A NEW TWO-STAGE SCORING NORMALIZATION APPROACH TO SPEAKER VERIFICATION

M. W. Mak and W. D. Zhang

Center for Multimedia Signal Processing, Dept. of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong, China

M. X. He

Ocean Remote Sensing Laboratory of Ministry of Education, Ocean Remote Sensing Institute, Ocean University of Qingdao, China

ABSTRACT

In speaker verification, the cohort and world models have been separately used for scoring normalization. In this work, we embed the two models in elliptical basis function networks and propose a two-stage decision procedure for improving verification performance. The procedure begins with normalization of an utterance by a world model. If the difference between the resulting score and a world threshold is sufficiently large, the claimant is accepted or rejected immediately. Otherwise, the score will be normalized by a cohort model, and the resulting score will be compared with a cohort threshold to make a final accept/reject decision. Experimental evaluations based on the YOHO corpus suggest that the two-stage method achieves a lower error rate as compared to the case where only one background model is used.

1. INTRODUCTION

Most speaker verification (SV) systems require a set of speaker models built from customers' speech and a speaker-dependent or speaker-independent background model for scoring normalization. The cohort model-based normalization approach (CMN) [1, 2] and the world model-based normalization approach (WMN) [2, 3] are two of the most popular techniques. Studies have shown that inclusion of these models not only improves speaker separability, but also allows decision thresholds to be set easily. Theoretically, these two approaches represent two different paradigms for decision-making, and each has its own strengths and weaknesses. For example, Furui [7] pointed out that CMN is robust to similar, same-gender impostors (wolves) but is vulnerable to attacks made by opposite-gender impostors, whereas WMN is robust to the general population but is vulnerable to attacks made by the wolves.

Gu and Thomas [5] combined CMN and WMN to obtain a hybrid score:

$$S_{comb} = \alpha S_{cohort} + (1-\alpha) S_{world} \qquad 1 \le \alpha \le 1 \qquad (1)$$

where S_{cohort} and S_{world} represent the CMN score and WMN score (in log-likelihood domain) respectively, and α is a combining factor. The linear combination, however, has its limitations because (a) S_{cohort} and S_{world} must be computed for every decision, (b) the evaluation of S_{cohort} is very costly, and (c) the value of α must be optimized experimentally.

In this paper, we propose a two-stage decision-making method (as shown in Fig. 1) to combine the CMN and WMN. Unlike the one-stage parallel combination suggested in [5], our method combines the two models in a two-stage sequential manner. The method is based on the idea that the world model can discriminate the speaker from the general population very well. Therefore, we do not need to use the cohort model when the claimant's score is very low or very high, because the claimant is very likely an impostor or a true speaker. We only need the cohort model when the claimant's score falls into an uncertain region. This may occur if the claimant is in fact an impostor but whose voice is very similar to that of the true speaker, or if the claimant is the true speaker whose voice exhibits unexpected changes. The experimental results based on the YOHO corpus show that this sequential combination approach achieves a lower error rate as compared to the case where either the world model or the cohort model is used for scoring normalization.

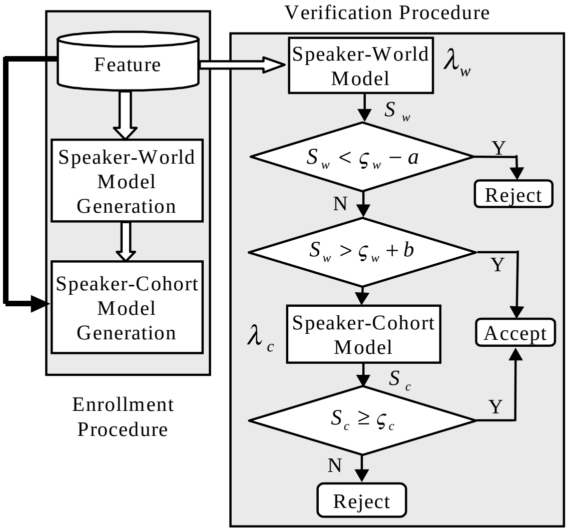


Figure 1. Enrollment and verification procedures of the two-stage SV system.

