On the Use of Score Pruning in Speaker Verification for Speaker Dependent Threshold Estimation

Javier R. Saeta[1], Javier Hernando[2]

[1] Biometric Technologies, S.L. Barcelona, Spain
j.rodriguez@biometco.com
[2] TALP Research Center. Universitat Politècnica de Catalunya, Spain
javier@talp.upc.es

Abstract

The use of a priori speaker-dependent thresholds has been shown convenient for speaker verification. However, their estimation is highly affected by the difficulty of obtaining data from impostors, the mismatched conditions, the scarcity of data in real applications and the need of setting the threshold a priori, during enrollment. In this context, possible outliers, i.e., those client scores which are distant with respect to mean in terms of Log-Likelihood Ratio (LLR), could lead to wrong estimations of client mean and variance. To overcome this problem, we propose here several methods based on pruning LLR scores with different statistical criteria. Before estimating the threshold, score pruning removes outliers and improves subsequent estimations. To solve the problem of impostor data, we also suggest a speaker dependent threshold estimation with only data from clients. Text-dependent and text-independent experiments have been carried out by using a telephonic multisession database in Spanish with 184 speakers, that has been recorded by the authors.

1. Introduction

In speaker verification, a utterance is compared to the speaker model and the speaker is accepted if the LLR is above a certain threshold and rejected if below. LLR can be normalized by the Universal Background Model (UBM) or a speaker cohort. To compare two systems, it is common to use the equal error rate (EER). However, in real applications, EER is less significant because a certain False Acceptance Rate (FAR) or False Rejection Rate (FRR) is usually required. To obtain a specific rate, we only have to adjust our threshold.

In development tasks, the threshold can be estimated a posteriori. However, in real applications, the threshold is to be established a priori. This limitation elicits several problems.

First, it is common to have only a few data from clients. The amount of data is essential for the correct training of the speaker model. The lack of data leads to bad threshold estimations because the influence of outliers will be higher in this case.

Second, it is difficult to obtain data from impostors, especially in phrase-prompted cases. The models are often estimated with data from clients, impostors or both of them. In real applications, it is complicated to acquire and to select the most representative data from impostors, for instance when using key words or sentences.

Apart from these considerations, we should also note that a speaker-dependent threshold should be used because it better reflects speaker peculiarities and intra-speaker variability than a speaker-independent threshold. A speaker dependent threshold can also be transferred to a speaker specific score normalization technique with a global threshold.

In this paper, we propose a new speaker-dependent threshold based only on data from clients, as a basis to test some algorithms to remove the non-representative LLR scores by means of Score Pruning (SP) techniques.

The speaker dependent threshold estimation method is a linear combination of mean and standard deviation from clients. The advantage of this method is that it does not use data from impostors. Besides, it mitigates the problem of data scarcity employing the same utterances used to train the model.

The main problem when we have only a few utterances to estimate the threshold is that some of them could produce non-representative scores, the 'outliers', i.e., client scores which are distant with respect to mean. This is common when there are background noises, distortions or strange articulatory effects, especially in mismatched conditions. The outliers affect mean and standard deviation estimation.

Score pruning techniques suppress the effect of non-representative scores, removing them and contributing to a better estimation of means and variances in order to set the speaker dependent threshold.

We describe the state-of-the-art in speaker dependent threshold estimation in Section 2. Section 3 shows the score pruning methods that are introduced in this paper and Section 4 presents the experimental results. Finally, we add some conclusions in Section 5.

2. Threshold estimation approaches

Several approaches have been proposed to automatically estimate a priori speaker dependent thresholds. Conventional methods have faced the scarcity of data and the problem of an a priori decision, using client scores, impostor data, a speaker independent threshold or some combination of them. In [1], we can find a threshold estimation as a linear combination of impostor scores mean ($\hat{M}_{\bar{X}}$) and standard deviation from impostors $\hat{\sigma}_{\bar{X}}$ as follows:

$$\Theta_x = \alpha \ (\hat{M}_{\bar{X}} - \hat{\sigma}_{\bar{X}}) + \beta \tag{1}$$

where α and β should be obtained empirically.

