

A PRIORI THRESHOLD DETERMINATION FOR PHRASE-PROMPTED SPEAKER VERIFICATION

W.D. Zhang, K.K. Yiu, M.W. Mak and C.K. Li

Dept. of Electronic and Information Engineering,
The Hong Kong Polytechnic University, Hong Kong.
enmwamak@polyu.edu.hk

M.X. He

Ocean Remote Sensing Institute,
Ocean University of Qingdao, China.

ABSTRACT

This paper presents a novel method to determine the decision thresholds of speaker verification systems using enrollment data only. In the method, a speaker model is trained to differentiate the voice of the corresponding speaker and that of a general population. This is accomplished by using the speaker's utterances and those of some other speakers (denoted as anti-speakers) as the training set. Then, an operation environment is simulated by presenting the utterances of some pseudo-impostors (none of them is an anti-speaker) to the speaker model. The threshold is adjusted until the chance of falsely accepting a pseudo-impostor falls below an application dependent level. Experimental evaluations based on 138 speakers of the YOHO corpus suggest that with a simulated operation environment, it is able to determine the best compromise between false acceptance and false rejection.

1. INTRODUCTION

The determination of decision thresholds is a very important problem in speaker verification. A small threshold could result in a vulnerable system while a large one could make the system annoying to users. Conventional threshold determination methods [1, 2] typically compute the distribution of inter- and intra-speaker distortions, and then chose a threshold to equalize the overlapping area of the distributions, i.e. to equalize the false acceptance rate (FAR) and false rejection rate (FRR). The success of this approach, however, relies on whether the estimated distributions match the speaker- and impostor-class distributions. Another approach derives the thresholds based on speaker models only [3]. Session-to-session speaker variability, however, contributes much bias to the thresholds, rendering the verification system unusable.

Due to the difficulty in determining a reliable threshold, it is not uncommon for researchers to report the equal error rate (ERR) of verification systems based on the assumption that an *a posteriori* threshold can be optimally adjusted during verification. A real-world application, however, is only realistic with *a priori* thresholds which should be determined during enrollment.

In recent years, research effort has been centered upon the normalization of speaker scores to minimize error rates. This includes the likelihood ratio scoring proposed by Higgins et al. [4], where verification decisions are based on the ratio of the likelihood that the observed speech is uttered by the true speaker to the likelihood that it is spoken by

an impostor. The *a priori* threshold is then set to 1.0, with the claimant being accepted (reject) if the ratio is greater (less) than 1.0. Higgins et al. [4] and more recent work [5, 6] based on likelihood normalization also show that including an impostor model during verification not only improves speaker separability, but also allows decision thresholds to be easily set. Although this approach helps to select an appropriate threshold, it causes the system to favor rejecting true speakers, resulting in a high FRR. For example, in [4], the FRR is more than 10 times larger than the FAR. A recent report [7] using similar normalization methods but different threshold setting procedures also found that the average of FAR and FRR is about 3 to 5 times larger than the ERR, suggesting that the ERR is an over optimistic estimate of the true system performance.

This paper proposes an *a priori* threshold determination method to address the above problem. The method is different from that of [4] in that rather than using a ratio speaker set formed by pooling the nearest reference speakers, we used two speaker sets, namely anti-speaker set and pseudo-impostor set, to determine the threshold. For each speaker, a speaker model is trained to differentiate the speech uttered by the speaker and the anti-speakers. Then, the pseudo-impostor set is used to determine the threshold. To enhance the capability of the speaker models without increasing the enrollment time, we sample the utterances of many anti-speakers and pseudo-impostors to form a training set (having the same size as the one without sampling) for building the speaker models and for determining the thresholds. Therefore, an operation environment for the speaker model is effectively simulated.

This paper is organized as follows. Section 2 outlines the speaker models and the verification procedure. The *a priori* threshold determination methods are explained in Section 3, and their performance are compared in Sections 4 and 5. Finally, we conclude our discussions in Section 6.

