SEGMENTAL NORMALIZATION FOR ROBUST SPEAKER VERIFICATION

Corinne FREDOUILLE, Jean-François BONASTRE, Teva MERLIN

LIA/CERI - Agroparc
339, chemin des Meinajaries BP1228 - 84911 Avignon Cedex 9 (France)
E-Mail : (corinne.fredouille,jean-francois.bonastre,teva.merlin)@lia.univ-avignon.fr

ABSTRACT

For the task of speaker verification, similarity measure normalization methods are relevant to cope with variability problems and with data and/or decision fusion problems.

The aim of this paper is to suggest a new method of normalization which combines classical world model based normalization techniques with ones based on a posteriori probability.

This original method presents the well-known advantages of the a posteriori probability based methods without requiring data and speaker specific processing.

In this paper, the proposed method is experimented in a framework of a temporal-segmental speaker verification system.

The results obtained on a subset of Switchboard-Nist98 database demonstrate the ability of this method to normalize similarity measures (in [0,1] probability domain) without decreasing performances.

1. INTRODUCTION

For the task of speaker verification, similarity measure normalization methods have proved their efficiency through several studies to cope with variability problems induced by the message content, noise and degradation issued from signal recording and transmission device. In the closely related domain of data fusion, normalization methods are also very useful since similarity measures, yielded by different recognizers and/or different temporal segments, have to be merged together.

In speaker verification, two main approaches, mutually exclusive, are suggested. The most frequently used solution consists of normalizing the similarity measure $f(s|X)$ by $f(s|\overline{X})$ where $f(s|\overline{X})$ represents the similarity measure between impostor model $\overline{X}$ and test signal *s*. In this context, $f(s|X)$ is replaced with the ratio $f(s|X)/f(s|\overline{X})$ [1] ([1],[2],[3],[4],[5]). A second approach consists of modelling the recognizer behaviour from a test data set dedicated to this task and of replacing original similarity measure $f(s|X)$ by the MAP[2] estimation defined as $p(X|s) = p(s|X) * p(X) / p(s)$ ([6],[7]). The main advantage of this solution is to propose a bounded score, between 0 and 1, which corresponds to the probability of the studied hypothesis in a specific context. Its main drawback relies on the requirement of a great amount of data in order to take the various disturbances seen above into account.

The aim of this paper is to present a new method of normalization which combines these two techniques in a speaker verification system.

Section 2 details the normalization method and suggests some of its advantages. In *Section 3*, the speaker verification system baseline is described. *Section 4* is dedicated to experiments where the estimation (on a development data set) and the validation (on an evaluation data set) of the normalization functions are detailed. In this section, the potential of the normalization function is also illustrated through speaker verification system performances. Finally, *Section 5* summarises the main results and outlines the potential advantages of the normalization method proposed.

2. NORMALIZATION METHOD

Let $f(s|X)$ and $R_S = f(s|X)/f(s|\overline{X})$ be, respectively, the similarity measure between model *X* and test signal *s* and the similarity measure ratio where $f(s|X)$ is normalized by the similarity measure for a world model, representing the population in general. The normalization method proposed here consists of replacing $f(s|X)$ with the a posteriori probability, denoted as $P(X = X_S|R_S)$ so that the claimant identity is correct given similarity measure ratio R_S; in other words, the a posteriori probability so that R_S is a target score (as opposed to a non target or impostor score). This probability, following the Bayes theorem is defined as:

$$P(X = X_S|R_S) = \frac{P(R_S|X = X_S).P(X = Xs)}{P(R_S|X = Xs).P(X = X_S) + P(R_S|X \neq X_S).P(X \neq X_S)} \quad (1)$$

where $P(R_S|X = X_S)$ (resp. $P(R_S|X \neq X_S)$) is the probability for ratio R_S given the probability density fonction of target scores (resp. impostor scores) estimated a posteriori on a separate development data set and $P(X = X_S)$ (resp. $P(X \neq X_S)$) is the a priori probability for a target score

[1] Different techniques are proposed to estimate $f(s|\overline{X})$: a posteriori probability, cohort model, world model.
[2] Maximum A Posteriori.

(resp. impostor score), which is assumed to be constant for all R_S.

This type of normalization is close to MAP normalization since it proposes bounded scores, dependent on operating conditions of the system (a priori probability) and meaningful since scores are probabilities. Nevertheless, the preliminary world model based-normalization allows to reduce the amount of tests and tuning conditions necessary to the estimation of MAP normalization function.