Three more speaker dependent threshold estimation methods similar to (1) are introduced in (2), (3) and (4) [2, 3]:

$$\Theta_x = \hat{M}_{\bar{X}} + \alpha \, \hat{\sigma}^2_{\bar{X}} \tag{2}$$

where $\hat{\sigma}^2_{\bar{X}}$ is the variance estimation of the impostor scores, and:

$$\Theta_x = \alpha\, \hat{M}_{\bar{X}} + (1-\alpha)\, \hat{M}_X \quad (3)$$

$$\Theta_x = \Theta_{SI} + \alpha\, (\hat{M}_X - \hat{M}_{\bar{X}}) \quad (4)$$

where $\hat{M}_X$ is the client scores mean, Θ_{SI} is the speaker independent threshold and α is a constant, different for every equation and empirically determined. Equation (4) is considered as a fine adjustment of a speaker independent threshold.

Another expression introduced in [4] encompasses some of these approaches:

$$\Theta_x = \alpha\, (\hat{M}_{\bar{X}} + \beta\, \hat{\sigma}_{\bar{X}}) + (1-\alpha)\hat{M}_X \quad (5)$$

where α and β are constants which have to be optimized from a pool of speakers.

Methods previously shown do not use the client variance because they consider that its estimation is unreliable when only a few data are available.

Other approaches to speaker dependent threshold estimation are based on a normalization of client scores (S_M) by mean ($\hat{M}_{\bar{X}}$) and standard deviation ($\hat{\sigma}_{\bar{X}}$) from impostor scores [5]. This approach is based on znorm [6]:

$$S_{M,norm} = \frac{S_M - \hat{M}_{\bar{X}}}{\hat{\sigma}_{\bar{X}}} \quad (6)$$

We should also make reference to another threshold normalization technique such as hnorm [7], which makes use of a handset-dependent normalization.

Some other methods are based on FAR and FRR curves [8]. Speaker utterances used to train the model are also employed to achieve the FRR curve. On the other hand, a set of impostor utterances is used to obtain the FAR curve. The threshold is adjusted to equalize both curves.

There are also other approaches [9] based on the idea of the difficulty of achieving impostor utterances which fit the client model, especially in phrase-prompted cases. In these cases, it is difficult to obtain the whole phrase from impostors. The alternative is to use some words from other speakers or different databases, to complete the whole phrase.

On the other hand, it is worth noting that there are other methods which use different estimators for mean and variance. In [10], we can observe two of them, classified according to the percentage of used frames. Instead of employing all frames, one of the estimators use 95% most typical frames discarding 2,5% maximum and minimum frame likelihood values. An alternative is to use 95% best frames, removing 5% minimum values. With the selection of a high percentage of frames and not all of them, we remove those frames which are out of range of typical frame likelihood values.

Finally, we should note that the use of impostors data to estimate the threshold creates difficulties in real applications. In general, it is not easy to obtain data from impostors for certain uses, for instance in phrase-prompted cases. Furthermore, it is very difficult to select the impostors in a right way, because they could become clients in the future. To solve these problems, we define a new speaker dependent threshold estimation [11] based on data from clients only. Like the expressions in Section 2, this is a linear combination of mean and standard deviation estimations, but in this case it uses only data from clients. It is very similar to (2), but employs standard deviation instead of variance and uses also the client mean from LLR scores. The client mean estimation is adjusted by means of the client standard deviation estimation and α, as follows:

$$\Theta_x = \hat{M}_X - \alpha\, \hat{\sigma}_X \quad (7)$$

where $\hat{M}_X$ is the client scores mean, $\hat{\sigma}_X$ is the standard deviation and α is a constant which has to be set experimentally on a development population. Equation (7) implements an adjustment of the mean through the standard deviation.

