A MONTE-CARLO METHOD FOR SCORE NORMALIZATION IN AUTOMATIC SPEAKER VERIFICATION USING KULLBACK-LEIBLER DISTANCES

Mathieu BEN, Raphaël BLOUET, Frédéric BIMBOT

IRISA/METISS, Campus Universitaire de Beaulieu
35042 Rennes Cedex, FRANCE, European Union
mben,rblouet,bimbot@irisa.fr

ABSTRACT

In this paper, we propose a new score normalization technique in Automatic Speaker Verification (ASV): the D-Norm. The main advantage of this score normalization is that it does not need any additional speech data nor external speaker population, as opposed to the state-of-the-art approaches. The D-Norm is based on the use of Kullback-Leibler (KL) distances in an ASV context. In a first step, we estimate the KL distances with a Monte-Carlo method and we experimentally show that they are correlated with the verification scores. In a second step, we use this correlation to implement a score normalization procedure, the D-Norm. We analyse its performance and we compare it to that of a conventional normalization, the Z-Norm. The results show that performance of the D-Norm is comparable to that of the Z-Norm. We then conclude about the results we obtain and we discuss the applications of this work.

1 Introduction

The task for an ASV system is to reject or accept a claimed identity by analysing a speaker's voice on a test utterance. Two phases are necessary for the system to be able to accomplish this task. First, it must learn the features of each speaker's voice. This is performed in the training phase during which the ASV system uses training utterances to estimate statistical models of each of the speakers, and a non-speaker model called the world model. In the second phase which is the test phase, the system analyses a test utterance pronounced by the speaker and uses the model corresponding to the claimed identity and the non-speaker (world) model to compute a verification score. Two categories of scores can be distinguished: client scores and impostor scores. In both cases, the score is compared to a decision threshold and the claimed identity is accepted or rejected.

In ASV, the verification score is based on the Log Likelihood Ratio (LLR) between the speaker and the world model distributions, computed on the test utterance. Before making the decision, the system needs to operate a score normalization because of unperfect estimates of the models, which leads to speaker-dependent biases in the client score distribution. In current ASV systems, the most frequently used score normalization techniques are the Z-Norm and the T-Norm (with handset-dependent variants: HZ-Norm and HT-Norm) [1]. These two score normalizations lead to better system performance but they need additional speech data or external speakers to be computed. In some applications, it is sometimes difficult or impossible to find this additional material. In a critical case when no extra data is available, or when it is not possible to use external speaker models, these normalization techniques can not be applied. In this paper, we focus on a new score normalization technique based on Kullback-Leibler distances and which only needs synthetic data.

In section 2, we study the links between scores and the KL distances in an ASV context. We estimate the KL distances with a Monte-Carlo method and we experimentally show that they are strongly correlated with the impostor scores. In section 3, we use these correlations to implement a new score normalization that we call "D-Norm" (for "Distance Normalization"). We first present current score normalization techniques. Then, we expose the D-Norm technique and we experimentally validate it. We analyse its performance and we compare it to that of the Z-Norm. In section 4 we draw a conclusion about the results we obtain and we discuss the applications of this work.

2 The KL distances in ASV

2.1 Definitions

In the field of information theory, the KL distances are defined as relative entropies between two probability density functions. They are not distances in a strict sense because they usually do not verify the symmetry condition. In a speaker recognition context we define two asymmetric distances between a speaker model p_{X_l} and the world model p_W, the "client" distance KL_{X_l} and the "impostor" distance KL_W, as follows:

$$\begin{aligned} KL_{X_l} &= E_{p_{X_l}}[\log \frac{p_{X_l}}{p_W}] \\ KL_W &= E_{p_W}[\log \frac{p_W}{p_{X_l}}] \end{aligned} \quad (1)$$

where $E_{p_\bullet}[\bullet]$ is the expectation under the law $p_\bullet$.

Classically, we also define a symmetrized distance $KL2$ as the sum of the two asymmetric distances:

$$KL2 = KL_{X_l} + KL_W \quad (2)$$

In a theoretical ideal case where the models would be perfectly estimated, KL_{X_l} should give the expected value of the client score, and KL_W should give the opposite of the expected value of the impostor score. In a real case, we only have estimates $\hat{p}_{X_l}$ and $\hat{p}_W$ of the speaker and world models, so we can only obtain the KL distances between these estimated probability functions. Nevertheless, we study these distances to see how they are linked to the scores.

