New Speaker-Dependent Threshold Estimation Method in Speaker Verification based on Weighting Scores

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

[1] Biometric Technologies, S.L..
08007 Barcelona, Spain,
j.rodriguez@biometco.com
[2] TALP Research Center, Universitat Politecnica de Catalunya,
Barcelona, Spain
javier@talp.upc.es

Abstract. Threshold estimation methods mainly deal with the scarcity of data and the difficulty of obtaining data from impostors in real applications. In this context, potential outliers, i.e., those client scores which are distant with respect to mean, could lead to wrong mean and variance client estimations, which are commonly used by somel threshold estimation methods. To alleviate this problem, some efficient threshold estimation methods based on weighting scores are proposed here. Before estimating the threshold, the set of client scores is totally or partially weighted, improving subsequent estimations. The weighting factor is obtained from a non-linear function that distributes scores according to their distance to the estimated mean. Text-dependent experiments have been carried out by using a telephonic multi-session database in Spanish. The database has been recorded by the authors and has 184 speakers.

1 Introduction

Speaker verification (SV) is a way to ensure that a person is who (s)he claims to be. In SV, the decision is normally based on the Log-Likelihood Ratio (LLR), which is given by the test utterance X, the speaker model Y and the non-speaker model $\overline{Y}$, as follows:

$$LLR(X) = Log\left(\frac{P(X \mid Y)}{P(X \mid \overline{Y})}\right) \quad \textbf{(1)}$$

An utterance is compared to the speaker model and it is considered as belonging to the speaker if the LLR surpasses a predefined threshold and rejected if not.

In order to compare two systems, it is common to use the Equal Error Rate (EER), obtained when the False Acceptance Rate (FAR) and the False Rejection Rate (FRR) are equal. However, in real applications, a specific FAR or FRR is usually required. In this case, it is necessary to tune the speaker thresholds to achieve the desired rates.

In a typical SV application, the user enrolls the system by pronouncing some utterances in order to estimate a speaker model. The enrollment procedure is one of the most critical stages of a SV process. At the same time, it becomes essential to carry

out a successful training process to obtain a good performance. The importance and sensitiveness of the process force us to pay special attention on it. Consequently, it is necessary to protect the enrollment procedure by giving the user some security mechanisms, like extra passwords or by providing a limited physical access.

In development tasks, the threshold is usually set a posteriori. However, in real applications, the threshold must be set a priori. Furthermore, a speaker-dependent threshold should be used because it better reflects speaker peculiarities and intra-speaker variability than a speaker-independent threshold. The speaker dependent threshold estimation method is a linear combination of mean, variance or standard deviation from clients and/or impostors.

Human-machine interaction can elicit some unexpected errors during training due to background noises, distortions or strange articulatory effects. An unknown channel aggravates the problem [1]. Furthermore, the more training data available, the more robust model can be estimated. However, in real applications, one can normally afford one or two enrollment sessions only. In this context, the impact of those utterances affected by adverse conditions becomes more important in such cases where a great amount of data is not available. Score pruning (SP) [2, 3] 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. The main problem it that in a few cases the elimination of certain scores can produce unexpected errors in mean or variance estimation.

In this paper, new threshold estimation methods based on weighting the scores to reduce the influence of the non-representative ones are introduced. The methods use a sigmoid function to weight the scores according to the distance from the scores to the estimated scores mean.

A theoretical view of the state-of-the-art is reported on the next section. New proposals are developed in section 3. The experimental setup and the evaluation with empirical results are described in section 4, followed by conclusions in section 5.

2 Theoretical Approach

In real speaker verification applications, the speaker dependent thresholds should be estimated a priori, using the speech collected during the speaker models training. Besides, the client utterances must be used to train the model and also to estimate the threshold because data is scarce. It is not possible to use different utterances for both stages. Finally, the threshold should be speaker dependent to include speaker peculiarities.

There have been several approaches to automatically estimate a priori speaker dependent thresholds. Some 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 to estimate acoustic parameters statistics. These are some of them [4, 5, 6]:

$$\Theta = \alpha \ (\mu_I - \sigma_I) + \beta \tag{2}$$

$$\Theta = \mu_I + \alpha \ \sigma_I^2 \tag{3}$$

$$\Theta = \alpha \ \mu_I + (1 - \alpha) \ \mu_C \tag{4}$$

$$\Theta = \Theta_{SI} + \alpha \ (\mu_C - \mu_I) \tag{5}$$

$$\Theta = \alpha \ (\mu_I + \beta \ \sigma_I) + (1 - \alpha)\mu_C \tag{6}$$

where μ_C is the client scores mean, μ_I is the impostor scores mean, σ_I is the standard deviation from impostors, Θ_{SI} is the speaker independent threshold, and α and β are constants which have to be optimized from a pool of speakers. Equation (5) is considered as a fine adjustment of a speaker independent threshold.