2. SPEAKER VERIFICATION

2.1. Speaker Models: EBF Networks

Elliptical basis function (EBF) networks have been used as the speaker models in this work. EBF networks can be considered as an extension of radial basis function networks [8]. The k th output ($k = 1, \dots, K$) of an EBF network with I inputs and J function centers has the form

$$y_k(\vec{x}_p) = w_{k0} + \sum_{j=1}^J w_{kj} \phi_j(\vec{x}_p) \quad p = 1, \dots, N \quad (1)$$

This work was supported by The H.K. Polytechnic University Grant No. G-S472.

where $\phi_j(\vec{x}_p) = \exp\left\{-\frac{1}{2\gamma_j}(\vec{x}_p - \vec{\mu}_j)^T \Sigma_j^{-1}(\vec{x}_p - \vec{\mu}_j)\right\}$. In (1), \vec{x}_p is the p th input vector, $\vec{\mu}_j$ and Σ_j are the mean vector and covariance matrix of the j th basis function respectively, w_{k0} is a bias term, and γ_j is a smoothing parameter controlling the spread of the j th basis function.¹ Typically, the mean vectors are found by the K-means algorithm and the covariance matrices are estimated by sample covariances or the EM algorithm. Once the basis function parameters are known, the output weights can be determined by a least squares approach using the technique of singular value decomposition (see [9] for details).

To apply EBF networks for speaker verification, each registered speaker is assigned an EBF network with two outputs. The first output is trained to output a ‘1’ for the speaker’s speech and a ‘0’ for other speakers’ utterances, and vice versa for the second output.

Of particular interest is that the EBF networks incorporate the idea of likelihood ratio scoring in their discriminative training procedure. An EBF network does not require a set of cohort or background speakers during verification; rather, it embeds the characteristics of the background speakers in its parameter estimation procedure during the enrollment stage.

2.2. Verification Procedure

For each verification session, the test vectors from the utterances of a claimant are concatenated to form a test sequence $\mathcal{T} = [\vec{x}_1, \vec{x}_2, \dots, \vec{x}_T]$. The sequence is then divided into a number of overlapping segments containing T_s consecutive vectors. For a segment \mathcal{T}_s , the normalized average outputs

$$z_k = \frac{1}{T_s} \sum_{\vec{x} \in \mathcal{T}_s} \frac{\exp\{\tilde{y}_k(\vec{x})\}}{\sum_{r=1}^2 \exp\{\tilde{y}_r(\vec{x})\}} \quad k = 1, 2 \quad (2)$$

corresponding to the speaker and anti-speaker classes are computed, where

$$\tilde{y}_k(\vec{x}) = \frac{y_k(\vec{x})}{P(C_k)}$$

represents the scaled output and $P(C_k)$ the prior probability of class C_k .

Verification decisions are based on the criterion:

$$\text{If } z_1 - z_2 \begin{cases} > \zeta & \text{accept the claimant} \\ \leq \zeta & \text{reject the claimant} \end{cases} \quad (3)$$

where $\zeta \in [-1, 1]$ is an *a priori* threshold that has been determined during enrollment (see Section 3 below). A verification decision is made for each segment, and the error rate (either FAR or FRR) is the proportion of incorrect verification decisions to the total number of decisions. Details of the verification procedure can be found in [9].

3. A PRIORI THRESHOLD DETERMINATION

To determine the *a priori* thresholds, we need to obtain the FAR and FRR as a function of thresholds using enrollment data only. We propose three methods to achieve this goal. They are denoted as Baseline, Pseudo-Impostor Based Threshold Determination (PIBTD) and Sampling Pseudo-Impostor Based Threshold Determination (SPIBTD) in this paper.

¹In this work, γ_j was set to 3.0 for all j .

3.1. Baseline

This method, being very similar to that of [4], is to form a baseline for comparison. Specifically, for each registered speaker in the system, five other speakers whose speech are closest to that of the speaker are selected from the population, forming an anti-speaker set. Note that this is analogous to the ratio speakers of [4]. The speech of the speaker and the anti-speakers are used to train the speaker model. Then, the same speech data from the anti-speakers are applied to the speaker model according to the above verification procedure. The FAR as a function of the threshold is obtained by adjusting the threshold ζ , resulting in an FAR curve. Similarly, the speaker’s utterances which have been used to train the speaker model are applied to the model to obtain an FRR curve. A typical example of these curves is shown in Fig. 1(a).

