COMPARISON OF NORMALIZATION TECHNIQUES FOR SPEAKER VERIFICATION

G. Gravier G. Chollet

ENST/SIG CNRS-URA 820
46 rue Barrault, 75634 Paris Cedex 13, France
{*gravier,chollet*}*@sig.enst.fr*

RÉSUMÉ

De récentes études montrent l'influence du combiné téléphonique sur les performances des systèmes de vérification du locuteur. Une solution à ce problème consiste à sélectionner un modèle de normalisation et/ou à ajuster le seuil de décision en fonction d'information *a priori* sur le combiné. Ce papier compare ces techniques pour de la vérification d'identité indépendante du texte sur Switchboard II. Bien que l'utilisation de connaissance *a priori* pour choisir un modèle de normalisation n'améliore pas la séparabilité, l'utilisation simultanée des deux techniques de normalisation s'avére efficace lorsque les mêmes *a priori* sont utilisés.

ABSTRACT

It has recently been shown that handset mismatch between training and testing has a great influence on speaker verification systems. To overcome the problem, knowledge of the handset as a prior can be used to select a background model and/or to adjust the decision threshold. This paper compares those two techniques as well as their combination. Experiments on text-independent speaker verification with the Switchboard II corpus show that, methods based on the adjustment of the decision threshold improve the stability of the decision boundary. Though the use of the priors to select a background model does not improve the separability of the hypothesis, it is shown that the use of both normalization methods is efficient if the priors are the same.

1 INTRODUCTION

In speaker verification, one has to decide whether some speech X was uttered by a claimed speaker λ or not. Most of the systems are based on similarity-domain normalization [3], which can be seen as H_0 versus H_1 statistical test. H_0 is the hypothesis that the claimant is the right speaker while H_1 is the hypothesis that the claimant is not the right speaker. The decision is taken according to

$$\frac{p(X|\lambda)}{p(X|\overline{\lambda})} \underset{H_1}{\overset{H_0}{\gtrless}} \beta$$

The likelihood under hypothesis H_1, $p(X|\overline{\lambda})$, is not directly computable and must therefore be approximated. It can be approximated using a cohort [10] of speakers, usually selecting speakers who are *close* to the claimant, or by a speaker-independent background model [2] , sometimes refered to as a *world* model.

The other important part of a speaker verification system is the choice of the decision threshold β which can be speaker dependent or not. In text-independent mode, the quality of speaker models may vary from one speaker to another because of the differences in the training material and therefore a speaker-independent decision threshold is not efficient unless some kind of normalization of the likelihood ratios is performed. A common solution to that problem is to base the normalization on the distribution of impostor likelihood ratios. This technique was first used by Furui to set speaker-dependent thresholds [4]. A likelihood ratio normalization scheme, which consists in matching the impostor log-likelihood ratio distribution to a zero mean, unit variance Gaussian distribution is proposed in [7]. This technique is refered to as *znorm*. All those methods can be seen as threshold normalization techniques.

More recently, normalization methods have been further adapted to deal with handset mismatch. This paper compares several of these new approach with respect to test hypothesis separability and threshold stability when switching from a development set to an evaluation set. The paper is organized as follows: in the next section, the experimental protocol is explained

and the speaker verification system is described. Finally, the choice of a background speaker-independent model is studied and different normalization schemes are compared with a fixed background model in section 3.

2 EXPERIMENTS

2.1 Database

Text-independent speaker verification experiments are carried out on the subset of Switchboard II corpus used for the NIST 97 speaker recognition campaign. The complete 97 evaluation data were divided into two sets of 46 male and 46 female speakers each. The first set is used as a development set while the second one is used for evaluations. 40 male and 40 female speakers among the remaining speakers of the complete NIST data are used to build *world* models as described below. Finally all the speakers unused previously are considered as potential impostors to estimate the impostor log-likelihood ratio distribution parameters. The training material consists of two minutes of speech recorded over a single session (and therefore a single handset), and tests are performed for 3 and 30 seconds segments since it is believed that some normalization techniques are more efficient for longer segments. The impostor log-likelihood ratio distributions were computed with 70 impostor trials in every case and, for similar experiments on the development and evaluation sets, the same impostor segments were used. For both development and evaluation sets, there is about 12,000 impostor trials and 1,000 true-speaker trials, without cross-sex trials are performed.

