and 2 utterances are used to test. The remaining 80 speakers are tested as the outset data.

In the front end of the feature extraction, audio signal is segmented into frames to extract the Mel-Frequency Cepstral Coefficients (MFCCs) with 20 dimensions. The training process includes training a GMM models with 128 Gaussian components and training JFA space matrices. The rank of the eigenvoice matrix is 150 and the dimension of the y-factor is 150 as well. The GMM model is the universal background model (UBM) that will be applied in space matrices training. The verification score is computed using PLDA method. T-Norm score normalization with cohort selection method and logistic regression method are used here, and the cohort number of each speaker if set to 40.

The baseline system is speaker recognition system with eigenvoice modeling with sparse training data [10]. The biggest difference compared with proposed system is the verification process. In our proposed structure, we apply the silhouettes in speaker verification confirm threshold setting. Besides, we propose a group based decision-making method which can make a better decision in speaker verification.

The comparison evaluation results are shown in table I and table III. It is apparent that the new confirm threshold setting method can get a better performance. With the silhouette threshold increasing, the precision increases and the FA rate as well as HTER drop. The group based decision-making method not only finds the similar speaker group for each register, but also reduces the FR rate and FA rate. And meantime, it increases the precision and the correct rate of the overall experiment.

5. Conclusion

The common threshold setting methods are statistics which use the mean and deviation of the score distribution under the assumption of Gaussian distribution. In this paper, we proposed a new verification threshold setting method using of silhouettes and a group based mechanism for speaker verification. The proposed approach takes the good use of silhouette which evaluates the similarity of the point to points in its own cluster compared to points in other clusters. Experiment results on TIMIT corpus show significant improvement compared to the baseline system.

In the future work, we will continue exploit other the threshold setting methods and decision-making mechanisms.

References

- [1] Reynolds, Douglas A. "Speaker identification and verification using Gaussian mixture speaker models ☆." *Speech Communication* 17.1–2 (1995):91-108.
- [2] Chen, Ke. "Towards better making a decision in speaker verification." *Pattern Recognition* 36.2 (2003):329-346.
- [3] Saeta, Javier R., and J. Hernando. "New Speaker-Dependent Threshold Estimation Method in Speaker Verification based on Weighting Scores." 2005:34--41.
- [4] Diaz-Amador, Roberto, E. Castillo-Guerra, and J. Cardenas-Barrera. "Adaptive management of speaker verification thresholds." *Electrical and Computer Engineering, 2008. CCECE 2008. Canadian Conference on IEEE*, 2008:001763-001766.
- [5] Dehak, Najim, et al. "Front-End Factor Analysis for Speaker Verification." *IEEE Transactions on Audio Speech & Language Processing* 19.4 (2011):788-798.
- [6] Bousquet, Pierre Michel, D. Matrouf, and J. F. Bonastre. "Intersession Compensation and

- Scoring Methods in the i-vectors Space for Speaker Recognition." *INTERSPEECH 2011, Conference of the International Speech Communication Association, Florence, Italy, August* DBLP, 2011:485-488.
- [7] Garcia-Romero, Daniel, and C. Y. Espy-Wilson. "Analysis of i-vector Length Normalization in Speaker Recognition Systems." *INTERSPEECH 2011, Conference of the International Speech Communication Association, Florence, Italy, August* DBLP, 2011:249-252.
- [8] Higgins, A., L. Bahler, and J. Porter. "Speaker verification using randomized phrase prompting." *Digital Signal Processing* 1.2 (1991):89-106.
- [9] Carey, M. J., E. S. Parris, and J. S. Bridle. "A speaker verification system using alpha-nets." *International Conference on Acoustics, Speech, and Signal Processing* IEEE, 2002:397-400.
- [10] Kenny, P., G. Boulianne, and P. Dumouchel. "Eigenvoice modeling with sparse training data." *IEEE Transactions on Speech & Audio Processing* 13.3 (2005):345-354.
- [11] Glembek, Ondrej, et al. "Comparison of scoring methods used in speaker recognition with Joint Factor Analysis." *IEEE International Conference on Acoustics, Speech and Signal Processing* IEEE, 2009:4057-4060.
- [12] Lindberg, J., et al. "Techniques For A Priori Decision Threshold Estimation In Speaker Verification." 1920:89--92.