A threshold is chosen so as to equalize the FAR and FRR. When the FAR and FRR curves do not cross each other, i.e., there exist a range of thresholds for which both FAR and FRR are zero, we select the middle of the zero crossing points of the FAR and FRR curves as the threshold. The reason behind this selection strategy is that the FAR and FRR curves are likely to shift to the right and left respectively during verification, as shown in Fig. 1(a). This is because the anti-speaker set, which contains 5 anti-speakers only, is unlikely to form a good representation of the impostor population. When the speech of a real impostor is applied, the speaker model produces a very different response.

3.2. Pseudo-Impostor Based Threshold Determination (PIBTD)

Using the same set of utterances for training the speaker model as well as for determining the threshold has a serious drawback. After training, the speaker model is likely to bias towards representing the training utterances of the speaker and anti-speakers. If the same set of utterances is applied to the speaker model during threshold determination, the threshold obtained is likely to be a biased one. To resolve this problem, PIBTD uses another set of speakers, called pseudo-impostor set, together with another set of utterances produced by the registered speaker for threshold determination. More specifically, after training the speaker model, five pseudo-impostors are randomly selected from the population and applied to the speaker model. These pseudo-impostors, being different from the anti-speakers and never seen by the speaker model before, are more likely to form a better representation of the impostor population. This prevents the FAR curve from shifting along the threshold axis drastically during verification, as shown in Fig. 1(b).

3.3. Sampling Pseudo-Impostor Based Threshold Determination (SPIBTD)

Obviously, the representation of the impostor population can be improved by increasing the number pseudo-impostors and anti-speakers. However, increasing the size of these sets will also lead to unrealistic enrollment time. SPIBTD aims at reducing the error rate and improving the robustness of the thresholds without increasing the enrollment time. The basic idea is to randomly select from the utterances of a large number of pseudo-impostors and anti-speakers to form a set of vectors for training the speaker model and for determining the threshold. In this way, the number of training vectors and the enrollment

time remain the same. Another advantage of this sampling strategy is that the resulting training vectors become more representative of the impostor population, for they are derived from many pseudo-impostors instead of five as in PIBTD. As shown in Fig. 1(c), this makes the position of FAR curves more predictable as compared to that of PIBTD. The reduction in the displacement between the FAR curves obtained during enrollment and verification means that we can use the FAR curve to determine the threshold. More specifically, the threshold is adjusted until the FAR obtained during enrollment falls below an application dependent level. In this work, we set this level to 0.1%.

4. EXPERIMENTAL EVALUATION

All of the 138 speakers (108 male, 30 female) in the YOHO corpus [10] have been used in the experiments. For each speaker in the corpus, there are 4 enrollment sessions with 24 utterances in each session, and 10 verification sessions of 4 utterances each. Each utterance is composed of three 2-digit numbers (e.g. 34-52-67). All sessions were recorded in an office environment using a high quality telephone handset and sampled at 8 kHz.

The enrollment process involves two steps. First, for each speaker in the corpus, 72 utterances from his/her first three enrollment sessions and 480 utterances from the 4 enrollment sessions of 5 anti-speakers (Baseline and PIBTD) were used to train a speaker model. For SPIBTD, the 480 utterances were randomly selected from 45 anti-speakers. Second, the *a priori* threshold was determined by using either the anti-speaker set (Baseline) or the pseudo-impostor set (PIBTD and SPIBTD)² together with the speaker's speech. The speaker's speech was derived from the training utterances (Baseline) or from other utterances for which the model has never seen before (PIBTD and SPIBTD).

Verification was performed using each speaker in the corpus as a claimant, with 45 impostors being randomly selected from the remaining speakers (excluding the anti-speakers and pseudo-impostors) and rotating through all speakers. The claimant's utterances, which were derived from his/her 10 verification sessions, were concatenated to form a sequence of features vectors. Similarly, the feature vectors of 45 impostors were randomly selected and then concatenated to form a test sequence. Verification decisions were made according to (3) with the segment length T_s in (2) being set to 300.³ This arrangement produces approximately 1000 genuine trials and 1000 impostor attempts for each speaker.

