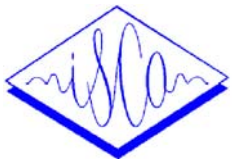
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

A Study on Model-Based Equal Error Rate Estimation for Automatic Speaker Verification

Jyh-Min CHENG and Hsiao-Chuan WANG

Department of Electrical Engineering,
National Tsing Hua University, Hsinchu 300, Taiwan, R.O.C.

d927908@oz.nthu.edu.tw, hcwang@ee.nthu.edu.tw

10.21437/Interspeech.2004-517

Abstract

Usually, we need a large number of testing samples to evaluate the performance of automatic speaker verification (ASV) systems. The equal error rate (EER) is a common measure for this purpose. It is derived according to the threshold determined by finding the verification score when the false rejection rate (FRR) equals to the false acceptance rate (FAR). In this paper, a method of model-based EER estimation for the ASV system is proposed. The goal is to estimate the EER directly from the speaker model parameters without running the speaker verification experiments using a large number of testing samples. The verification scores are computed using the model parameters, and then both FRR and FAR are derived. With a small number of testing samples, we can adjust the score distribution to estimate the EER of the ASV system. The experimental result shows that the proposed method is effective and very promising for feedback loop design of an ASV system.

1. Introduction

The automatic speaker verification (ASV) refers to the task of verifying the speaker's identity from their voices. In the conventional ASV system, the decision rule of acceptance or rejection is based on the score of a testing utterance and a predefined threshold [1]. There are two phases for an ASV system to accomplish this task. In the training phase, it learns each client's voice features from the training utterances to generate a statistical model. Also, a single speaker independent model called the world model or the universal background model (UBM) [2] is generated. In the testing phase, the ASV system analyzes an incoming utterance, and then uses the claimed speaker model and the UBM to compute the verification score. Then the score is compared with a preset threshold to determine whether the speaker is accepted or rejected. The Gaussian mixture model (GMM) method has been successfully used for this purpose [1] [3].

The speech data of a specific speaker is trained with EM method [4] to derive the corresponding GMM. The verification score is based on computing the log likelihood ratio (LLR) of the client speaker model and the UBM. The score distributions are then used to compute the false rejection rate (FRR) and the false acceptance rate (FAR). The equal error rate (EER), where FRR equals to FAR, is a common measure of the performance of the ASV system. Usually, we need a large number of testing samples to evaluate the performance of the ASV system, i.e., to compute the EER.

A model-based framework of classification error rate estimation has been proposed for automatic speech recognition [5]. It aims at predicting the run-time performance of a hidden Markov model (HMM) based recognition system for a given task vocabulary and grammar without the need of running recognition experiments using a separate set of testing samples. This motivates the idea of evaluating the performance of an ASV system without running the speaker verification experiments.

In this paper, we derive an algorithm for estimating the EER of an ASV system directly by the parameters of the corresponding GMMs. It computes LLR scores of the true speaker and the imposters, and then calculates the FRR and the FAR. We used an approximation in [5] to compute mixture densities in the log-of-sum of the integration of LLR score. Finally, we use a small number of testing samples to adjust the distribution of LLR scores. Then we derive an estimation of the EER which is close to the EER obtained from running speaker verification experiment with large number of testing samples.

In the following sections, we will first introduce the computation of the LLR scores directly by the model parameters. Then, we show the observations of the distribution of model-based LLR scores based on NIST 1999 Speaker Evaluation data. Finally, we demonstrate an experiment of using a small number of testing samples to adjust the distribution of LLR scores.

2. Computation of Model-Based LLR Scores

This section introduces the algorithm for estimating the LLR score in terms of the model parameters without using a large number of testing samples.

2.1. Definitions

A verification score is defined as the following equation related to the test speaker model, the client speaker model, and the UBM,

$$\Lambda(p;q,r) = \int_{x} p(\mathbf{x}) \log \frac{q(\mathbf{x})}{r(\mathbf{x})} d\mathbf{x} \qquad (1)$$

where $\Lambda(\cdot)$ is the score with respect to test speaker model p, client speaker model q, and the UBM r. $\mathbf{x}$ is the incoming feature vector. Equation (1) can also be expressed as,

$$\Lambda(p;q,r) \quad (2)$$
$$= \int_x p(x)\log q(x)dx - \int_x p(x)\log r(x)dx$$
$$= L(p;q) - L(p;r)$$

where $L(\cdot)$ is the log likelihood score. Then $\Lambda(\cdot)$ is called the log likelihood ratio (LLR) score.