3. Score pruning

The presence of outliers can elicit wrong estimations of mean and variance of client scores. The influence of outliers becomes even more significant if the standard deviation or the variance are multiplied by a constant, like in expressions (1) and (2). The threshold of some speakers is probably wrong fixed due to the outliers. In this way, our goal is to minimize their presence.

Pruning is a technique which has been previously applied to frames [12, 13, 14]. It has been used in the parameterization stage to cut off certain frames in order to improve the performance of speaker recognition. We use here the concept of score pruning [4] as a suitable method to remove outliers and obtain better estimations of means and variances.

For this purpose, we introduce an algorithm that sets mean and standard deviation estimations. It begins to consider the most distant score with respect to the mean, and will continue with the second most distant if necessary. The main questions here will be: 1) how to decide the elimination of a score, and 2) when to stop the algorithm.

To solve the first question, we use a parameter to control the difference between the standard deviation estimation with and without the most distant score, the potential outlier. We define Δ as the percentage of variation of the standard deviation from which we consider to discard a score. Δ will decide if the score is considered as an outlier or not. If the percentage of variation exceeds Δ, we confirm this score as an outlier.

In the case we have decided that a score is non-representative, we recalculate mean and standard deviation estimations without it. At this point, we look for the next most distant score. A second question appears: when to stop the iterations. To answer this question is necessary to define $\sigma_{\min}$ as the flooring standard deviation, i.e., the minimum standard deviation from which we decide to stop the process. If $\sigma_{\min}$ is reached, the algorithm stops.

This algorithm will be referred to as SP1 in order to distinguish it from posterior variants. To tune SP1, we introduce SP2. The difference with SP1 is that if the

percentage is lower than Δ, but the standard deviation is still higher than the predefined maximum standard deviation, σ_{max}, this score is also considered as an outlier. Furthermore, if the variation of the standard deviation is higher than Δ or than σ_{max}, and σ_{min} has not been reached yet, we start a new iteration.

The algorithm proposed here is similar to the one introduced in [4]. In this case, we add some threshold values like a maximum and minimum standard deviation and some additional conditions to link these values. We consider that it is necessary to establish some kind of threshold values to better control the pruning, apart from the stopping condition Δ, because our experiments have shown to us that an excessive pruning elicits a decrease in performance.

The iterative algorithms SP1 and SP2 will be compared in this work with other two non-iterative methods, that will be referred as SP3 and SP4. They remove a fixed percentage of scores. SP3 automatically employs the most typical scores and discards a percentage of α most distant scores with respect to the mean. SP4 removes a percentage β of maximum and minimum scores. SP3 and SP4 are similar to the method of frame discarding used in [10].

Our goal is to compare the proposed methods to the baseline. It is worth noting that SP1 and SP2 are iterative score pruning methods, whereas SP3 and SP4 are fixed score pruning methods.

4. Experiments

4.1. Experimental setup

The database used in this work [11] has been recorded by the authors and has been specially designed for speaker recognition. 184 speakers were recorded by fixed-line and/or mobile telephones.

Utterances are processed in 25 ms frames, Hamming windowed and pre-emphasized. The feature set is formed by 12^{th} order Mel-Frequency Cepstral Coefficients (MFCC) and the normalized log energy. Delta and delta-delta parameters are computed to form a 39-dimensional vector for each frame. Cepstral Mean Subtraction (CMS) is also applied.

Left-to-right HMM models with 2 states per phoneme and 1 mixture component per state are obtained for each digit. Client and world models have the same topology. Gaussian Mixture Models (GMM) of 32 mixture components are employed to model spontaneous speech.

The speaker verification is performed in combination with a speech recognizer for connected digits recognition in text-dependent experiments. During enrollment, those utterances catalogued as "no voice" are discarded. This selection ensures a minimum quality for the threshold setting.