2. EBFN-BASED VERIFICATION

2.1 Speaker Models: EBF Networks

Elliptical basis function (EBF) networks, which have been successfully applied to speaker verification [8], were used as the speaker models in this study. The *k*th output (*k* = 1,…,*K*) of an EBFN with *I* inputs and *J* function centers has the form

$$y_k(\vec{x}_p) = w_{k0} + \sum_{j=1}^{J} w_{kj} \phi_j(\vec{x}_p) \qquad p = 1, \ldots, N \qquad (2)$$

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 (2), $\vec{x}_p$ 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 that controls the spread of the *j*-th basis function. Typically, the mean vectors are found by the K-means algorithm and the covariance matrices are estimated by the sample covariance or the EM algorithm. Once the basis function parameters are known, the output weights can be determined by the least squares approach using the technique of singular value decomposition.

2.2 Scoring Normalization

EBF networks incorporate the idea of scoring normalization in their discriminative training procedure. In [8], each registered speaker is assigned an EBF network with two outputs. The first output is trained to output a '1' for the customer' speech and a '0' for other speakers' utterances, and vice versa for the second output. Two sets of data are required to train the model, one of them being derived from the customer and another from some other speakers (denoted as anti-speakers). The random selection of a large number of anti-speakers from the general population enables the network to embed a world model in the speaker model. Therefore, the network is capable of discriminating the customer from the general population, and we denote the network as the speaker-world model. On the other hand, the selection of a small number of anti-speakers whose speech is similar to the customer enables the network to embed a cohort model in the speaker model. The network will therefore be capable of discriminating the customer from the 'close' impostors, and this network is referred to as the speaker-cohort model.

During verification, the feature vectors derived from the utterances of a claimant are concatenated to form a vector sequence $\Im = [\vec{x}_1, \vec{x}_2, ..., \vec{x}_T]$, and the sequence is presented to the speaker-world (or speaker-cohort) model. The normalized average outputs

$$S_k = \frac{1}{T} \sum_{\vec{x} \in \Im} \frac{\exp\{\tilde{y}_k(\vec{x})\}}{\sum_{r=1}^{2} \exp\{\tilde{y}_k(\vec{x})\}} \qquad k = 1,2 \qquad (3)$$

corresponding to the speaker and anti-speaker (or cohort) classes are computed, where $\tilde{y}_k(\vec{x}) = y_k(\vec{x}) / P(C_k)$ represents the scaled output and $P(C_k)$ represents the prior probability of class C_k. The normalized score of the claimant's utterances can be obtained by

$$S = S_1 - S_2 \qquad (4)$$

where S_1 is the score of the claimant and S_2 is the normalization term given by the world model or the cohort model. If only one of the two normalization methods is used in the SV system, S should be compared with an *a priori* threshold $\zeta \in [-1,1]$, which is determined during enrollment [4] (also see Section 2.3 below) to make a decision.

Unlike the hidden Markov model-based and Gaussian mixture model-based speaker verification systems in which the cohort model is constructed during verification, the EBFN approach embeds the characteristics of the cohort speakers in its parameter estimation procedure during enrollment. The normalization terms are produced as part of the network outputs during verification. Hence, the EBFN approach is more computationally efficient when compared to the cohort model-based approach.

3. TWO-STAGE APPROACH

To determine the *a priori* thresholds, we need to obtain the false acceptance rate (FAR) and the false rejection rate (FRR) as a function of the thresholds using enrollment data only. Our previous work [4] has shown that the speaker-world model constructed from a large amount of training data from the anti-speakers can help determine a reliable threshold. To avoid annoying the true speakers with too many false alarms, the threshold must be set at a value that is as small as possible. However, with a small threshold, the speaker-world model is likely to favour accepting similar, same-gender impostors (wolves). Therefore, a cohort model is necessary to address this problem.

