SPEAKER VERIFICATION WITHOUT BACKGROUND SPEAKER MODELS

Chun-Nan Hsu, Hau-Chung Yu, Bo-Hou Yang

Institute of Information Science, Academia, Sinica, Nankang 115, Taipei City, Taiwan
chunnan@iis.sinica.edu.tw

ABSTRACT

Speaker verification concerns the problem of verifying whether a given utterance has been pronounced by a claimed authorized speaker. This problem is important because an accurate speaker verification system can be applied to many security applications. In this paper, we present a new algorithm for speaker verification called OSCILLO. By applying tolerance interval analysis in statistics, OSCILLO can verify a speaker's ID without *background speaker models*. This greatly reduces the space requirement of the system and the time for both training and verification. Experimental results show that OSCILLO can achieve error rates comparable or better than the GMM-based system with background speaker models for three benchmark databases: TCC-300, TIMIT and NIST 2000.

1. INTRODUCTION

Speaker recognition is the process of automatically recognizing who is speaking on the basis of information obtained from speech waves. This technique will make it possible to verify the identity of the person accessing the system, that is, implementing access control by voice in various applications.

There are two categories of the problems in speaker recognition [7] --- speaker identification and speaker verification. Speaker identification deals with a close-set classification problem, where the systems are supposed to assign the unknown speaker's identity to one of a set of known speakers, while speaker verification systems must accurately reject an unknown speaker, and therefore, must deal with an open-set classification problem [4].

Previously, Reynolds et al. proposed a speaker verification system using Gaussian mixture models (GMM) [6][8]. Their systems require a set of *background speaker models*, which are constructed from a large speech database of speakers with a variety of demographic backgrounds that match the population of the potential imposter speakers. In many real world situations, however, it may not be feasible to obtain such a database. Moreover, there is no systematic approach to the construction of background speaker models. The proposed approaches in [6] and [8] are heuristic and depend heavily on the quality of the speech database. The need of background speaker models also makes many security applications of speaker verification infeasible. For example, it is highly desirable on the market if portable devices such as cellular phones, and PDAs can be equipped with a speaker verification system so that these devices can only be accessed by their owner. But it is not cost effective if the device must carry a large database for constructing background speaker models. The other option is to let the users submit their speech sample to some site where training will be performed, but that will deter users with privacy concerns.

In this paper, we present a new algorithm for speaker verification without background speaker models. With this new algorithm, a speaker verification system can be trained using only the speech samples from the claimed speaker. This greatly reduces the space requirement and the time for both training and verification, and thus, makes many security applications feasible.

This paper is organized as follows: Section 2 reviews the Gaussian mixture model for speaker recognition; Section 3 introduces tolerance interval analysis; Section 4 describes the baseline system of speaker verification, a GMM-based algorithm proposed in previous work by Reynolds et al. [6]; Section 5 presents our algorithm OSCILLO; Section 6 reports the experimental results that compare OSCILLO and the baseline system; the last section contains our conclusion.

2. GUASSIAN MIXTURE MODEL

Mixture models are a type of the density model that comprises of a number of component functions, usually Gaussian. These component functions are combined to provide a multimodal density. The distributions of feature vectors were extracted from a speaker's speech modeled by a Gaussian mixture density. This is a method that has been proved to be one of the most successful approaches to close-set, text-independent speaker identification [5].

A Gaussian mixture density is a weighted sum of M component densities, and is given by the equation

$$p(\vec{x} \mid l) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \qquad (1)$$

where $\vec{x}$ is a D-dimensional random vector, $b_i(\vec{x})$, $i=1,\ldots,M$, are the component densities and p_i, $i=1,\ldots,M$, are the mixture weights. Each component density is a D-variate Gaussiam function of the form:

$$b_i(x) = \frac{1}{(2p)^{D/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(\vec{x}-\vec{m}_i)' \Sigma_i^{-1}(\vec{x}-\vec{m})_i\right\} \quad (2)$$

with mean vector $\vec{m}_i$, and covariance matrix Σ_i. The mixture weights satisfy the constraint that $\sum_{i=1}^{M} p_i = 1$.

