IDIAP RESEARCH REPORT

A Unified Framework for Score Normalization Techniques Applied to Text Independent Speaker Verification

Johnny Mariéthoz [1] Samy Bengio [2]

IDIAP–RR 04-62

December 7, 2005

published in
IEEE Signal Processing Letters, Volume 12

1 IDIAP, CP 592, 1920 Martigny, Switzerland, marietho@idiap.ch
2 IDIAP, CP 592, 1920 Martigny, Switzerland, bengio@idiap.ch

IDIAP Research Report 04-62

A Unified Framework for Score Normalization Techniques Applied to Text Independent Speaker Verification

Johnny Mariéthoz Samy Bengio

December 7, 2005

published in
IEEE Signal Processing Letters, Volume 12

Abstract. The purpose of this paper is to unify several of the state-of-the-art score normalization techniques applied to text-independent speaker verification systems. We propose a new framework for this purpose. The two well-known Z- and T-normalization techniques can be easily interpreted in this framework as different ways to estimate score distributions. This is useful as it helps to understand the various assumptions behind these well-known score normalization techniques, and opens the door for yet more complex solutions. Finally, some experiments on the Switchboard database are performed in order to illustrate the validity of the new proposed framework.

Contents

1 Introduction

Text-independent speaker verification systems have evolved through time [3]. The first systems had reasonable performance only in controlled conditions (no noise, same channel, same gender, etc). Over the years, researchers have improved their systems for unmatched conditions, thanks largely to score normalization techniques. In this paper, we propose a unified framework that explains several score normalization techniques used in text-independent speaker verification. Furthermore, an implementation of two of the most common techniques, the so-called T- and Z-normalization [1], is proposed here in this novel framework. While the two approaches are not strictly equivalent, in practice they give similar results. In fact, this new framework can be used to understand the assumptions that are implicit when using T- and Z-normalization. Moreover, it can also used to develop new normalization techniques. The paper is organized as follows. In section 2 we present the classical framework used in speaker verification. In section 3 a new framework is proposed for score normalization. T- and Z-norm implementations in this framework are then given in sections 4.1 and 4.2. Sections 5 and 8 show that the T- and Z-norm using this new framework are equivalent to their classical implementation. Finally we draw some conclusions.

2 Classical Framework used in Speaker Verification

Classical speaker verification models are based on a statistical framework. We are interested in $P(S_i|\mathrm{X})$ the probability that speaker S_i has pronounced sentence X. Using Bayes theorem, this can be expressed as follows:

$$P(S_i|\mathrm{X}) = \frac{p(\mathrm{X}|S_i)P(S_i)}{p(\mathrm{X})}. \tag{1}$$

In order to decide whether or not S_i has pronounced X, we compare $P(S_i|\mathrm{X})$ to the probability that any other speaker has pronounced X, denoted $P(\bar{S}_i|\mathrm{X})$. When $P(\bar{S}_i|\mathrm{X})$ is the same for all Clients, which is the assumption made in this paper, we replace it by a speaker independent model $P(\Omega|\mathrm{X})$ where Ω represents the *World* of all the speakers. The decision rule is then:

$$\text{if } P(S_i|\mathrm{X}) > P(\Omega|\mathrm{X}) \text{ then X was uttered by } S_i. \tag{2}$$

Using equation (1), inequality (2) can be rewritten as:

$$\frac{p(\mathrm{X}|S_i)}{p(\mathrm{X}|\Omega)} > \frac{P(\Omega)}{P(S_i)} = \delta_i \tag{3}$$

where the ratio of the prior probabilities is usually replaced by a threshold δ_i since it does not depend on X and is furthermore usually common for all speakers (hence δ). Taking the logarithm of (3) leads to the *log likelihood ratio* (LLR):

$$\mathrm{llr}_i = \log \frac{p(\mathrm{X}|S_i)}{p(\mathrm{X}|\Omega)} > \log \delta_i = \Delta_i \approx \Delta\ . \tag{4}$$