2.2. Derivation of the LLR Scores

The log likelihood score $L(\cdot)$ in the mixture mode is computed by,

$$L(p;q) = \sum_{i=1}^{M} w_{pi} \int_x N(x;\mu_{pi},\Sigma_{pi}) \cdot \log\left(\sum_{j=1}^{M} w_{qj} N(x;\mu_{qj},\Sigma_{qj})\right)dx \quad (3)$$

where $N(\cdot)$ is the Gaussian density and M is the corresponding number of mixtures. w, μ, and Σ are the mixture weight, mean, and covariance, respectively.

In order to solve the difficulty in computing the log-of-sum, we make an approximation,

$$L(p;q) = \sum_{i=1}^{M} w_{pi} \log\left(\sum_{j=1}^{M} w_{qj} \cdot \exp\left(\int_x N(x;\mu_{pi},\Sigma_{pi}) \cdot \log\left(N(x;\mu_{qj},\Sigma_{qj})\right)dx\right)\right) \quad (4)$$

The last term in the exponential of equation (4) can be then computed as the expected value of $\log(N(x;\mu_{qj},\Sigma_{qj}))$ given the distribution of $N(x;\mu_{pi},\Sigma_{pi})$,

$$E_{pi}\left[\log\left(N(x;\mu_{qj},\Sigma_{qj})\right)\right] \quad (5)$$
$$= -\frac{D}{2}\log(2\pi) - \frac{1}{2}\log\left(|\Sigma_{qj}|\right)$$
$$-\frac{1}{2}E_{pi}\left[(x-\mu_{qj})^T \Sigma_{qj}^{-1}(x-\mu_{qj})\right]$$

where D is the vector dimension, and the last term in equation (5) can be simplified by,

$$E_{pi}\left[(x-\mu_{qj})^T \Sigma_{qj}^{-1}(x-\mu_{qj})\right] \quad (6)$$
$$= E_{pi}\left[(x-\mu_{pi})^T \Sigma_{qj}^{-1}(x-\mu_{pi})\right]$$
$$+(\mu_{pi}-\mu_{qj})^T \Sigma_{qj}^{-1}(\mu_{pi}-\mu_{qj})$$
$$= tr(\Sigma_{pi}\Sigma_{qj}^{-1}) + (\mu_{pi}-\mu_{qj})^T \Sigma_{qj}^{-1}(\mu_{pi}-\mu_{qj})$$

where $tr(\cdot)$ is the trace of matrix $(\cdot)$. Then, the log likelihood score can be computed by the following equation,

$$L(p;q) = \sum_{i=1}^{M} w_{pi} \log\left(\sum_{j=1}^{M} w_{qj} \frac{1}{(2\pi)^{D/2}|\Sigma_{qj}|^{1/2}} \cdot \exp\left(-\frac{1}{2}\left(tr(\Sigma_{pi}\Sigma_{qj}^{-1}) + (\mu_{pi}-\mu_{qj})^T \Sigma_{qj}^{-1}(\mu_{pi}-\mu_{qj})\right)\right)\right) \quad (7)$$

By Equations (2) and (7), we could compute the LLR score directly using the model parameters without running the verification experiments.

3. Model-Based EER Estimation for ASV.

The EER estimation of an ASV system is based on the computation of LLR scores. Usually a large number of testing samples is necessary for evaluating FRR and FAR so that the EER can be determined. Here we try to evaluate EER by using model parameters

3.1. FRR vs. FAR

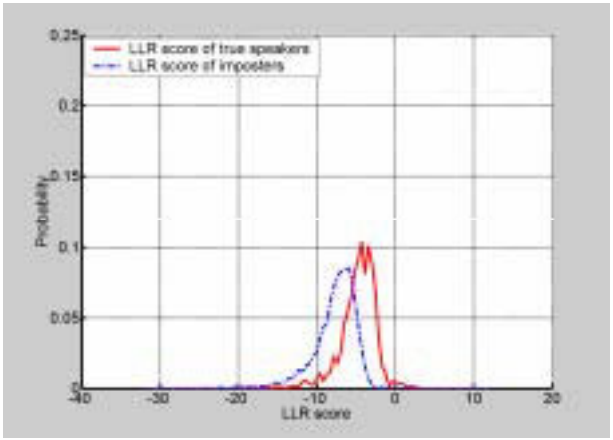
In an ASV system, there are two types of error, i.e., FRR and FAR. In equation (1), q and r are the claimed speaker model and the UBM, respectively. For the FRR, the test speaker model p is very similar to the claimed speaker model, q,