Another approach introduced by the authors in [7] uses only data from clients:

$$\Theta = \mu_C - \alpha \ \sigma_C \tag{7}$$

where μ_C is the client scores mean, σ_C is the standard deviation from clients and α is a constant empirically determined. Equation (7) is very similar to (3), but uses standard deviation instead of variance and the client mean instead of impostors mean.

Other approaches to speaker dependent threshold estimation are based on a normalization of client scores (S_M) by mean (μ_I) and standard deviation (σ_I) from impostor scores [8]. This approach is based on Znorm [9]:

$$S_{M,\,norm} = \frac{S_M - \mu_I}{\sigma_I} \tag{8}$$

It should also be mentioned another threshold normalization technique such as Hnorm [10], which makes use of a handset-dependent normalization.

Some other methods are based on FAR and FRR curves [11]. Speaker utterances used to train the model are also employed to obtain 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 [12] based on the difficulty of obtaining impostor utterances which fit the client model, especially in phrase-prompted cases. In these cases, it is difficult to secure the whole phrase from impostors. The solution is to use the distribution of the 'units' of the phrase or utterance rather than the whole phrase. The units are obtained from other speakers or different databases.

On the other hand, it is worth noting that there are other methods which use different estimators for mean and variance. With the selection of a high percentage of frames and not all of them, those frames which are out of range of typical frame likelihood values are removed. In [13], two of these methods can be observed, classified according to the percentage of used frames. Instead of employing all frames, one of the estimators uses 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.

3 New Speaker Dependent Threshold Estimation Methods

In this paper, a new threshold estimation method that weights the scores according to the distance d_n from the score to the mean is introduced. It is considered that a score which is far from the estimated mean comes from a non-representative utterance of the speaker. The weighting factor w_n is a parameter of a sigmoid function and it is used here because it distributes the scores in a non-linear way according to their proximity to the estimated mean. The expression of w_n is:

$$w_n = \frac{1}{1 + e^{-C d_n}} \tag{9}$$

where w_n is the weight for the utterance m, d_n is the distance from the score to the mean and C is a constant empirically determined in our case.

The distance d_n is defined as:

$$d_n = \left| s_n - \mu_s \right| \tag{10}$$

where s_n are the scores and μ_s is the estimated scores mean.

The C constant defines the shape of the sigmoid function and it is used to tune the weight for the sigmoid function defined in Equation (9). A positive C will provide increasing weights with the distance while a negative C will give decreasing values. A typical sigmoid function, with C=1 is shown in Figure 1:

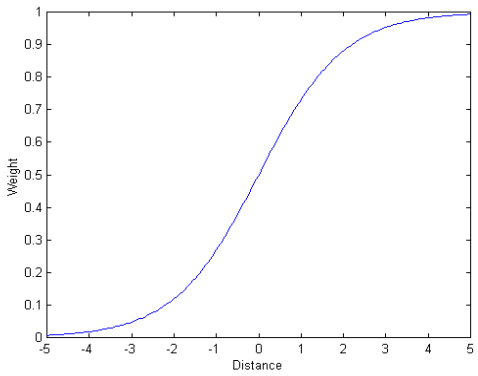

Fig. 1. Sigmoid function

The average score is obtained as follows:

$$s_T = \frac{\sum_{n=1}^{N} w_n s_n}{\sum_{n=1}^{N} w_n} \tag{11}$$

where w_n is the weight for the utterance n defined in (9), s_n are the scores and s_T is the final score.

The standard deviation is also weighted in the same way as the mean. This method is called Total Score Weighting (T-SW).

On the other hand, it is possible to weight only a certain percentage of scores –the least representative- and not all of them. This method is called Partial Score Weighting (P-SW).

4 Experiments

4.1 Database

The database used in this work has been recorded by the authors and has been especially designed for speaker recognition. It includes land-line and mobile telephone sessions. 184 speakers were recorded by phone, 106 male and 78 female. It is a multi-session database in Spanish, with 520 calls from the Public Switched Telephone Network (PSTN) and 328 from mobile telephones. One hundred speakers have at least 5 or more sessions. The average number of sessions per speaker is 4.55. The average time between sessions per speaker is 11.48 days.

Each session includes:

a) 4 different sequences of 8-digit numbers, repeated twice.
b) 2 different sequences of 4-digit numbers, repeated twice.
c) 6 different isolated words.
d) 5 different sentences.
e) 1 minute long read paragraph.
f) 1 minute of spontaneous speech.

4.2 Experimental setup

In our experiments, 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.

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

Clients have a minimum of 5 sessions. It yields 100 clients. They are used 4 sessions for enrollment and the rest of sessions to perform client. Speakers with more than one session and less than 5 sessions are used as impostors. 4- and 8-digit utterances are employed for enrollment and 8-digit for testing. Verbal information verification [14] is applied as a filter to remove low quality utterances. The total number of training utterances per speaker goes from 8 to 48. The exact number depends on the number of utterances discarded by the speech recognizer. During test, the speech recognizer discards those digits with a low probability and selects utterances which have exactly 8 digits.