3. THE SPEAKER VERIFICATION SYSTEM

3.1. Speaker models and similarity measures

The speaker verification system is based on EM-trained (Expectation-Maximization [8]) Gaussian Mixture Models (GMM [9]) to represent acoustical feature vectors of each speaker. Let x be a *p*-dimensional feature vector of speech signal uttered by speaker X_S, the mixture density is defined as:

$$p(x|X_S) = \sum_{i=1}^{M} p_S^i N_S^i(x, \mu_S^i, \Sigma_S^i) \quad (2)$$

where p_S^i and $N_S^i(x,\mu_S^i,\Sigma_S^i)$ are the mixture weights, which satisfy the constraint $\sum_{i=1}^{M} p_S^i = 1$, and the i-th uni-modal gaussian density, summarised by mean vector μ_S^i, and covariance matrix Σ_S^i.

In this experimental context, a 16 gaussian mixture summarised by full covariance matrices is used to estimate speaker and world models. Each speaker model is trained on 2 minute long speech signal, and gender dependent world models on about 50 minute long speech signal each.

3.2. Segmental framework (and acoustic parameterization)

The signal is characterised each 10 ms by a 16 cepstrum coefficients. Cepstral mean substraction is applied to cepstral vectors in order to operate a blind deconvolution.

A frame level likelihood ratio, denoted as $R(y_t|X)$ is computed for each test signal frame y_t.

Then, a segmental likelihood ratio is obtained by computing a geometric mean over *T* frames (*T* frame long segments) as following:

$$R(y_{t+1}\cdots y_{t+T}|X) = \left(\prod_{i=t+1}^{T} R(y_i|X)\right)^{\frac{1}{T}} \quad (3)$$

4. EXPERIMENTS

4.1. Data sets

The method proposed in this paper is experimented on a data set extracted from NIST/NSA 1998 evaluation campaign. This subset is composed of recordings issued from Switchboard database and built from concatenated telephone conversation segments.

Experiments are conducted on three different data subsets defined by the ELISA consortium[3]. These subsets are:

- A recording set for the gender dependent world model training, composed of recordings of 30 second long speech signal uttered by 100 male speakers and 100 female speakers.
- A development data set (denoted as Dev data set) used for the normalization function learning, which is composed of 100 male speakers and 100 female speakers (50 client and 50 impostor speakers for each gender). Each verification test is 30 second long (30s NIST test condition). Finally, test stage includes about 600 target trials and 4400 impostor trials.
- A validation data set (denoted as Eval data set) with same size and same structure as the previous one, but on a different speaker population.

NB: It can be noticed there is no overlapping between these three data sets. The 2 sets, Dev and Eval, are made up of two subsets, one for the speaker model training and another for the test stage. Each speaker model is trained from about 2 minutes of speech signal (2s NIST 98 training condition).

4.2. Likelihood ratio distributions and normalization functions

In this experimental context, the normalization method detailed in *Section 2* is applied at the segmental level. This means the normalization function is estimated from target and non target segmental log likelihood ratio distributions, computed on Dev data set. This normalization function is gender dependent.

NB: in the following subsections, all the figures refer to female speaker population. Although not represented here, similar behaviours are observed with male speaker population.

Segmental likelihood ratio distributions

Figure 1 represents the probability density functions (pdf) of female target and non target segmental log likelihood ratios stemming from verification tests carried out on Dev data set.

[3] The Elisa consortium is composed of European research laboratories, working on a reference plateform for speaker recognition system evaluation.

It can be observed that Target and Non Target pdf means are close with significant standard deviations, which involves a large overlapping between the two distributions.

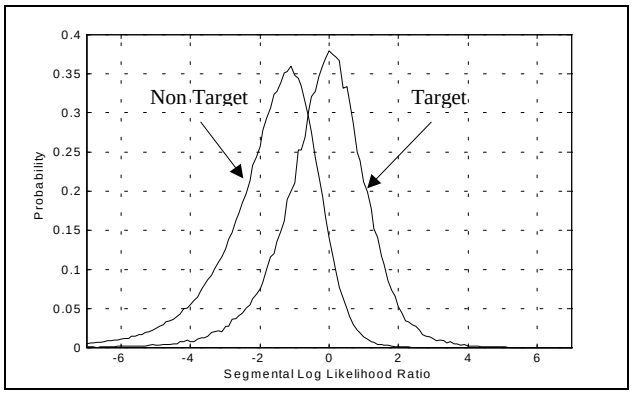

Figure 1: pdf of female target and non target segmental log likelihood ratios obtained on Dev data set.