In text-dependent experiments, tests are carried out with 8-digit utterances. We apply here verbal information verification [15] as a filter to remove low quality utterances. The speech recognizer discards those digits with a low probability and selects utterances which have exactly 8 digits.

Our text-dependent experiments have been carried out with digits, using speakers with a minimum of 5 sessions. It yields 100 clients. We use 4 sessions for enrollment and the rest of sessions to perform client tests. Speakers with more than one session and less than 5 sessions were used as impostors. 8-digit utterances are employed for enrollment and for testing. The total number of training utterances per speaker goes from 4 to 32. The exact number depends on the number of utterances discarded by the speech recognizer.

In text-independent experiments, one minute long spontaneous speech utterances are used to train and to test the model. The number of training sessions is the same as in the text-dependent case and we use one utterance per session. The choice of impostor data is the same as in the text-dependent case.

It is important to note that fixed-line and mobile telephone sessions are used indistinctly to train or test. This factor increases the error rate.

4.2. Verification results

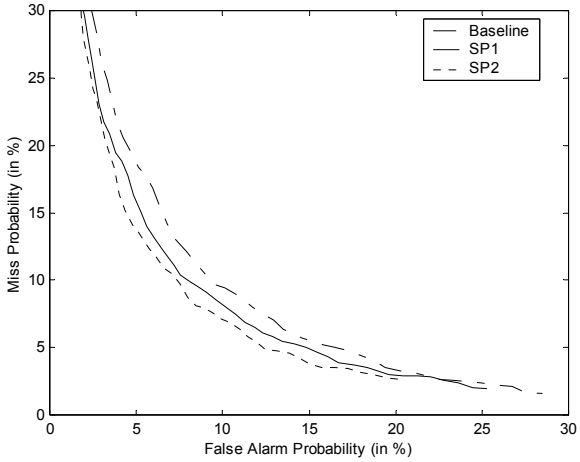

Figure 1: DET curves for iterative methods in text-dependent speaker verification with 100 clients.

Figure 1 shows the DET curves for baseline, SP1 and SP2 speaker-dependent threshold estimation methods. As we can see, SP2 performs better than baseline and SP1. Both score pruning methods have an EER lower than baseline. It remarks the importance of pruning the outliers.

EER (%)		TD (digits)	TI (free speech)
Baseline	-	9.6	20.3
SP1	Iterative	9.0	17.6
SP2	Iterative	8.3	16.9
SP3	Non-iterative	10.3	-
SP4	Non-iterative	10.1	-

Table 1: EER for text-dependent and text-independent experiments with baseline and score pruning methods.

Table 1 shows EERs for text-dependent and text-independent experiments. The error rates for SP3 and SP4 are not present because there are only 4 client scores for text-independent experiments.

As we can see from the table, the iterative score pruning methods have lower error rates than non-iterative ones. Even

more, non-iterative score pruning performs worse than the baseline. The percentage which gives the best results for non-iterative methods discards 15-20% of scores. These methods, based on [2], have a higher error that the baseline in our experiments, because they remove scores with a fixed percentage and they probably remove significant scores, and not only outliers. This leads to the loss of data and consequently increases the error in estimations.

SP2 is the method with the lowest EER and considerably reduces the baseline error. SP1 also reduces the error with respect to the baseline. This is a common feature for both text-dependent and text-independent experiments.

We have also carried out experiments with threshold estimation methods described in (1), (2) and (3) for text-dependent cases. They perform slightly better than our baseline threshold estimation method based on data from clients only, although not all of them perform better if we apply score pruning techniques to the baseline of our method, and what is more critical, they need data from impostors. The method described in (3), which uses mean estimation from clients and impostors - but not standard deviation or variance, has become the method with the lowest EER.