If the normalized score is significantly larger (or smaller) than the threshold, we can accept (or reject) the claimant with a small chance of false identification. However, when the claimant's score falls into an uncertain region around the threshold, it becomes difficult to determine whether the claimant is in fact an impostor whose voice is very similar to the true speaker or whether the claimant is the true speaker but whose voice suffers from unexpected changes. As the speaker-cohort model has the ability to discriminate the speaker from similar, same-gender impostors, a combination of the speaker-world and speaker-cohort models is expected to improve the accuracy of the verification decision.

Fig. 1 illustrates the two-stage sequential approach that we propose. The process begins with the training of a speaker-world model to discriminate the speaker from the general population. The general population is composed of 32 female speakers and 45 male speakers in the YOHO corpus. Another 10 male speakers (called pseudo-impostors) are randomly selected from the remaining 60 male speakers to help determine the *a priori* thresholds (see [4] for the details of threshold determination). The speech of 55 male speakers (45 male anti-speakers and 10 pseudo-impostors) was presented to the speaker-world model, and the speaker-specific threshold ζ_w was adjusted until the FAR reached an application dependent level (the level was set to 0.5% in the experiments).

Then, we fed the utterances of 55 male speakers used in the training and threshold determination to select 15 cohort speakers and to train another network, namely the speaker-cohort model. This model is to discriminate the speaker from the wolves. The utterances of 25 male anti-speakers (whose voices are similar to the speaker's) were fed into the speaker-cohort model to determine another threshold ζ_c, based on a fixed pre-defined FAR (again 0.5% was used in the experiments).

During verification, we presented the claimant's utterances to the speaker-world model. If the score was smaller than $\zeta_w - a$, we would reject the claimant. If it was higher than $\zeta_w + b$, we would accept the claimant. The parameters *a* and *b* were application dependent. For a system that requires a high level of security, *a* should be small to minimize false acceptances, and *b* should be large to reduce the chance of accepting similar, same-gender impostors. This defines an uncertain region within which the speaker-cohort model is called upon to make the final verification decision. More specifically, when the output of the speaker-world model falls into $[\zeta_w - a, \zeta_b + b]$, we would feed the utterances to the speaker-cohort model. If the output of the speaker-cohort model is larger than ζ_c, we accept the claimant; otherwise, we would reject the claimant (see Fig. 1).

Note that the speaker-cohort model is constructed from the speech of the speaker and 15 cohort speakers; its ability to discriminate the speaker's speech from those of similar, same gender impostors should be much better than the speaker-world model. The existence of the cohort model means that the system is more robust to the attacks made by these impostors, thus improving the security of the system. With this guideline in mind, we set *a* to zero and *b* to 0.15 to simulate a highly secured system in our experiments. For these values of *a* and *b*, we found that there is about 20% chance that the cohort model will be used.

On the other hand, a user friendly system requires a low FRR, which suggests that we need to accommodate unexpected changes in the customers' speech, especially when the customers' speech sounds like an impostor's. If this situation occurs, we should delay the decision and allow the speaker-cohort model to make the final decision. Without the speaker-cohort model, these customers would be rejected immediately by the speaker-world models, causing a high FRR. These easily rejected speakers are commonly referred to as goats. As the speaker-cohort models are designed to discriminate the true speakers from highly similar impostors, they may also be good at detecting the goats. To allow the speaker-cohort models to detect goats, we need to define a new uncertain region into which the score of all goats will fall. To achieve this, *a* should be large and *b* should be small. Therefore, we set *a* to 0.15 and *b* to 0.0.

4. EXPERIMENTS

All of the 138 speakers (106 male, 32 female) in the YOHO corpus [6] were used in the experiments. For each speaker in the corpus, there were four enrollment sessions with 24 utterances in each session, and ten verification sessions of four 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. In our experiments, only 106 male speakers were used as speakers (customers) because the number of female speakers was too small. If all of the female speakers were used as anti-speakers, none of them would be left for threshold determination and evaluation. If we used all the male speakers to test the female speaker models, the results would be biased.