3 Unified Framework for Score Normalization

Most state-of-the-art text-independent speaker verification systems use linear score normalization functions of the form:

$$\mathrm{llr}_{i_{norm}} = \frac{\mathrm{llr}_i - \mu}{\sigma} > \Delta \tag{5}$$

where μ and σ are respectively the mean and the standard deviation of a normal distribution of LLRs. These parameters are then estimated differently for each type of score normalizations. This paper proposes a unified framework for all kinds of normalization of the form of (5), and also other non-linear functions. We further propose an implementation for the two well-known T- and Z-normalization techniques.

We have seen that in text-independent speaker verification we are interested in the probability that a speaker S_i has pronounced a sentence X. Let us now consider the LLR as an additional random variable, and let us introduce it in the original framework by looking at $P(S_i|\mathrm{X}, \mathrm{llr}_i)$, the probability that a speaker S_i has pronounced a sentence X and obtained an LLR of llr_i. Using the same approach as in section 2, we obtain:

$$P(S_i|\mathrm{llr}_i, \mathrm{X}) > P(\Omega|\mathrm{llr}_i, \mathrm{X}). \tag{6}$$

Using inequality (6) and the Bayes theorem, it can then be rewritten as:

$$p(\mathrm{llr}_i, \mathrm{X}|S_i)\frac{P(S_i)}{p(\mathrm{llr}_i, \mathrm{X})} > p(\mathrm{llr}_i, \mathrm{X}|\Omega)\frac{P(\Omega)}{p(\mathrm{llr}_i, \mathrm{X})} \; . \tag{7}$$

Applying some simplifications to inequality (7) yields:

$$\frac{p(\mathrm{llr}_i, \mathrm{X}|S_i)}{p(\mathrm{llr}_i, \mathrm{X}|\Omega)} > \frac{P(\Omega)}{P(S_i)} \; . \tag{8}$$

Using inequality (8) and the conditional law of probabilities gives:

$$\frac{p(\mathrm{llr}_i|\mathrm{X}, S_i)}{p(\mathrm{llr}_i|\mathrm{X}, \Omega)}\frac{p(\mathrm{X}|S_i)}{p(\mathrm{X}|\Omega)} > \frac{P(\Omega)}{P(S_i)} \; . \tag{9}$$

Taking the logarithm of inequality (9), we finally obtain:

$$\mathrm{llr}'_i = \log\frac{p(\mathrm{llr}_i|\mathrm{X}, S_i)}{p(\mathrm{llr}_i|\mathrm{X}, \Omega)} + \mathrm{llr}_i > \log\frac{P(\Omega)}{P(S_i)} \approx \Delta \; . \tag{10}$$

Comparing equation (10) of this new framework with the original equation (4), we can see that a new term appears. It is the log of the ratio of two likelihoods estimated by two score distributions. The numerator represents the distribution of LLRs for a given access X and for client S_i. The denominator represents the distribution of LLRs for a given access X and for all impostors Ω. We will see that, depending on how these two distributions are estimated, we can obtain classical score normalization techniques such as T-norm (when estimated on a test access) or Z-norm (when estimated for each client S_i).

4 Relation to Existing Normalization Techniques

4.1 T-norm

The T-norm, as introduced in [1] and [6], estimates μ and σ as the mean and the standard deviation of the log likelihood ratios (LLRs) using models of a subset of impostors, for a particular test access X_0.

$$\mu_N = \frac{1}{N}\sum_n \mathrm{llr}_n(\mathrm{X}_0) \tag{11}$$

$$\sigma_N = \sqrt{\frac{1}{N}\sum_n (\mathrm{llr}_n(\mathrm{X}_0) - \mu_N)^2} \tag{12}$$

where N is the number of impostor models and llr_n is the score for the n^{th} impostor model for the particular access X_0. Using (5) we obtain:

$$\mathrm{llr}_{i_{t-norm}} = \frac{\mathrm{llr}_i - \mu_N}{\sigma_N} > \Delta \; . \tag{13}$$

Let us now show how it is possible to perform T-normalization using our new framework under reasonable assumptions. We also show in the Appendix a comparison of our framework and the T-norm implementation found in the literature.