LP-derived cepstral coefficients were used as acoustic features. For each utterance, the silent regions were removed by a silent detection algorithm based on the energy and zero crossing rate of the signal. The remaining signals were pre-emphasized by a filter with transfer function $1 - 0.95z^{-1}$. Twelfth-order LP-derived cepstral coefficients were computed using a 28 ms Hamming window at a frame rate of 14 ms. These feature vectors were used to train a set of speaker models (EBF networks) with 12 inputs, two outputs, and 24 centers, where 8 centers were contributed by the speaker and the remaining 16 by the

²Five pseudo-impostors were used in PIBTD, whereas in SPIBTD, the pseudo-impostor set is constructed by selecting the utterances of 45 pseudo-impostors randomly.

³This is approximately equal to the length of 4 utterances. Thus, the results can be compared with those of [4].

anti-speakers.

5. RESULTS AND DISCUSSION

Fig. 1 depicts FARs and FRRs as a function of thresholds for Baseline, PIBTD, and SPIBTD. Some interesting results can be observed from these figures. First, Fig. 1(a) shows that there is a large displacement between the FAR curve corresponding to enrollment and that corresponding to verification when anti-speakers' utterances were used to determine the FAR curve during enrollment. Second, when pseudo-impostors were used to obtain the FAR curve during enrollment, the displacement is considerably reduced, as shown in Fig. 1(b). Third, the displacement is further reduced for SPIBTD (Fig. 1(c)) where random utterances were sampled from a large number of pseudo-impostors, suggesting that the verification performance of the system can be reliably predicted. Fig. 1(c) also suggests that using the FAR curve to determine the threshold is better than choosing the middle of zero FAR and FRR as the threshold. This is because the position of the FRR curves is difficult to predict.

Method	Enrollment		Verification		
	FAR	FRR	FAR	FRR	ERR
Baseline	0.00	0.00	5.45	4.67	1.59
PIBTD	0.20	0.20	0.64	11.44	0.95
SPIBTD	0.24	0.15	1.03	4.86	0.66

Table 1: Error rates (in %) obtained by different methods.

Table 1 summarizes the FARs, FRRs, and ERRs obtained by Baseline, PIBTD and SPIBTD. The ERRs were obtained by adjusting the thresholds *a posteriori*. All results are based on the average of 138 speakers. Although the baseline method yields zero FAR and FRR during enrollment, they are significantly larger during verification. This agrees with our claim that the baseline method has difficulty in predicting the verification performance. While the FAR and FRR during enrollment are the same for PIBTD, they differ widely during verification. This phenomenon is also demonstrated in [4]. The large difference is mainly due to the fact that the pseudo-impostor set is too small to provide a good representation of the impostor population, resulting in some underestimated thresholds. SPIBTD, on the contrary, yields a much lower FRR as compared to PIBTD. This is because it samples the utterances of a large number of pseudo-impostors, thus providing a much better impostor representation. The ERR of SPIBTD is also about half of that of [4], suggesting that sampling the utterances from a large number of anti-speakers is able to produce a more robust speaker model.

Fig. 2 depicts the FAR and FRR of individual speakers. Of particular interest is that for the baseline method, most of the speakers have a low FAR but a high FRR or vice versa, while for PIBTD, most speakers have a high FRR. This confirms the results in Table 1. Fig. 2(c) shows that the FAR and FRR of most speakers are smaller than 10%, again confirming the results in Table 1 and suggesting that SPIBTD is more robust.

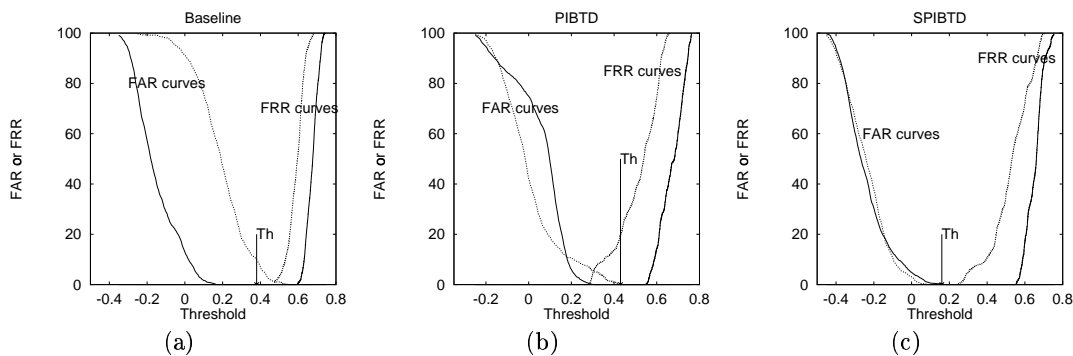


Figure 1: Variations of FARs and FRRs with respect to decision thresholds during enrollment (solid) and verification (dots) using (a) Baseline, (b) PIBTD, and (c) SPIBTD. The label “Th” denotes the *a priori* threshold found by these methods.