2.2 Speaker verification system

The text-independent speaker verification system used for those experiments is similar to the standard LP-SOSM system described in [5]. It is based on second-order statistical measures on the long term cepstrum [1] and implemented as a likelihood ratio test. Feature vectors consist in 16 cepstral coefficients derived from linear prediction of order 16, plus delta cepstral coefficients and delta log-energy. Performances are measured in terms of detection cost according to $\mathrm{dcf} = 0.1\,\mathrm{pfr} + 0.99\,\mathrm{pfa}$, where pfr and pfa are respectively the false rejection and false alarm probabilities, or in terms of DET curves [8]. The detection cost function is mainly used to study decision boundary stability while the DET curves are a practical way of looking at the hypothesis separability.

2.3 Experiments

The two previous NIST campaigns pointed out the fact that speaker verification systems are very sensitive to handset mismatches between test and training despite the use of channel compensation methods such as cepstral mean subtraction or RASTA filtering. Therefore, two methods have been developed to tackle the problem. Heck proposed to use handset-dependent background models [6] since a segment may score poor with the speaker model and well with the *world* model, not because it is not from the genuine speaker but because it comes from a non-training handset. Considering that handset labels for test and training segments are priors, this can be seen as a similarity-domain normalization which includes the priors. Another technique called *hnorm*, based on a handset-dependent *znorm*, is described in [9]. This is similar to adapting the decision threshold to the speaker according to the test segment handset.

The first experiment studies the benefit of including the priors to select the most appropriate background model. For these experiments, the priors are (1) no cross-sex trials and (2) train and test segment handsets are known. Therefore gender-dependent background models can be used for the first prior and gender-dependent, handset-dependent models can be used for (1) and (2). The second experiment focuses on the threshold stability with respect to the different threshold normalization techniques mentioned previously. Finally, the combination of similarity-domain and threshold normalization methods is studied.

3 RESULTS

3.1 Similarity-domain normalization

In order to include the priors in the selection of a background model, several *world* models were estimated. The first one (**w6**) was built using 60 minutes of speech from 30 male and 30 female speakers of the Switchboard I corpus. The other model parameters are all estimated using 40 minutes of speech from 40 different speakers, using speakers of the Switchboard II corpus (i.e. same corpus as the development and evaluation sets). **w7** is a simple speaker-independent background model, the 40 speakers being selected so that there is as many male as female speakers and as many carbon as electret segments. **wg** refers to a gender-dependent normalization experiment. A female and a male *world* model was build with as many carbon as electret segments. Finally, **wgh** refers to a gender-dependent, handset-dependent normalization. Since these experiments focus on the test hypothesis sepa-