The complete Gaussian mixture density is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities. These parameters are collectively represented by the notation:

$$l = \{p_i, \vec{m}_i, \Sigma_i\}, \; i=1,\ldots,M.$$

For speaker identification, each speaker is represented by a GMM and is referred to by his/her model l.

3. TOLERANCE INTERVAL ANALYSIS

For any fixed region *R* of a given population, we define the coverage of *R* as the proportion of the population that lies in *R*, that is, the proportion of the population covered by *R*. Formally, the coverage of *R* is

$$C(R) = P(X \in R),$$

where *X* is drawn at random from the population.

A tolerance region is a random region having a specified probability, $1-a$, such that its coverage is at least a specified value, *c*. Various names are given to $1-a$ and *c* in the literature. We shall call $1-a$ the *confidence level* and *c* the *tolerance proportion*, the latter because in some situation it is the minimum proportion of the population that is considered tolerable to cover. We also consider a "*c tolerance region with confidence* $1-a$*.*" For any continuously distributed random variable, we can show that the probability *P* that the coverage *C* spanned by the largest to smallest of *n* independent samples is at least *c* is [2][4]:

$$P(C > c) = \sum_{k=0}^{n-2} \binom{n}{k} c^k (1-c)^{n-k} \quad (3)$$

In Figure 1, we can see that as the training sample size increases, the coverage increases when the confidence is fixed.

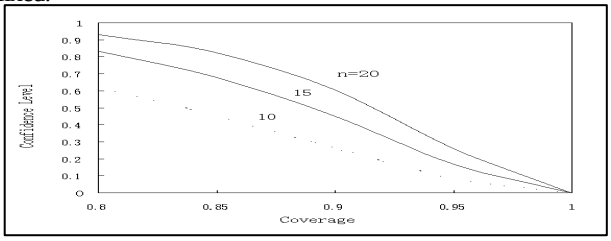

Figure 1. Tolerance interval analysis.

4. BASELINE SYSTEM

The general approach proposed by Reynolds et al. [6] for the speaker verification system is to apply a likelihood ratio test to an input utterance to determine if the claimed speaker should be accepted or rejected. Given an utterance $X = \{x_1,\ldots,x_T\}$, a claimed speaker identity with corresponding model l_C and anti-model $l_{\overline{C}}$, the likelihood ratio is defined by

$$\frac{\Pr(\text{X is from the claimed speaker})}{\Pr(\text{X is not from the claimed speaker})} = \frac{P_r(l_C \mid X)}{P_r(l_{\overline{C}} \mid X)} = \frac{P_r(X \mid l_C)/P_r(X)}{P_r(X \mid l_{\overline{C}})/P_r(X)}. \quad (4)$$

Discarding the constant prior probabilities for claimant and imposter speakers, the likelihood ratio in the log domain becomes

$$\Lambda(X) = \log p(X \mid l_C) - \log p(X \mid l_{\overline{C}}) \quad (5)$$

The term $p(X \mid l_C)$ is the likelihood of the utterance given that it is from the claimed speaker and $p(X \mid l_{\overline{C}})$ is the likelihood given that it is not from the claimed speaker. The likelihood ratio is compared to a threshold Θ and the test speaker is accepted if $\Lambda(X) > \Theta$ and rejected if $\Lambda(X) \le \Theta$. The likelihood ratio essentially measures how much better the claimant's model scores for the test utterance compared to some non-claimant model. The decision threshold is then set to adjust the trade-off between rejecting true claimant utterances (false reject errors) and accepting non-claimant utterance (false accept errors).

Reynolds et al. proposed an approach to the selection of the background speaks [6]. In this approach, GMM of all speakers in the database are created using training data and pair-wise distances between the speaker models are computed. The background speakers selected as the maximally-spread close set and far set. We use the approach as our baseline system in our experiment.

For training data $X = \{X_1,\ldots, X_n\}$, the threshold value is obtained by the following steps:

1. Calculate $\Lambda(X_i)$ by equation (5), where $1 \le i \le n$.
2. Place $\Lambda(X_i)$ in one sorted list and set the point on the list at which the false accept rate equals to the false reject rate as the threshold value.

Applying likelihood ratio test for speaker verification has been presented in [1]. More recently, Reynolds et al. has proposed an improved yet similar method for constructing background speaker models [8].