It is worth noting that land-line and mobile telephone sessions are used indistinctly to train or test. This factor increases the error rate.

4.3 Verification results

In this section, the experiments show the performance of the new threshold estimation methods.

The following table shows a comparison of the EER for threshold estimation methods with client data only, without impostors and for the baseline SDT method defined in Equation (4):

	Baseline	SP	T-SW	P-SW
SDT	5.89	3.21	3.03	3.73

Table 3: Comparison of threshold estimation methods in terms of EER.

As it can be seen in Table 3, the T-SW method performs better than the baseline and even than the SP method. The P-SW performs better than the baseline too, but not than the SP. The results shown here correspond to the weighting of the scores which distance to the mean is bigger than the 10% of the most distant score. It has been found that the minimum EER is secured when everyone of the scores is weighted. It means that the optimal case for the P-SW method is the T-SW method.

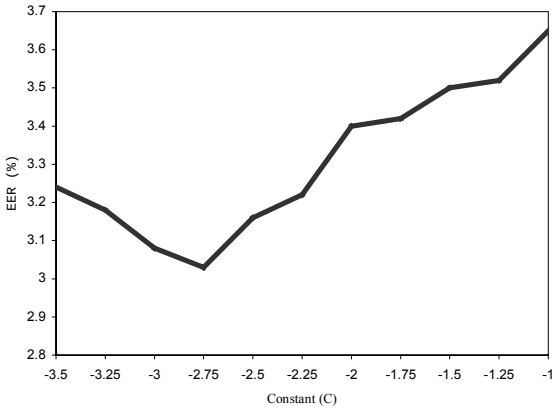

Fig. 2. Evolution of the EER with the variation of C

In Figure 2, it is shown the EER with respect to the C constant. It has been shown that the system performs better for a C = -2.75.

Figure 3 shows the function of the distance and the weight for the best C = -2.75. The weight decreases exponentially with the distance:

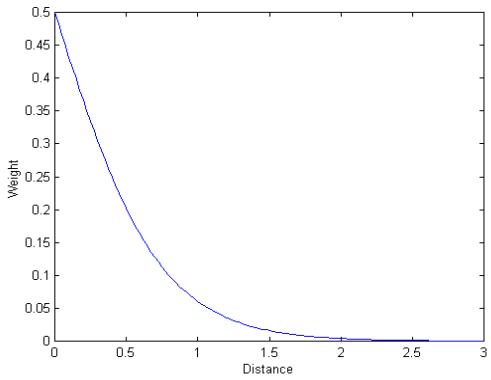

Fig. 3. Variation of the weight (w_n) with respect to the distance (d_n) between the scores and the scores mean

5 Conclusions

Some speaker dependent threshold estimation methods use mean and variance estimations. The presence of outliers elicits some unexpected errors. The SP methods try to

mitigate this problem by removing the outliers, but another problem arises when only a few scores are available. In these cases, the suppression of some scores worsens estimations. For this reason, weighting threshold methods introduced here use the whole set of scores but weighting them in a non-linear way according to the distance to the estimated mean. Weighting threshold estimation methods based on the sigmoid function improve the baseline speaker dependent threshold estimation methods when using data from clients only. The T-SW method is even more effective than the SP ones.

References

1. Kimball, O., Schmidt, M., Gish, H., and Waterman, J., "Speaker Verification with Limited Enrollment Data", Proc. Eurospeech'97, pp. 967-970, 1997.
2. Saeta, J.R. and Hernando, J., "On the Use of Score Pruning in Speaker Verification for Speaker Dependent Threshold Estimation", 2004: A Speaker Odyssey, The Speaker Recognition Workshop", pp. 215-218, 2004.
3. Chen, K., "Towards Better Making a Decision in Speaker Verification", Pattern Recognition, 36, pp. 329-346, 2003.
4. Furui, S., "Cepstral Analysis for Automatic Speaker Verification", IEEE Trans. Speech and Audio Proc., vol. 29(2): 254-272, 1981.
5. 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, 1998.
6. 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.
7. 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.
8. Mirghafori, N. and Heck, L., "An Adaptive Speaker Verification System with Speaker Dependent A Priori Decision Thresholds", Proc. ICSLP'02, pp. 589-592.
9. Gravier, G. and Chollet, G., "Comparison of Normalization Techniques for Speaker Verification", Proc. RLA2C, Avignon, 1998, pp. 97-100.
10. Reynolds, D.A, "Comparison of Background Normalization Methods for Text-Independent Speaker Verification", Proc. Eurospeech'97, pp. 963-966.
11. 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.
12. 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.
13. Bimbot, F. and Genoud, D., "Likelihood Ratio Adjustment for the Compensation of Model Mismatch in Speaker Verification", Proc. Eurospeech'97, pp. 1387-1390.
14. Li, Q., Juang, B.H., Zhou, Q., and Lee, C.H., "Verbal Information Verification", Proc. Eurospeech'97, pp. 839-842.