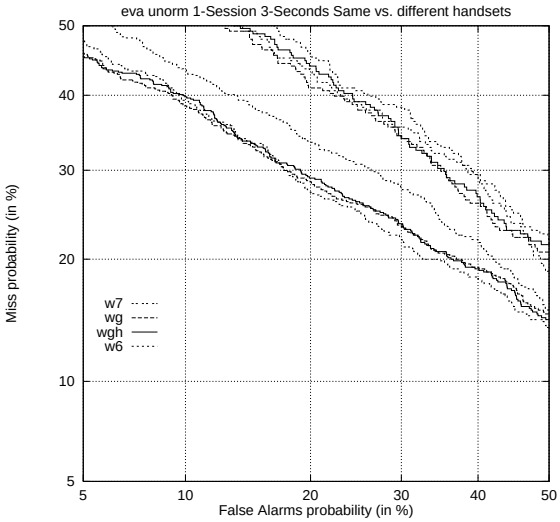

Figure 1: 3 second test

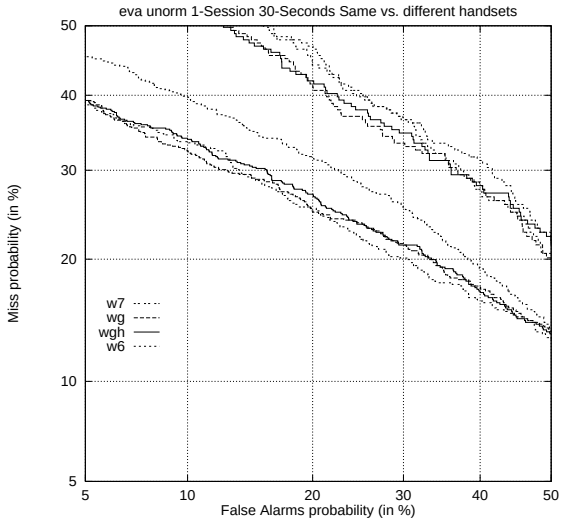

Figure 2: 30 second test

rability, no decision is actually taken and no threshold normalization method is used. Figures 1 and 2 show the DET curve on the evaluation set for 3 and 30 seconds test segments respectively. The DET curves are given separately for training and non-training handsets.

The two figures clearly show that the use of a background model estimated on the same corpus as the test data significantly improves the separability of the test hypothesis. It can be seen that the improvement mainly comes from training handsets since for non-training handsets, the **w6** curves are very close to the others. The use of the priors to select the background model does not significantly improve the separability in this case. For both test segment duration, a gender specific *world* model is slightly better than the general case and performs as well as a gender and handset specific background model. The improvements obtained with world models including the priors on the non-training handsets are more evident on the 30 second test segments where **wg** and **wgh** curves are below **w7**.

3.2 Threshold normalization

For this experiment, the background model used is **w7** and the threshold normalization techniques mentioned above are compared in terms of threshold stability. For each experiment, thresholds are determined on the development set to match the minimum of the detection cost function and hard decision is performed on the evaluation set. Thresholds depend on the test segment length. Figure 3 shows the hard decision and the minimum a-posteriori detection costs, without normalization (*unorm*), *znorm* and *hnorm*. The left bar corresponds to the hard decision detection cost while the right one corresponds to the minimum. For each bar the lowest part indicates the cost of errors due to false alarms.

Threshold normalization methods appears to be more efficient for the 30 second test segments than for the 3 second ones. For the 30 second case, the operating point is close to the optimal one for *znorm* and *hnorm*. One point which is not revealed by the figure is that the comparison of the a-posteriori thresholds obtained on the development set shows that they vary less with the test segment length when using either *znorm* or *hnorm*. For exemple, the threshold was experimentally set to 504 (resp. 4288) for 3 (resp. 30) second tests without normalization and to 1.75 (resp. 1.62) when using *znorm*. Surprisingly, the minimum of the detection cost function is slightly lower without normalization.

3.3 Mixed normalization

Both similarity-domain and threshold normalization are now varied. Figure 4 shows the detection costs for background models **wg** and **wgh**, using either *znorm* or *hnorm*. In the figure, the threshold normalization method is refered to by its first letter.

One interesting point to note is that, for both test durations, the combinations **wg**/*znorm* and **wgh**/*hnorm* gives hard decision scores very close to the minimum a-posteriori. Since no-cross sex impostor trials are done, one can consider the *znorm* as being gender-dependent. This result points out the fact that, when using both similarity-domain and threshold nor-