2.2 Estimation of the KL distances with a Monte-Carlo method

Direct computation of the KL distances according to equations (1) and (2) should become impossible with multi-dimensional data and complex statistical laws $\hat{p}_{X_l}$ and $\hat{p}_W$, which is the case in ASV. Therefore, we use a Monte-Carlo method to estimate the KL distances by synthetizing data that follow the statistical laws of the speaker and the world models. According to equations (1), the synthetic data $\tilde{y}_n$ should follow $\hat{p}_{X_l}$ to compute an estimate $\widehat{KL}_{X_l}$ of the "client" distance, and they should follow $\hat{p}_W$ to compute an estimate $\widehat{KL}_W$ of the "impostor" distance. The expected values of the LLR are estimated as the mean on a large number N of synthetic data:

$$\widehat{KL}_{X_l} = \frac{1}{N}\sum_{n=1}^{N} \log \frac{\hat{p}_{X_l}(\tilde{y}_n^{X_l})}{\hat{p}_W(\tilde{y}_n^{X_l})}, \quad \tilde{y}_n^{X_l} \sim \hat{p}_{X_l}$$

$$\widehat{KL}_W = \frac{1}{N}\sum_{n=1}^{N} \log \frac{\hat{p}_W(\tilde{y}_n^{W})}{\hat{p}_{X_l}(\tilde{y}_n^{W})}, \quad \tilde{y}_n^{W} \sim \hat{p}_W$$

$$\widehat{KL2} = \widehat{KL}_{X_l} + \widehat{KL}_W \quad (3)$$

The main state-of-the-art approach consists in using Gaussian Mixture Models (GMMs) to approximate the speakers and world statistical laws [2]. The probability density $p(y)$ of an m-component GMM for p-dimensional acoustic vectors y is defined as:

$$p(y) = \sum_{k=1}^{m} w_k \mathcal{G}_k(y) \quad (4)$$

where $\mathcal{G}_k$ is a Gaussian function with mean vector μ_k and covariance matrix Σ_k, and w_k is the relative weight of function $\mathcal{G}_k$ in the mixture ($\sum_{k=1}^{m} w_k = 1$).

In the Monte-Carlo method, we generate GMM synthetic data in two steps:

1. We randomly draw an index k by respecting the weights $\{w_k\}$. For that, we generate a random variable $k \in \{1,..,m\}$ which follows a multinomial law with parameters $\{w_k\}$.
2. We generate a zero-mean, unit-covariance Gaussian variable $\tilde{y}'_n$ with a Box-Muller algorithm [3] and we scale it to the parameters μ_k and Σ_k of the Gaussian function $\mathcal{G}_k$:

$$\tilde{y}_n = {\Sigma_k}^{\frac{1}{2}} \tilde{y}'_n + \mu_k.$$

In this way, we can generate synthetic data $\tilde{y}_n^{X_l}$ and $\tilde{y}_n^{W}$ that follow, respectively, the GMMs $\hat{p}_{X_l}$ and $\hat{p}_W$ of the speaker and world models. For each synthetic data we calculate the corresponding LLR with the scoring module of the ASV system and, by repeating this procedure N times, we get the estimates of the KL distances given by equations (3).

2.3 Application

2.3.1 Database and ASV system features

For these experiments, we use the database of the NIST'00 evaluation campaign [4]. This database contains phone conversations of american students, with both male and female speakers using electret or carbon handsets. The ASV system that we use is the IRISA/ELISA baseline system for the NIST'01 evaluation [5]. The acoustic analysis of this system gives acoustic vectors with 32 coefficients: the first 16 cepstral coefficients and their respective deltas. The statistical models are based on 128-component, diagonal covariance matrices GMMs and we use gender- and handset-type-dependent world models. The parameters of the speaker models are adapted from the world model using training data, with an EM algorithm and a Maximum A Posteriori (MAP) criterion.

2.3.2 Experimental protocol

For each speaker X_l in the NIST'00 database, we calculate a client mean score $\overline{S}_{client}$ and an impostor mean score $\overline{S}_{imp}$, using all the client accesses and the impostor accesses available against the claimed identity X_l. In the whole NIST'00 database, there is an average of 50 impostor accesses per speaker, and an average of 5 client accesses per speaker. We also calculate, for each speaker X_l, the corresponding Monte-Carlo estimated KL distances. We use 10000 synthetic acoustic vectors for each estimation. This corresponds approximately to the amount of data available in an utterance of about 2 minutes, which is the duration of the training utterances for the NIST'00 evaluation.

We want to highlight possible correlations between the KL distances and the mean scores. One way to observe these correlations is to plot the distributions of points representing the KL distances versus the mean scores for each speaker. In section 2.3.3 we only study these distributions for the symmetrized distance $\widehat{KL2}$ because it shows stronger correlation than $\widehat{KL}_{X_l}$ and $\widehat{KL}_W$.