$$\Lambda(p;q,r) = \int_x q(x)\log\frac{q(x)}{r(x)}dx \quad (8)$$

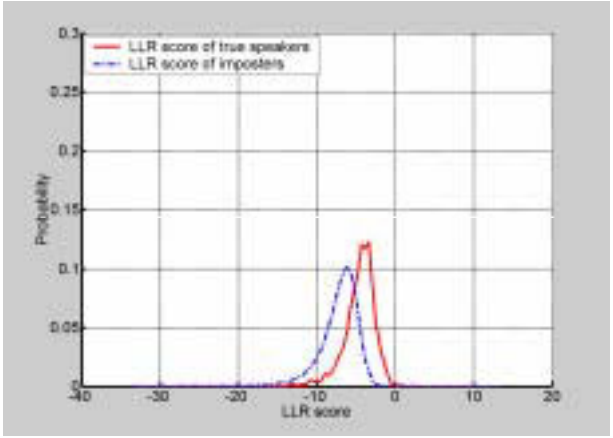
For the FAR, the test speaker model p is the imposter mixture densities, we approximate it as the UBM, r,

$$\Lambda(p;q,r) = \int_x r(x)\log\frac{q(x)}{r(x)}dx \quad (9)$$

We can observe the distributions of the LLR scores either derived from running verification experiments using a large number of testing samples in Figure 1 or derived from model parameters in Figure 2. It is clear that there is no overlapping of two distributions in Figure 2. That means EER is zero, or can not be calculated.

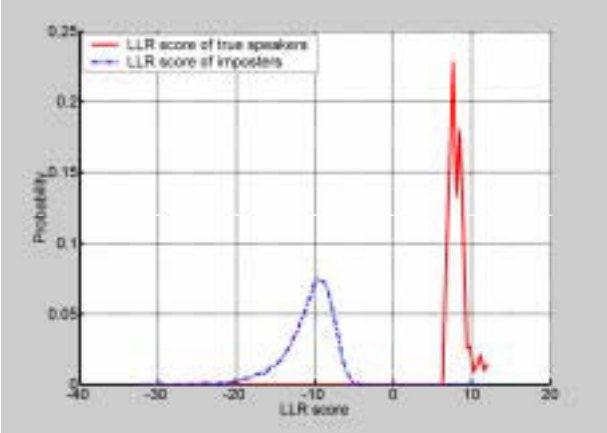


(a) male speakers

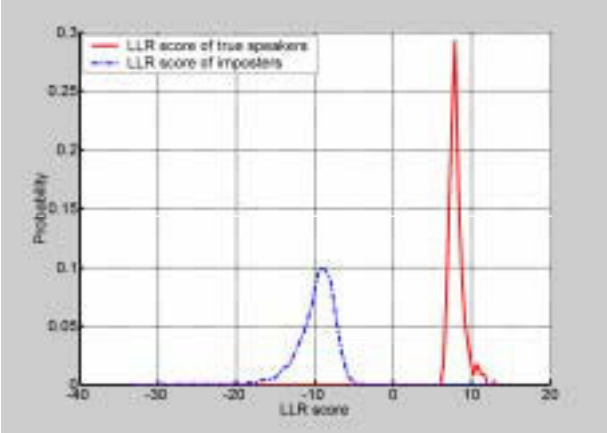


(b) female speakers

Figure 1: Distributions of data-based LLR scores



(a) male speakers



(b) female speakers

Figure 2: Distributions of mode-based LLR scores

3.2. Adjustment of LLR score distributions

In Figure 2, we can see that both the mean of true speakers' LLR scores and the mean of imposters' LLR scores are shifted related to those in Figure 1. That means there are biases of LLR scores derived from model parameters as comparing with the LLR scores obtained from running verification experiments. For the scores of true speakers, the score distribution is even distorted. Therefore, we need a small number of testing samples to adjust the distributions of model-based LLR scores so that they can be close to those derived from running verification experiments. Here we suggest to adjust the mean of model-based LLR score distribution or to adjust both mean and variance of model-based LLR score distribution.

3.3. Experiments

3.3.1. Database

In our experiments, we use the evaluation data of the NIST 1999 Speaker Verification task. It provides 230 male speakers and 309 female speakers. Each speaker has two minutes training data with two different channels recorded in 8 kHz sampling rate and ITUT G.711 μ-law encoded format. In the testing set, there are 1447 and 1972 utterances for male speakers and female speakers, respectively.

3.3.2. Experimental arrangement

The speech data is pre-emphasized with a factor of 0.97. The analysis window is of 32 ms length. The frame shift is 16 ms. The 15-order Mel-frequency cepstral coefficients (MFCCs) extracted from each frame form a feature vector for generating speaker models. Speaker models are represented in GMM of 1024 mixture components. The UBM is generated using data of 60 speakers for each sex in NIST 1999 Training data. The small set of testing samples for score distribution adjustment is about 10% of whole testing samples. This small set of testing samples is randomly selected from the whole test samples.