5. Conclusions

The automatic estimation of speaker dependent thresholds has revealed as a key factor in speaker verification enrollment. Threshold estimation methods deal mainly with the sparseness of data and the difficulty of obtaining data from impostors in real-time applications. These methods are currently a linear combination of the estimation of means and variances from clients and/or impostor scores. When we have only a few utterances to create the model, the right estimation of means and variances from client scores becomes a real challenge.

In this paper, we have proposed a new speaker-dependent threshold estimation method and several score pruning techniques to alleviate the problem of the selection of impostor data, the low number of utterances and the presence of outliers. Experiments from our database with 184 speakers have shown a reduction in error rates with score pruning techniques. Furthermore, lower error rates have been obtained for iterative score pruning methods, whereas non-iterative methods perform worse than the baseline.

Future work will consist in applying score pruning techniques to existing threshold estimation methods that use data from impostors.

6. References

1. Furui, S., "Cepstral Analysis for Automatic Speaker Verification", *IEEE Trans. Speech and Audio Proc*., vol. 29(2): 254-272, 1981.
2. Pierrot, J.B., Lindberg, J., Koolwaaij, J., Hutter, H.P., Genoud, D., Blomberg, and M., Bimbot, F., "A Comparison of A Priori Threshold Setting Procedures for Speaker Verification in the CAVE Project", *Proc. ICASSP'98,* pp. 125-128.
3. Lindberg, J., Koolwaaij, J., Hutter, H.P., Genoud, D., Pierrot, J.B., Blomberg, M., and Bimbot, F., "Techniques for A Priori Decision Threshold Estimation in Speaker Verification", *Proc. RLA2C, Avignon 1998,* pp. 89-92.
4. Chen, K., "Towards Better Making a Decision in Speaker Verification", *Pattern Recognition* 36 (2), pp. 329-346, 2003.
5. Mirghafori, N. and Heck, L., "An Adaptive Speaker Verification System with Speaker Dependent A Priori Decision Thresholds", *Proc. ICSLP'02,* pp. 589-592.
6. Gravier, G. and Chollet, G., "Comparison of Normalization Techniques for Speaker Verification", *Proc. RLA2C, Avignon,* 1998, pp. 97-100.
7. Reynolds, D.A, "Comparison of Background Normalization Methods for Text-Independent Speaker Verification", *Proc. Eurospeech'97,* pp. 963-966.
8. Zhang, W.D., Yiu, K.K., Mak, M.W., Li, C.K., and He, M.X., "A Priori Threshold Determination for Phrase-Prompted Speaker Verification", *Proc. Eurospeech'99,* pp. 1203-1206.
9. Surendran, A.C. and Lee, C.H., "A Priori Threshold Selection for Fixed Vocabulary Speaker Verification Systems", *Proc. ICSLP'00,* vol. II, pp.246-249.
10. Bimbot, F. and Genoud, D., "Likelihood Ratio Adjustment for the Compensation of Model Mismatch in Speaker Verification", *Proc. Eurospeech'97,* pp. 1387-1390.
11. Saeta, J.R. and Hernando, J., "Automatic Estimation of A Priori Speaker Dependent Thresholds in Speaker Verification", Proc. 4th International Conference in Audio- and Video-based Biometric Person Authentication (AVBPA), ed. Springer-Verlag, pp. 70-77, 2003.
12. Besacier, L. and Bonastre, J.F., "Frame Pruning for Speaker Recognition", *Proc. ICSLP'98,* pp. 765-768.
13. Besacier, L. and Bonastre, J.F., "Time and Frequency Pruning for Speaker Identification", *Proc. RLA2C, Avignon 1998,* pp. 106-109.
14. Besacier, L. and Bonastre, J.F., "Frame Pruning for Automatic Speaker Identification", *Proc. Eusipco'98,* pp. vol I, pp. 367-370.
15. Li, Q., Juang, B.H., Zhou, Q., and Lee, C.H., "Verbal Information Verification", *Proc. Eurospeech'97,* pp. 839-842.