2.3.3 Practical observations

We present here the results obtained with the female speakers using an electret handset in the NIST'00 database, which correspond to a population of 506 speakers. Figure 1 presents the distributions of points ($\overline{S}_{client}$, $\widehat{KL2}$), noted with (+), and ($\overline{S}_{imp}$, $\widehat{KL2}$), noted with (•), for all these speakers. We also plot with dashes, the linear regression lines of the two distributions. We force

these lines to cross the (0,0) point which is a theoretical point: for a null distance, the speaker and the world model are strictly identical and then, the mean scores are necessarily null. On this figure, we can observe a

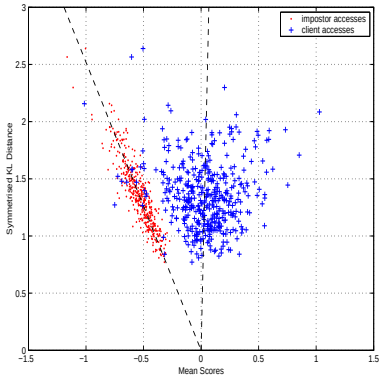

FIG. 1 –. *Distributions of points* $(\overline{S}_{client}, \widehat{KL2})$ *and* $(\overline{S}_{imp}, \widehat{KL2})$

strong correlation between $\widehat{KL2}$ and $\overline{S}_{imp}$: the correlation coefficient between these two quantities is -0.9. Approximately, $\widehat{KL2}$ and $\overline{S}_{imp}$ are linked with a linear relation:

$$\widehat{KL2} \approx -\alpha \overline{S}_{imp} \tag{5}$$

where $-\alpha$ is the slope coefficient of the linear regression line of the ($\bullet$) distribution (on figure 1, we have $-\alpha \approx -2.5$). No correlation appears between $\widehat{KL2}$ and $\overline{S}_{client}$.

These results show that the symmetrized KL distances in ASV can give information about the mean of the impostor scores. In section 3, we use this correlation to implement a new score normalization that we call the "D-Norm" (for "Distance-Normalization").

3 D-Norm: a new score normalization

3.1 Current score normalizations

Currently, the most frequently used score normalizations are the Z-Norm (Zero Normalizaton) and the T-Norm (Test Normalization) which have handset-type-dependent variants, the HZ-Norm and the HT-Norm [1]. These normalizations act on scores to rescale the impostor score distribution of each speaker to a normal distribution (zero mean, unit variance). For each speaker X_l, a mean $\mu_{imp}^{X_l}$ and a standard deviation $\sigma_{imp}^{X_l}$ of the impostor scores are estimated. A score S computed on a test utterance with the claimed identity X_l is then normalized as follows:

$$S_{norm} = \frac{S - \mu_{imp}^{X_l}}{\sigma_{imp}^{X_l}} \tag{6}$$

For the Z-Norm, the parameters $\mu_{imp}^{X_l}$ and $\sigma_{imp}^{X_l}$ for each speaker are estimated "off-line", before the test phase, using an external population of impostors. The T-Norm uses parameters $\mu_{imp}^{X_l}$ and $\sigma_{imp}^{X_l}$ that are estimated by calculating impostor scores on the test utterance itself. This normalization needs additional speaker models and is calculated "on-line", during the test phase, which leads to a large increase of the test duration. For the HZ-Norm and the HT-Norm, the parameters $\mu_{imp}^{X_l}$ and $\sigma_{imp}^{X_l}$ are estimated dependently from the handset type. All these normalizations either need additional speech data or external speaker models to be computed. In the following sections, we present a new score normalization that does not need such additional material.

3.2 Functioning of the D-Norm

We want to use the observed correlation between $\overline{S}_{imp}$ and $\widehat{KL2}$ to normalize scores. We exploit the approximative linear relation (5) between $\widehat{KL2}$ and $\overline{S}_{imp}$ to rescale the impostor score ditribution of each speaker to a (quasi-)constant mean. A verification score S for an access with claimed identity X_l is then normalized as follows:

$$S_{D-Norm} = \frac{S}{\widehat{KL2}_{(X_l)}} \tag{7}$$

where $\widehat{KL2}_{(X_l)}$ is the symmetrized distance corresponding to the speaker X_l. In this way, the mean impostor score for speaker X_l is scaled to a constant value approximately equal to $-1/\alpha$ (see section 2.3.3).