3.3.3. Experimental results

The experiments are arranged to investigate the goodness of the model-based EER estimation. This can be accomplished by measuring the difference between the model-based EER estimation and the EER obtained by running the verification experiment using a large number of testing samples, i.e., the whole testing samples. The EER measured by running the verification experiment using whole testing samples is considered as a result of reference experiment. For the performance evaluation, EERs are computed by using Equations (2) and (7) with the adjustment of model-based LLR score distributions. The adjustment is based on a small number of testing samples randomly selected from the whole testing samples. The adjustment can be on mean only or on both mean and variance. The experiments on male speakers and female speakers are conducted separately. For each experiment, we consider two conditions. One assumes that the distribution of speakers' testing data is known so that we can proportionally select the testing data for each speaker from the whole testing samples. The other assumes no information about the distribution of speakers' testing data so that the testing data are randomly selected without considering the distribution of speakers. The speech features are presented in two cases. One is MFCCs without mean subtraction (denoted by F-1), and the other is MFCCs with mean subtraction (denoted by F-2).

The results of reference experiment (denoted by Exp-1) are shown in Table 1. They are far away from that implied by Figure 2. In other words, the estimation of EERs using model parameters without making any adjustment is impossible. The reason comes from that the approximation based on the assumption deriving from Equation (3) to Equation (4) is an inappropriate one.

Table 1: EERs of reference experiment (%)

	F-1		F-2		Ave.
	male	female	male	female	
Exp-1	24.04	24.18	22.24	23.85	23.58

The experimental results with the adjustment of model-based LLR score distribution on mean only (denoted by Exp-2) or on both mean and variance (denoted by Exp-3) are

summarized in Table 2 and Table 3. The goodness of the model-based EER estimation can be measured by its difference from the EER obtained from reference experiment.

Table 2: EER with knowing the distribution of speakers' data (%)

	F-1		F-2		Ave.
	male	female	male	female	
Exp-2	23.87	18.77	25.38	22.74	22.69
	(0.17)	(5.41)	(3.14)	(1.11)	(2.46)
Exp-3	29.71	17.36	27.43	27.49	25.50
	(5.67)	(6.82)	(5.19)	(3.64)	(5.33)

*( ) denote the deviation from EER of reference experiment in Table 1

Table 3: EER without knowing the distribution of speakers' data (%)

	F-1		F-2		Ave.
	male	female	male	female	
Exp-2	29.57	22.17	32.83	25.57	27.54
	(5.53)	(2.01)	(10.59)	(1.72)	(4.96)
Exp-3	30.65	28.16	27.83	27.83	28.62
	(6.61)	(3.98)	(5.59)	(3.98)	(5.04)

*( ) denote the deviation from EER of reference experiment in Table 1.

In average, the model-based estimation of EER with knowing the distribution of speakers' data is more close to the result of Exp-1 than that without knowing the distribution of speakers' data. The adjustment on the score distribution mean only is superior to the adjustment on both mean and variance. The use of MFCCs presented with mean subtraction or without mean subtraction does not make too much different.

4. Conclusions

We propose a preliminary study on the model-based equal error rate (EER) estimation for automatic speaker verification (ASV). Experimental results show that we can effectively estimate the performance of the ASV system based on its model parameters and using a small number of testing samples. The adjustment of model-based LLR score distribution is necessary. The use of MFCCs or mean subtracted MFCCs does not make too much different. This technique can be used for feedback loop design in an ASV system.

5. Acknowledgements

This research was partially sponsored by the National Science Council, Taiwan, under contract number NSC-92-2213-E-007-036.

6. References

[1] Reynolds D., "Speaker identification and verification using Gaussian mixture models," Speech Communication,. Vol. 17, pp. 91-108, 1995.

[2] D. A. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," Proceedings of the EUROSPEECH, Rhodes, Greece, Sept. 1997.

[3] Reynolds D., Rose R.C., "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Transactions on Speech and Audio Processing, Vol. 3, pp. 72-83, Jan. 1995.

[4] Guorong Xuan, Wei Zhang, Peiqi Chai, "EM algorithm of Gaussian mixture and hidden Markov model," IEEE International Conference on Image Processing, Vol. 1, pp. 145-148, 2001.

[5] C.-S. Huang, H.-C. Wang and C.-H. Lee, "A Study on Model-Based Error Rate Estimation for Automatic Speech Recognition," IEEE Transactions on Speech and Audio Processing, Vol. 11, No. 6, Nov. 2003