The enrollment process involved two steps. In the first step, for each male speaker in the corpus, 24 utterances from his first enrollment session and 1800 utterances from the first enrollment session of 75 anti-speakers (32 females and 45 males) were used to train a speaker-world model. To reduce the training time, we randomly selected one-eighth of the feature vectors of the anti-speakers. Therefore, only 225 utterances from the anti-speakers were used for training. In the second step, the *a priori* threshold ζ_w was determined by using the anti-speakers and the pseudo-impostors. Second, the utterances of the anti-speakers and pseudo-impostors were fed to the speaker-world model, and their scores were sorted in decreasing order. The first 15 speakers in the sorted list were selected to form the cohort speakers set and were used together with the speaker's speech to train the speaker-cohort model. Again, we selected one-third of the feature vectors from the cohort speakers to reduce the training time. The first 25 speakers in the sorted list were selected to determine the threshold ζ_c.

Verification was performed by using each of the male speakers in the corpus as a claimant, and 50 impostors randomly selected from the remaining speakers (excluding the anti-speakers and pseudo-impostors) and by rotating all the speakers. The claimant's utterances, which were derived from his 10 verification sessions, were concatenated to form a sequence of features vectors. Similarly, the feature vectors of 50 impostors were randomly selected and then concatenated to form a test sequence. Verification decisions were made with the segment length *T* in (3) being set to 300. This was approximately the length of 4 utterances. This arrangement produced 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 the speaker-world models and speaker-cohort models (EBF networks) with 12 inputs, two outputs, and 24 (20) centers for the speaker-world (speaker-cohort) model, of which 8 centers were contributed by the speaker and the remaining 16 (12) by the anti-speakers (cohort-speakers).

5. RESULTS AND DISCUSSION

Table 1 summarizes the FRRs, FARs and verification time obtained by the SV systems using different normalization methods. All results are based on the average of 106 male speakers.

Normalization Method	FRR	FAR	Time
Speaker-world model only	11.47	1.03	T_w
Speaker-cohort model only	7.53	3.25	T_c
2-stage, 2-models, *a*=0.0, *b*=0.15	12.82	0.73	$T_w + 0.2T_c$
2-stage, 2-models, *a*=0.15, *b*=0.0	6.87	2.25	$\mathrm{T}_w + 0.3T_c$

Table 1: Average error rates (in %) obtained and theoretical verification time taken by different normalization methods.

Table 1 shows that the FAR of the speaker-world model is much smaller than that of the speaker-cohort model, which suggests that the former is more capable of discriminating the speaker from a general population. This capability is important because, in some applications, it is important to reduce the FAR to make the SV system more robust to impostor attacks.

The third and fourth rows of Table 1 show the results of the two-stage decision-making approach. Obviously, the two-stage approach (with a=0 and b=0.15) reduces the FAR by about 29%, from 1.03% to 0.73%. This substantial reduction only causes a slight increase in the FRR, from 11.47% to 12.82%—an 11.7% increase on average. Apparently, a net improvement in verification performance has been achieved. The improvement is mainly due to the fact that the FRRs of most speakers remain unchanged and the increase in FRRs is very small.

The reason for the substantial reduction in FAR and the small increase in FRR can be found in Fig. 2. Fig. 2(a) plots the FARs of 106 speakers obtained by the two-stage SV system (with a=0, b=0.15) against the FARs of the one-stage SV system (using the speaker-world model only). Likewise, Fig. 2(b) plots the corresponding FRRs. Fig. 2(a) demonstrates that the FARs achieved by the two-stage system are either identical to or smaller than that in the one-stage system, which suggests that the speaker-cohort model is robust to highly similar impostors (wolves). Fig. 2(b) shows that although there is an increase in the FRR for some speakers, the increase is minimal, and the FRR of most speakers remains unchanged.

Figs. 2(c) and 2(d) show that we can further reduce the FRRs, without significantly increasing the FARs, by shifting the uncertain region. Of particular interest is that none of the speakers in the two-stage system has an FRR that is higher than that of the one-stage system. Although the reduction in the average FRR among the 106 speakers is substantial, the increase in the FAR is not significant at all, as can be seen from Fig. 2(c).