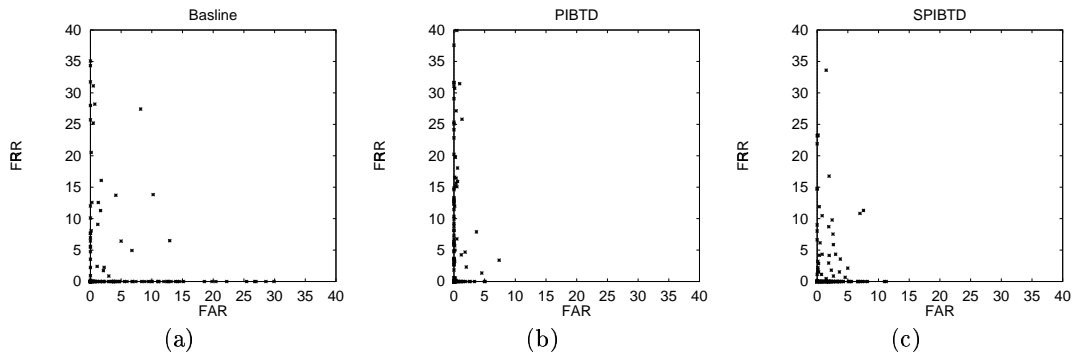


Figure 2: FARs and FRRs of all speakers during verification using (a) Baseline, (b) PIBTD, and (c) SPIBTD.

6. CONCLUSIONS

This paper addresses the problem of determining *a priori* thresholds for speaker verification. Conventional approaches have been compared with a new one. It was shown that robust thresholds can be obtained by simulating an operation environment as close as possible to the real one. Our proposed method is able to predict the verification performance accurately by using enrollment data only, leading to more reliable thresholds. The proposed method is also able to find a better balance between FARs and FRRs.

7. ACKNOWLEDGMENT

This work was supported by The H.K. Polytechnic University Grant No. G-S472.

8. REFERENCES

- [1] S. Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Trans. on Acoustics Speech and Signal Processing*, ASSP-29(2):254–272, 1981.
- [2] D. K. Burton. Text-dependent speaker verification using vector quantization source coding. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, ASSP-35(2):133–143, 1987.
- [3] J. M. Naik, L. P. Netsch, and G. R. Doddington. Speaker verification over long distance telephone lines. In *Proc. ICASSP’89*, 1989.
- [4] A. Higgins, L. Bahler, and J. Porter. Speaker verification using randomized phrase prompting. *Digital Signal Processing*, 1:89–106, 1991.
- [5] T. Matsui and S. Furui. Likelihood normalization for speaker verification using a phoneme- and speaker-independent model. *Speech Communications*, 17:109–116, 1995.
- [6] C. S. Liu, C. H. Lee, B. H. Juang, and A. E. Rosenberg. Speaker recognition based on minimum error discriminative training. In *Proc. ICASSP’94*, volume 1, pages 325–328, 1994.
- [7] J. B. Pierrot et al. A comparison of *a priori* threshold setting procedures for speaker verification in the CAVE project. In *Proc. ICASSP’98*, pages 125–128, 1998.
- [8] J. Moody and C. J. Darken. Fast learning in networks of locally tuned processing units. *Neural Computation*, 1:281–194, 1989.
- [9] M. W. Mak and C. K. Li. Elliptical basis function networks and radial basis function networks for speaker verification: A comparative study. In *IJCNN’99*, July 1999.
- [10] Jr. J. P. Campbell. Testing with the YOHO CD-ROM voice verification corpus. In *ICASSP95*, pages 341–344, 1995.