Normalization function

As explained in *Section 2*, the normalization method is based on a posteriori target and non target likelihood ratio distributions (figure 1 for female speaker population) and a priori probability for target and non target score. The latter are fixed according to expected target and non target test trials. In this context, $P(X = X_S) = 0.1$ and $P(X \neq X_S) = 0.9$ are chosen.

Figure 2 provides the gender dependent normalization functions (here female speaker population) estimated from the target and non target segmental log likelihood ratio distributions obtained on Dev data set (Figure 1).

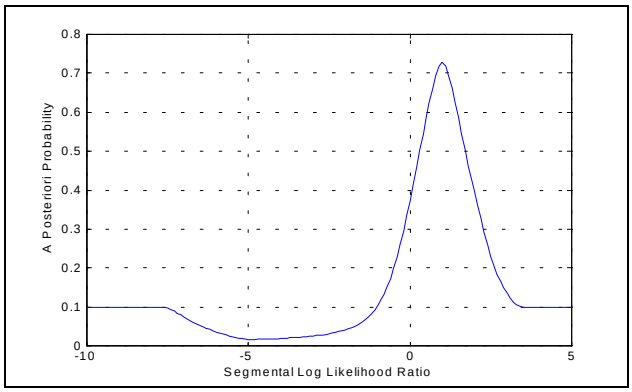

Figure 2: female speaker normalization function estimated from log likelihood ratio pdf obtained on Dev data set.

As expected, this normalization function is defined by three main parts:

- the first one attributes a target representative probability to log likelihood ratios included in [-1;3.5];
- the second one, related to ratios in [-8;-1.1], refers to Non Target log likelihood ratios since they are associated with probability smaller than 0.1 (a priori probability of Target Score);
- the last part, in [-inf;-8] and [3.6;+inf], deals with non informative ratios (e.g. unusual ratio values) set to the a priori target score probability: 0.1.

It is interesting to notice that the first and second parts represent a straightforward way to discriminate Target against Non Target scores.

Normalized segmental likelihood ratio distributions on Dev and Eval data sets

Figure 3 shows the gender dependent (female speaker population) pdf of the normalized target and non target log likelihood ratios once applied the normalization function (figure 2) to Dev data set, which was used to learn it.

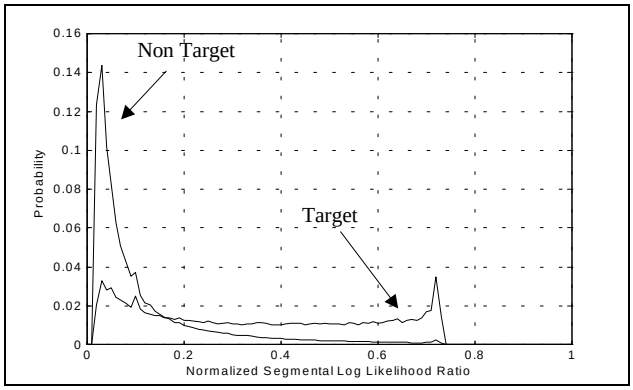

Figure 3: pdf of normalized female target and non target segmental log likelihood ratios obtained on Dev data set.

Different aspects have to be pointed out:

- The Non Target log likelihood ratios are mainly concentrated in the interval [0;0.1] as expected according to the normalization function.
- The Target log likelihood ratio pdf is composed of two main parts: the first one in the interval]0.1;0.75] (0.75 refers to normalization function optimum) where segments are correctly labelled as target segments with a certain confidence according to the probability associated with. The second part is formed of ratios with probability in [0;0.1] which should be labelled as Non Target ratios and may correspond to error-prone segments due to a lack of speaker specific information.

These normalization functions have to be applied to target and non target log likelihood ratios of a separate data set (Eval data set) in order to confirm their behaviour and to be validated.

The log likelihood ratio pdf obtained before and after normalization (not illustrated here) on Eval data set present similar characteristics to those of Dev data set.

This shows the normalization method proposed is not disturbing when it is applied on speakers and data different from those used to learn it.

4.3. Speaker Verification Results

To motivate the use of the normalization method presented here, this latter has been tested through the segmental speaker verification system.