Given the framework described in section 3, we must define two distributions, which will be here defined as Normal, as follows:

$$\hat{p}(\mathrm{llr}|\mathrm{X}, S_i) = \mathcal{N}(\mathrm{llr}_{\mathrm{S}_i}; \mu_{\mathrm{S}_i}, \sigma_{\mathrm{S}_i}) \tag{14}$$
$$\hat{p}(\mathrm{llr}|\mathrm{X}, \Omega) = \mathcal{N}(\mathrm{llr}_{\mathrm{S}_i}; \mu_{\Omega}, \sigma_{\Omega}) \tag{15}$$

where $\mu_{\mathrm{S}_i}, \sigma_{\mathrm{S}_i}$ are the parameters of the client distribution and $\mu_\Omega, \sigma_\Omega$ are the parameters of the impostor distribution. To obtain the T-norm we make the assumption that the standard deviations are equal: $\sigma_N = \sigma_{\mathrm{S}_i} = \sigma_\Omega$. We thus obtain:

$$\begin{aligned} \log \frac{\hat{p}(\mathrm{llr}|\mathrm{X}, S_i)}{\hat{p}(\mathrm{llr}|\mathrm{X}, \Omega)} &= -\frac{1}{2\sigma_N^2}\left((\mathrm{llr}_{\mathrm{S}_i} - \mu_{\mathrm{S}_i})^2 - (\mathrm{llr}_{\mathrm{S}_i} - \mu_\Omega)^2\right) \\ &\quad - \log \frac{\sqrt{2\pi\sigma_N^2}}{\sqrt{2\pi\sigma_N^2}} \\ &= \frac{\mu_{\mathrm{S}_i} - \mu_\Omega}{\sigma_N^2}\left(\mathrm{llr}_{\mathrm{S}_i} - \frac{\mu_{\mathrm{S}_i} + \mu_\Omega}{2}\right) . \end{aligned} \tag{16}$$

If we now define the means as:

$$\begin{aligned} \mu_{\mathrm{S}_i} &= \mathrm{llr}_{\mathrm{S}_i} \\ \mu_\Omega &= \mu_N \end{aligned} \tag{17}$$

we obtain

$$\mathrm{llr}_{\mathrm{S}_i} + \frac{(\mathrm{llr}_{\mathrm{S}_i} - \mu_N)^2}{2\sigma_N^2} > \Delta . \tag{18}$$

Note that equations (17) and (18) are valid only when $\mathrm{llr}_{\mathrm{S}_i} > \mu_N$. A reasonable thing to do is to reject directly without any normalization a claimed speaker if its obtained LLR is smaller than the average of LLRs over a subset of impostors. The consequence of this on the T-norm equation is to force the threshold Δ in (13) to be positive.

4.2 Z-norm

The basis of Z-norm [1] is to test a speaker model against example impostor utterances and to use the corresponding LLR scores to estimate a speaker specific mean and standard deviation:

$$\mu_J = \frac{1}{J}\sum_j \mathrm{llr}_{\mathrm{S}_i}(\mathrm{X}_j) \tag{19}$$
$$\sigma_J = \sqrt{\frac{1}{J}\sum_j (\mathrm{llr}_{\mathrm{S}_i}(\mathrm{X}_j) - \mu_J)^2} \tag{20}$$

where J is the number of impostor accesses. Using a similar approach to that in section 4.1, the estimate of the two distributions needed for the proposed unified framework becomes:

$$\hat{p}(\mathrm{llr}|\mathrm{X}, S_i) = \mathcal{N}(\mathrm{llr}_{\mathrm{S}_i}; \mu_{\mathrm{S}_i}, \sigma_{\mathrm{S}_i}) \tag{21}$$
$$\hat{p}(\mathrm{llr}|\mathrm{X}, \Omega) = \mathcal{N}(\mathrm{llr}_{\mathrm{S}_i}; \mu_{\Omega}, \sigma_{\Omega}) \tag{22}$$

with, again, the same standard deviation, $\sigma_J = \sigma_{\mathrm{S}_i} = \sigma_\Omega$. If we now define the means as follows:

$$\begin{aligned} \mu_{\mathrm{S}_i} &= \mathrm{llr}_{\mathrm{S}_i} \\ \mu_\Omega &= \mu_J \end{aligned} \tag{23}$$

then using equations (23) and (16) we obtain:

$$\text{llr}_{\text{S}_i} + \frac{(\text{llr}_{\text{S}_i} - \mu_J)^2}{2\sigma_J^2} > \Delta \ . \tag{24}$$

Finally, as explained for the T-norm at the end of section 4.1, we also need to reject a claimed access if ($\text{llr}_{\text{S}_i} < \mu_J$).

5 Experiments

The goal of these experiments is to show that the proposed framework can indeed be used to perform T-norm or Z-norm while obtaining the same performance as the original methods, and, gaining some insight about the underlying assumptions.

5.1 Performance Measure in Speaker Verification

The performance of a speaker verification system is usually represented in terms of false acceptance rate (FAR, the number of false acceptances divided by the number of impostor accesses) and false rejection rate (FRR, the number of false rejections divided by the number of client accesses). A summary of these two values is often given by the half total error rate (HTER, the average of FAR and FRR), or by the equal error rate (EER, the point where FAR is equal to FRR[1]). It is also possible to represent graphically the performance using DET curves [5] which, similarly to ROC curves, show FAR with respect to FRR for various values of the threshold of equation (4), but with a Normal scale transformation. More recently, a new Expected Performance Curve (EPC) has been proposed by [2], which has shown to provide a fairer comparison between models. The procedure optimizes a convex combination of the individual performance measures, $ep = \alpha FAR + (1 - \alpha) FRR$, for various values of $\alpha \in [0, 1]$ on the validation set during the training procedure used for parameter selection, and then plots HTER on the test set as a function of α. In this way, each point on the graph contains its underlying *a priori* threshold selection procedure and is thus comparable to similar points (same α) coming from other models. Thus, this curve can be seen as an *a priori* DET curve.

5.2 Database and Protocol

The comparison was done on a subset of the database that was used for the *NIST 2000 Speaker Recognition Evaluation*, which comes from the Switchboard-2 Phase 1 and 2 Corpus collected by the Linguistic Data Consortium. This data was used as an *evaluation* set while the World model and the *development* data come from previous NIST campaigns. While in the original database two different handsets were used (carbon and electret), in the subset selected for this paper, we only used data from electret handsets. This protocol was first proposed by the ELISA consortium as a reference for the NIST 2001 evaluation. We separated the data into male and female data, in order to create two different World models. The male World model was trained on 137 speakers for a total of 1.5 hours of speech, while the female World model was trained on 218 speakers for a total of 3 hours of speech. After that, the two World models were merged: the new World model has the same mean and variance vectors as the concatenation of the two gender dependent World models and the weights are normalized in order to satisfy the constraint that they should sum to 1. For both development and evaluation Clients, approximately 2 minutes of telephone speech were used to train the models and each test access was less than 1 minute long. The development population consisted of 45 males, with 417 males in the evaluation set. The total number of accesses in the development population was 2441 and 27893 for the evaluation population with a proportion of 10% of true target accesses.

[1]Note that EER is an *a posteriori* measure in the sense that the underlying threshold is necessarily chosen on the test set.