Table 1 also shows the theoretical verification time required by the different scoring methods. If we combine the world and cohort models in a one-stage approach (as in [5]), the time taken will be the sum of the time required by each of the two models, i.e. $T_w + T_c$. However, the cohort model in our system will be used for only about 20% to 30% of time time. Hence, the verification time is reduced. The reduction is more substantial when compared with the conventional cohort-based scoring methods. This is because in the latter, the cost of finding a cohort model during verification is a linear function of the cohort size, whereas in our approach, the cohort model has already been computed and embedded in the EBFNs during enrollment.

6. CONCLUSION

The speaker-world model and the speaker-cohort model focus on different regions of the feature space. Better results are expected by integrating them into a single framework. This paper has described a two-stage decision-making approach that combines these models. The results demonstrate that the combined model can reduce error rates, and that the two-stage sequential approach probably will take up a shorter verification time as compared to the one-stage parallel combination approach.

7. ACKNOWLEDGEMENT

This work was supported by The H.K. Polytechnic University Grant No. 1.42.37.A410.

8. REFERENCES

[1] A. E. Rosenberg, et al, “The use of cohort normalized scores for speaker verification,” ICSLP, pp. 599-602, 1992.

[2] A. E. Rosenberg and S. Parthasarathy, “Speaker background models for connected digit password speaker verification,” ICASSP, pp. 81-84, 1996.

[3] T. Matsui and S. Furui, “Likelihood normalization for speaker verification using a phoneme- and speaker-independent model,” Speech Communication, vol. 17, pp. 109-116, 1995.

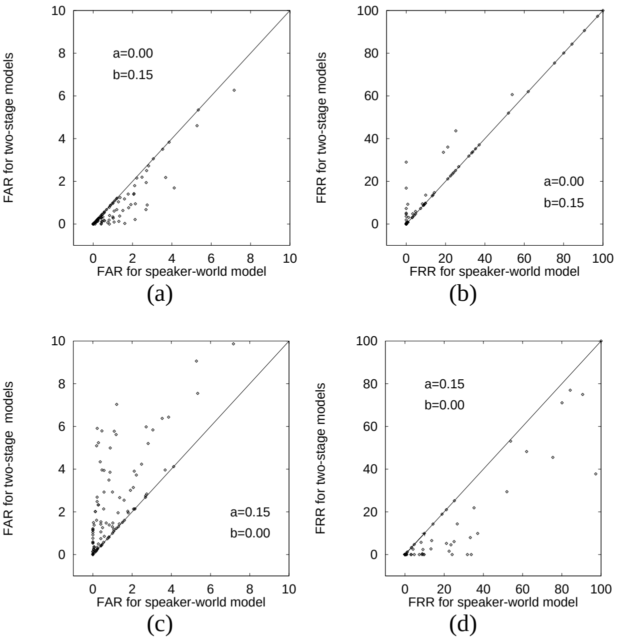
[4] W. D. Zhang, et al, “A priori threshold determination for phrase-prompted speaker verification,” Eurospeech’99, Vol. 2, pp. 1023-1026, September, 1999.

[5] Y. Gu and T. Thomas, “A hybrid score measurement for HMM-based speaker verification,” ICASSP’99.

[6] J.r. J. P. Campbell, “Testing with the YOHO CD-ROM voice verification corpus,” ICASSP, pp. 341-344, 1995.

[7] S. Furui, “An overview of speaker recognition technology,” *Automatic speech and speaker recognition,* Eds. C. H. Lee, F. K. Soong and K. Kuldip, Kluwer Academic Pub, 1996.

[8] M. W. Mak and S.Y. Kung, “Estimation of Elliptical Basis Function Parameters by the EM Algorithms with Application to Speaker Verification," IEEE Trans. on Neural Networks, Vol. 11, No. 4, pp. 961-969, July 2000..



(a) (b)

(c) (d)

Figure 2 Graphs showing the relationship between the error rates obtained by the one- and two-stage models. (a) FAR and (b) FRR of the two-stage model (a=0.0, b=0.15) against the FAR and FRR of the speaker-world model. (c) and (d) are identical to (a) and (b), except for a=0.15 and b=0.0.