Figure 4 refers to Det curves [10] obtained, on one hand, by using classical world model based normalization (denoted as Classical Norm. on the figure) and, on the other hand, by integrating the log likelihood ratio gender dependent normalization method (denoted as Ratio Norm. on the figure). Tests have been performed on Eval data set.

It can be observed a very slight difference in performance between the two kinds of normalization.

This shows that the original normalization method presented here can be easily integrated to a speaker verification system without decreasing performances. The main advantages to this point are:

- the normalization function tuned and applied on two entirely separate data sets (Dev data set to learn it and Eval data set for testing) does not disturb speaker verification system performances.
- the normalized scores resulting from verification tests are bounded (distributed in interval [0,1]) and meaningful since they correspond to a posteriori probabilities. This may facilitate decision threshold tuning for speaker verification system or score fusion in the framework of segmental and/or multi recognizer system [11][12].

5. CONCLUSION

We suggest a new similarity measure normalization for speaker verification. This normalization method combines classical world model based normalization methods with a posteriori probability based ones.

This new method allows the well known advantages of a posteriori probability based methods without requiring data and speaker specific processing.

The results obtained demonstrate the ability of this method to normalize similarity measures (in [0,1] probability domain) without decreasing performances.

Further studies will have to demonstrate the potentiality of this original method for tuning the decision threshold for speaker verification or for the open-problem of data fusion in the framework of segmental and/or multi recognizer speaker recognition system.

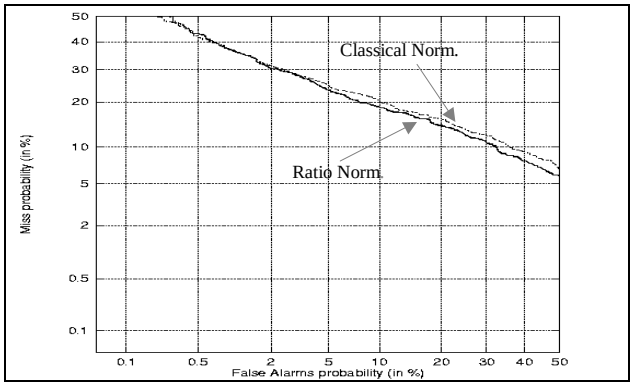

Figure 4: Det curves obtained on Eval data set by using Classical world model based normalization (Classical Norm.) and by integrating segmental log likelihood ratio dependent gender normalization (Ratio Norm.).

6. REFERENCES

[1] Higgins A., Bahler L., Porter J., "Speaker verification using randomized phrase prompting", *Digital Signal Processing*, 1991, Vol. 1, pages 89-106.

[2] Rosenberg A. E., "The use of cohort normalized scores for speaker verification", *Proc. International Conference on Speech and Language Processing*, 1992, pages 599-602.

[3] Carey M. J., Parris E. S., "Speaker verification using connected words", *Proc. Institute of Acoustics*, 1992, Vol. 14, pages 95-100.

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

[5] Gravier G., Chollet G., "Comparison of normalization techniques for speaker verification", *Proc. Speaker Recognition and its Commercial and Forensic Applications (RLA2C)*, April, 1998, pages 97-100.

[6] Matsui T., Furui S., "Likelihood normalization for speaker 7verification using a phoneme- and speaker-independent model", *Speech Communication*, August, 1995, pages 109-116.

[7] Tran D., Minh D., Wagner M., Van Le T., "A proposed decision rule for speaker identification based on a posteriori probability", *Proc. Speaker Recognition and its Commercial and Forensic Applications (RLA2C)*, April, 1998, pages 85-88.

[8] Dempster A., Larid N., Rubin D., "Maximum likelihood from incomplete data via the EM algorithm", *J. Roy. Stat. Soc.*, 1977, Vol. 39, pages 1-38.

[9] Reynolds D. A., "Speaker identification and verification using gaussian mixture speaker models", *Speech Communication*, August, 1995, pages 91-108.

[10] Martin A., Doddington G., Kamm T., Ordowski M., Przybocki M., "The DET curve in assessment of detection task performance", *Proc. Eurospeech*, 1997.

[11] Besacier L., Bonastre J. F., "Subband architecture for automatic speaker recognition on partially corrupted speech ", *Proc. COST 254 Workshop on Emerging Techniques for Communication Terminals*, July, 1997.

[12] Besacier L., Bonastre J. F., "Frame pruning for speaker recognition", *Proc. International Conference on Acoustics Speech and Signal Processing*, May, 1998.