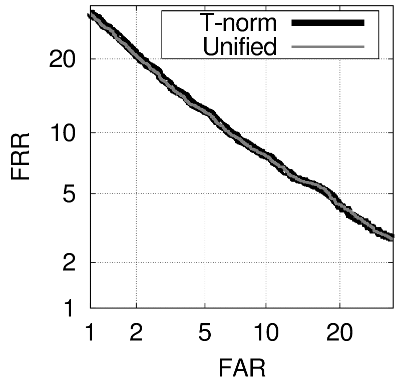

Figure 1: DET curves on the NIST 2000 evaluation set for the T-norm and unified framework T-norm systems.

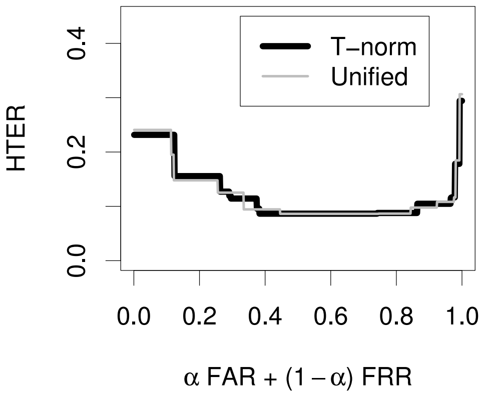

Figure 2: EPC curves on the NIST 2000 evaluation set for the T-norm and unified framework T-norm systems.

5.3 Experimental Results

To verify the validity of our framework and the underlying assumptions, we first compared the standard Z- and T-normalizations and the version derived from the proposed framework. Figure 1 and 3 present the results using DET curves since these curves are often used in the literature. Unfortunately, as explained in section 5.1, DET curves do not take into account the threshold estimation procedure. We thus also present results using EPC in Figure 2 and 4. In both cases the two curves match each other. These results show that the two approaches are equivalent. [2]

6 Conclusion

In this paper, we have presented a new unified framework for text-independent speaker verification score normalization techniques. We have shown that the T- and Z-normalizations can be formalized using this new framework. Theoretical and empirical results show that the implementation found in the literature for T- and Z-norm are equivalent to our implementation. This helps to interpret T- and Z-norm as a way to estimate score distributions using two Normal distributions with the same variance. These normalization techniques have a very simple form in this framework and we can thus

[2]In fact they are perfectly equal if we remove llr_{S_i} in equation (18) and (24).

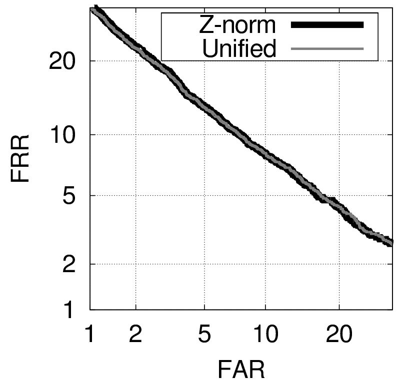

Figure 3: DET curves on the NIST 2000 evaluation set for the Z-norm and unified framework Z-norm systems.

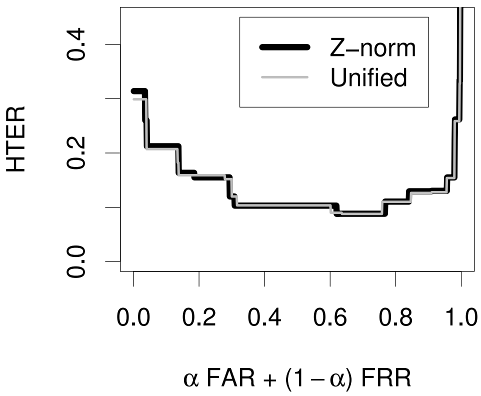

Figure 4: EPC curves on the NIST 2000 evaluation set for the Z-norm and unified framework Z-norm systems.

hope to find an even better estimate of LLR distributions. Indeed, there is no reason to force the LLR distribution to be Normally distributed, as done for the T- and Z-norm. Using our framework it is possible to approximate these distributions using more complex models, such as Mixtures of Gaussians for example. We hope that this framework will be used to propose new score normalization methods and also to improve understanding of this type of algorithms.