3.3 Experimental validation

Figure 2 should be compared with figure 1. It presents the distributions of the mean scores after being normalized with the D-Norm. This figure shows that the

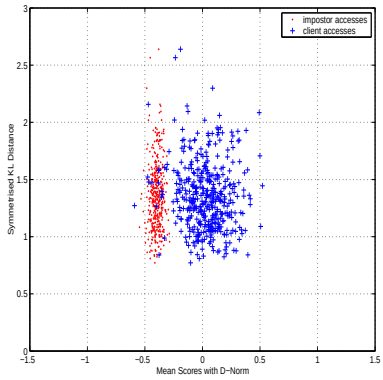

FIG. 2 –. *Distributions of points* $(\overline{S}_{client}, \widehat{KL2})$ *and* $(\overline{S}_{imp}, \widehat{KL2})$ *after D-Norm*

D-normalised mean scores are now much less correlated with the symmetrized distances and so, much less

dependent from the speakers. For each speaker, the D-normalised impostor mean score is approximately scaled to $-1/\alpha \approx -0.4$.

The curves on figure 3 show the behaviour of the mean scores distributions, before normalization (top), and after D-Norm (bottom). They show that with D-Norm, the distribution of the impostor mean scores is more concentrated than before normalization. This should lead to a better separation of the two score distributions (client and impostor) and therefore decrease the error rates of the ASV system.

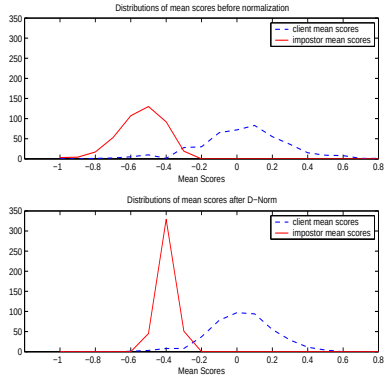

FIG. 3 –. *Distributions of mean scores*

3.4 Performance of the D-Norm

Performance for an ASV system is evaluated with the Detection Error Tradeoff (DET) curve which plots the miss rate versus the false alarm rate of the system.

On figure 4, we present the DET curves of the baseline ASV system of IRISA/ELISA, for the female speakers with an electret handset in the NIST'00 evaluation. The curves without normalization (No norm), with the D-Norm and with the Z-Norm are plotted. In this experiment, performance of the system without normalization, with the D-Norm and with the Z-Norm are equivalent for operating conditions with a low false alarm rate. Around the Equal Error Rate (EER) point the D-Norm and the Z-Norm consistently improve performance of the system without normalization and they nearly perform the same. Only for operating conditions with a low miss rate does the Z-Norm slightly outperform the D-norm.

4 Conclusion

A new approach for score normalization in ASV has been proposed in this paper. The D-Norm is based on the use of the Kullback-Leibler distances between the speaker model and the world model. These distances are estimated with a Monte-Carlo method so the D-Norm does not need any additional speech data or speaker models. This leads to an easier and faster normalization procedure as compared to those of the state of the art

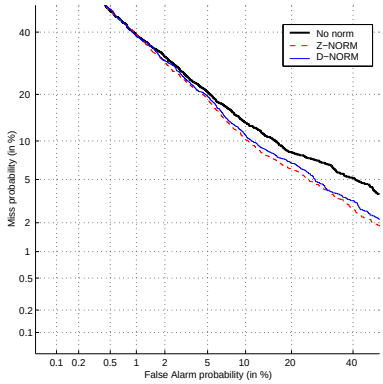

FIG. 4 –. *DET curves. Performance of the IRISA/ELISA baseline ASV system, without normalization (No norm), with the D-Norm, and with the Z-Norm*

(Z-Norm, T-Norm). The experiments we have presented show that the D-Norm performs comparably to the Z-Norm and could advantageously replace it in some specific cases when no additional data is available. In a future work, the results we obtained must be consolidated on a larger set of experiments and the D-Norm performance should be compared with other normalizations (HZ-Norm, T-Norm, HT-Norm). Furthermore, in cases, not considered in this paper, when additional material is available, we could associate the D-Norm with the other normalizations, which could further improve performance of the ASV system.

5 References

[1] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas. Score normalization for test-independent speaker verification systems. *Digital Signal Processing Vol 10,num 1-3*, 2000.

[2] A. Reynolds, T.F. Quatieri, and R.B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing Vol 10,num 1-3*, 2000.

[3] C.P. Roberts and G. Casella. *Monte Carlo Statistical Methods*, chapter 2, page 46. Springer-Verlag New York, Inc, 1999.

[4] National Institute of Standards and Technology. The 2000 nist speaker recognition evaluation. http://www.nist.gov/speech/tests/spk/2000/index.html.

[5] I. Magrin-Chagnolleau, G. Gravier, and R. Blouet. Overview of the 2000-2001 elisa consortium research activities. In *Proceedings of 2001: A Speaker Odyssey-The Speaker Recognition Workshop*, 2001.