5. OSCILLO SYSTEM

Given an utterance sample from a claimed speaker *A*, assuming that GMM is reliable, then it is likely that the output score of a GMM model built from speaker *A*'s utterance sample is grater than the score from a model built from another speaker's utterance sample. Figure 2 illustrates this idea.

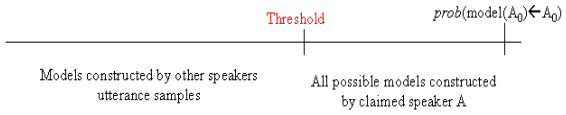

Figure 2. Preliminary of the OSCILLO system

The key assumption of this idea is that for the specific sample A_0, it is likely that:

$$\forall i, P_r(A_0 \mid \text{mod}(A_i)) > P_r(A_0 \mid \text{mod}(B)),$$

where mod(*B*) is the model constructed by another person's utterance sample while $\text{mod}(A_i)$ is the model constructed by the claimed speaker's utterance sample. A_i is an utterance sample of claimed speaker *A*, $0 \le i \le n$. If we can find the lower bound of $\Pr(A_0 \mid \text{mod}(A_i))$, then we can take this value as a threshold to determine if a given utterance has been pronounced by the claimed speaker.

We use the tolerance interval analysis [2] described in Section 4 to estimate the number of samples required and find the lower bound of $\Pr(A_0 \mid \text{mod}(A_i))$. Consider the size of the speech databases in our experiment, we set 20 as the number of independent samples to obtain a tolerance region with the coverage at least 0.8 and confidence level 0.93 (see Figure 1). Each Gaussian mixture model is trained by 1000 feature vectors extracted from 10 seconds speech. Therefore, each speaker needs to provide approximately 3.5 minutes of speech to construct 20 GMM models. We denote the 20 samples to train the GMM models as $\{A_0, ..., A_{19}\}$. Let A_0 be the specific utterance sample[1]. The OSCILLO algorithm of speaker verification is given below (see also Figure 3).

1. For the input test utterance *B*, we first construct a Gaussian mixture model mod(*B*).
2. Calculate the posterior probability of mod(*B*) for the specific utterance sample A_0.
3. Verify that the input utterance is from the claimed speaker if the posterior probability is greater than the threshold and not from the claimed speaker otherwise.

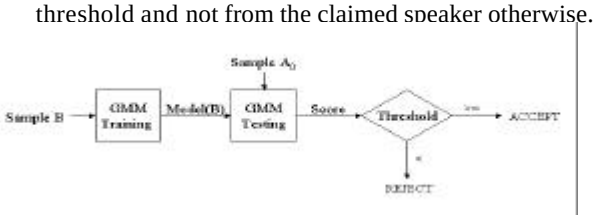

Figure 3. Speaker Verification algorithm OSCILLO

6. SPEAKER VERIFICATION EXPERIMENTS

6.1 Database

Three databases used for the experiments are given in Table 1. They are TCC-300, TIMIT and NIST 2000 [3]. TCC-300 is a collection of microphone speech databases produced by three universities in Taiwan: National Taiwan University, National Cheng-Kung University, and National Chiao-Tung University, from 300 speakers (150 males, 150 females).

Database	# speakers	# speaker used	channel	threshold
TCC-300	300	100	microphone	25
TIMIT	630	168	microphone	12
NIST	1060	50	telephone	15

Table 1. Databases for the experiments

A 30-ms Hamming windows was applied to the-speech every 10ms, allowing us to obtain 100 feature vectors from 1-s of speech data. For each speech frame, both a twelfth-order MFCCs (mel-frequency cepstral coefficients) and a log energy analysis were performed. We filter the feature vectors whose log energy values were lower than a threshold (see the fifth column of Table 1) as silence-removing processing.

From tolerance interval analysis, we segmented all the vectors into 20 samples . Each sample contains 1000 feature vectors. But for TIMIT and NIST, the speech data per speaker is not sufficient for 20 segments. Therefore, we have to create more samples by randomly selecting start points for the vectors. The Gaussian mixture model is constructed in diagonal covariance matrices and 32 component densities.