7 Acknowledgments

This research has been partially carried out in the framework of the Swiss NCCR project (IM)2. It was also supported in part by the IST Program of the European Community, under the PASCAL Network of Excellence, IST-2002-506778, funded in part by the Swiss OFES. All experiments were performed using the *Torch* package [4]. We would also like to thank Iain McCowan and David Grangier for fruitful discussion and proofreading.

8 Appendix

In this Appendix, we show the difference between the T-norm implementation found in the literature and our implementation using a unified framework. This demonstration can also be applied to Z-normalization.

The new implementation is given by:

$$\mathrm{llr}_{\mathrm{S}_i} + \frac{(\mathrm{llr}_{\mathrm{S}_i} - \mu_N)^2}{2\sigma_N^2} > \Delta \tag{25}$$

The classical method to implement T-norm is equivalent to the second term of the left side of equation (25) since:

$$\begin{aligned}
\frac{(\mathrm{llr}_{\mathrm{S}_i} - \mu_N)^2}{2\sigma_N^2} &> \Theta \\
(\mathrm{llr}_{\mathrm{S}_i} - \mu_N)^2 &> \Theta * 2\sigma_N^2 \\
(\mathrm{llr}_{\mathrm{S}_i} - \mu_N)^2 - 2\Theta * \sigma_N^2 &> 0 \\
\Big[(\mathrm{llr}_{\mathrm{S}_i} - \mu_N - \sqrt{2\Theta} * \sigma_N) & \\
\cdot (\mathrm{llr}_{\mathrm{S}_i} - \mu_N + \sqrt{2\Theta} * \sigma_N)\Big] &> 0
\end{aligned} \tag{26}$$

and if $\mathrm{llr}_{\mathrm{S}_i} > \mu_N$ then we can simplify (26) further into:

$$\begin{aligned}
\mathrm{llr}_{\mathrm{S}_i} - \mu_N - \sqrt{2\Theta} * \sigma_N &> 0 \\
\frac{\mathrm{llr}_{\mathrm{S}_i} - \mu_N}{\sigma_N} &> \sqrt{2\Theta}\,.
\end{aligned} \tag{27}$$

This inequation has a real solution only when $\Theta > 0$, which is true if $\mathrm{llr}_{\mathrm{S}_i} > \mu_N$. This assumption is reasonable: we do not want to accept an access if the LLR on the client model is smaller than the average LLR obtained over a subset of impostors. Given this reasonable assumption we can see the standard T-norm as a simplification of the T-norm using our new unified framework.

References

[1] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas. Score normalization for text-independent speaker verification systems. *Digital Signal Processing*, 10:42–54, January/April/July 2000.

[2] S. Bengio and J. Mariéthoz. The expected performance curve: a new assessment measure for person authentication. In *Proceedings of Odyssey 2004: The Speaker and Language Recognition Workshop*, 2004.

[3] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovsk-Delacrtaz, and D. Reynolds. A tutorial on text-independent speaker verification. *EURASIP Journal on Applied Signal Processing*, 4:430–451, 2004.

[4] R. Collobert, S. Bengio, and J. Mariéthoz. Torch: a modular machine learning software library. Technical Report IDIAP-RR 02-46, IDIAP, 2002.

[5] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The DET curve in assessment of detection task performance. In *Proceedings of Eurospeech'97, Rhodes, Greece*, pages 1895–1898, 1997.

[6] J. Navratil and Ganesh N. Ramaswamy. The awe and mystery of t-norm. In *Proc. of the European Conference on Speech Communication and Technology*, pages 2009–2012, 2003.