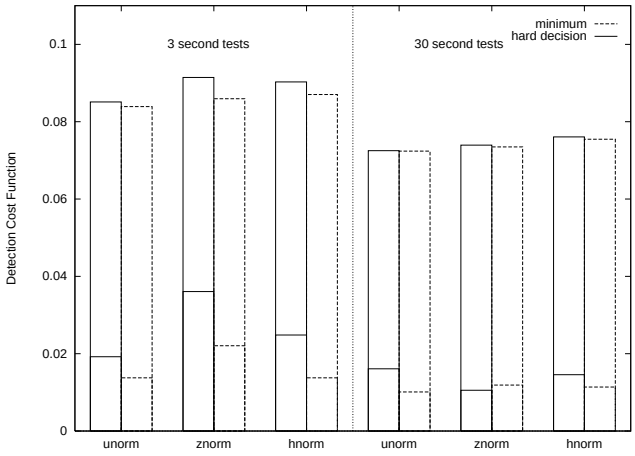

Figure 3: threshold normalization

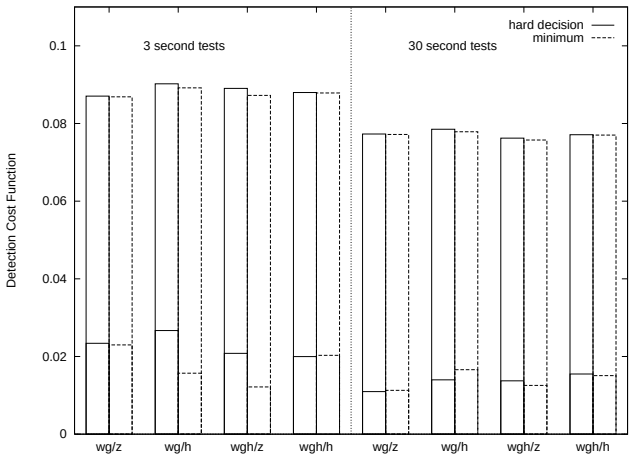

Figure 4: mixed normalization

malizations, it is better to take into account the same priors for each normalization method. Comparing with the previous experiment on **w7**, it can be seen that for the 3 second tests, the composite schemes **wg**/*znorm* and **wgh**/*hnorm* leads to a lower detection cost than **w7** with either *znorm* or *hnorm*, which is not true for the 30 second tests. The selection of a background model according to the priors and the use of the corresponding threshold normalization method is efficient for short test segments, while it is more suitable to only use a threshold normalization method for longer segments.

4 CONCLUSION

The comparison of normalization methods in text-independent speaker verification with second order statistical measures shows that no significant improvement is obtained in terms of test hypothesis separability when using the priors on gender and handset to select the background model. However, the stability of the decision threshold is increased using *znorm* or *hnorm* methods. It has experimentally been shown that the decision boundary is very stable when the same priors are used for the selection of the speaker-independent background model and for the estimation of the impostor log-likelihood ratio distribution.

References

[1] F. Bimbot et al. Second-order statistical measures for text-independent speaker identification. *Speech Communication*, 17:177–192, 1995.

[2] M. J. Carey and E. S. Parris. Speaker verificastion using connected words. In *Proc. Institute of Acoustics*, volume 14, pages 95–100, 1992.

[3] Sadaoki Furui. *An overview of speaker recognition technology*, chapter 2. Kluwer Academic Publishers, 1996.

[4] Sadaoki Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Trans. on ASSP*, 29(2), April 81.

[5] G. Gravier et al. Model dependent spectral representations for speaker recognition. In *EuroSpeech*, 1997.

[6] Larry P. Heck and Mitchel Weintraub. Handset-dependent background models for robust text-independent speaker recognition. In *ICASSP*, 1997.

[7] Kung-Pu Li and Jack E. Porter. Normalizations and selection of speech segments for speaker recognition scoring. In *ICASSP*, pages 595–597, 1988.

[8] A. Martin et al. The det curve in assessment of detection task performance. In *Eurospeech*, volume 4, pages 1895–1898, 1997.

[9] D. A. Reynolds. Comparison of background normalization methods for text-independent speaker verification. In *Eurospeech*, pages 963–966, 1997.

[10] Aaron E. Rosenberg et al. The use of cohort normalized scores for speaker verification. In *International Conference on Speech and Language Processing*, pages 599–602, 1992.