In the experiments, we found that the selection of the specific utterance sample may affect the performance of the system for telephone speech database, because the samples are so noisy that they yield inaccurate GMM models. We applied the following method to select the

[1] How to select this sample will be discussed in Section 6.

most suitable sample as the specific utterance sample. For each sample, the deviation is obtained by the equation below,

$$s = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - n m^2}{n}}, \quad m = \frac{1}{n}\sum_{i=1}^{n} x_i$$

where x_i is the score of the given sample from the models constructed by the other 19 samples. The sample that minimizes the deviation s is selected. Meanwhile, the speech data in microphone speech databases are quite clean and the performance will not be affected no matter which sample is selected.

6.2 Experimental Result

We performed experiments to compare OSCILLO and the baseline system described in Section 4. Since the baseline system requires samples to construct background speaker models, while OSCILLO does not but needs sufficiently large samples for each claimed speaker, it is difficult to compare the two systems based on exactly the same training data. However, we have tried our best to minimize the difference given the available databases. Table 3 show the samples used for baseline system (top) and OSCILLO (bottom), respectively.

Table 4 shows the experimental results in error rates. The results clearly reveal that OSCILLO outperformed the baseline system for TCC-300. By applying the sample selection method, we improved the error rates of OSCILLO from 7% to 4.8 % in FR and from 22% to 0.9% in FA. Note that the figures for TIMIT by the baseline system are cited from [6]. In general, the results show that OSCILLO achieves error rates comparable or better than the baseline system that requires background speaker models.

We also compare the time spent for training and verification of the two systems for TCC-300 (see Table 5). The baseline system takes much time to construct the background speaker models. As a result, OSCILLO can be trained faster than the baseline system by about 54 times. OSCILLO also takes much less time for verification.

7. CONCLUSION

In this paper, we present OSCILLO, which is based on tolerance interval analysis for speaker verification. This algorithm provides a preliminary step toward a speaker verification system without any background speaker model.

Our future work includes applying different techniques to optimize speaker models and reduce the channel effect to improve the performance for telephone speech data.

Database	# speaker	# true tests per speaker	# imposter tests per speaker	Total # true tests	Total # imposter tests
TCC-300	100	3	297	300	29700
TIMIT	168	2	334	336	55778
NIST	50	2	98	100	49000
TCC-300	100	5	495	500	49500
TIMIT	168	10	167	1680	28056
NIST	50	10	441	500	22050

Table 3. Number of trials for the baseline system (top) and number of samples for OSCILLO(bottom).

	FR (%)		FA (%)	
	Baseline	OSCILLO	Baseline	OSCILLO
TCC-300	16.6	5.6	1.2	0.1
TIMIT	0.24*	5	0.24*	0.3
NIST	24	4.8	6.5	0.9

Table 4 Comparison in FR (false reject rate) and FA (false accept rate)

Database	Training Time		Verification Time	
TCC-300	Baseline	OSCILLO	Baseline	OSCILLO
	53.7 minutes	59.65 seconds	20 seconds	3.665 seconds

Table 5. Comparison of training and verification time

REFERENCES

[1] L. Bahler Higgins and J. Porter." Speaker verification using randomized phrase prompting." *Digital Signal Processing*, 1 , 1991.

[2] John W.Pratt and Jean D. Gibbsons. *Concepts of Nonparametric Theory*, Springer Series In Statistics, Spring Verlag, 1981.

[3] Mark Przybocki and Alvin Martin. "2000 NIST speaker recognition evaluation." *Speech Communications*, 31, 2000.

[4] Edward C. Real and Andrew H. Baumann. "Open set classification using tolerance intervals." Presented at *the 34th Asilomar Conference on Signals, Systems and Computers*, October 2000, Monterey, CA, USA.

[5] Douglas A. Reynolds. "Robust text-independent speaker identification using gaussian mixture speaker models ." *IEEE Transactions on speech and audio processing*, 3(1), Jan. 1995.

[6] Douglas A. Reynolds. "Speaker identification and verification using Gaussiam mixture speaker models." *Speech Communication*, 17, 1995.

[7] Douglas A. Reynolds. "An overview of speaker recognition technology". ICASSP 2002.

[8] Douglas A. Reynolds, Thomas F. Quatieri and Robert B. Dunn. "Speaker verification using adapted Gaussian mixture models." *Digital Signal Processing*, Academic Press, 2000.