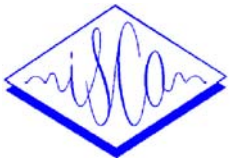

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

SIMILARITY NORMALIZATION METHOD BASED ON WORLD MODEL AND A POSTERIORI PROBABILITY FOR 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 issues. The aim of this paper is to suggest a new normalization method, which combines classical world model-based normalization techniques with a posteriori probability-based ones. This method presents the well-known advantages of the a posteriori probability-based methods without requiring data and speaker specific processing. Here, it is experimented through a temporal-segmental, multi-recognizer speaker verification system. The results obtained on a subset of the Switchboard-Nist98 database demonstrate the ability of this method to normalize similarity measures (in probability domain) without decreasing performance. The second advantage of this method is borne out by the performance of the multi-recognizer system, which reveals that this normalization is able to make the fusion step easier without requiring any weighting function even if individual recognizer performance is dissimilar.

1. INTRODUCTION

For the task of speaker verification, several similarity measure normalizations have been proposed to cope with variability problems induced by message content, by noise and degradation due to signal recordings and transmission channels, and by mismatch across training and testing conditions (telephone lines, handset types…).

In a multi-recognizer system, normalization methods are also relevant since similarity measures, yielded by different classifiers and/or different temporal segments, have to be merged together.

Two mutually exclusive approaches are generally used. The frequently used solution consists in normalizing the similarity measure, denoted as $f(s|X)$, between claimed speaker model X and test signal s with the similarity measure, denoted as $f(s|\overline{X})$ [1], between anti-speaker model $\overline{X}$ and test signal s.

In this context, $f(s|X)$ is replaced with ratio $f(s|X)/f(s|\overline{X})$ ([1][2][3][4][5][6]).

The second approach consists in characterizing the system's behavior from a test data set dedicated to this task and in replacing original similarity measure $f(s|X)$ with the MAP[2] estimation defined as $p(X|s) = p(s|X) * p(X)/p(s)$ ([7],[8]). Accordingly, this solution presents the advantage of "shifting" similarity measures into the probability domain where bounded scores refer to the probability of the studied hypothesis in a specific context. Nevertheless, the MAP estimation requires a great amount of data in order to take the various disturbances (mentioned above) into account.

This paper aims to present an original normalization method which combines these two techniques for speaker verification.

This normalization is especially appropriate to a multi-recognizer system since it is able to take the intrinsic recognizer performance into account and is experimented in this way.

Section 2 details the normalization method and outlines some of its advantages. In *Section 3*, the speaker verification system baseline is described. *Section 4* deals with experiments where the potentiality of the normalization function is illustrated through a multi-recognizer speaker verification system. *Section 5* and *6* summarize the main results and underline the potential advantages of the normalization method proposed.

2. NORMALIZATION METHOD

Let $f(s|X)$ be the similarity measure between model X and test signal s and $R_S = f(s|X)/f(s|\overline{X})$ be the similarity ratio where $f(s|X)$ is normalized by a world model, representing the population in general.

The principle of the normalization method presented here is to replace similarity measure $f(s|X)$ with the a posteriori probability so that similarity ratio R_S is a target score (as opposed to a non-target or impostor score).

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

[2] Maximum A Posteriori.

According to the Bayes rule, this a posteriori probability, denoted as $P(X = X_S|R_S)$ is defined by:

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

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

This normalization offers advantages similar to the MAP normalization approach. These two methods both propose probabilistic scores (bounded in [0,1] interval) which are dependent on the environmental conditions of the system through the a priori probabilities.

Conversely, the preliminary world model-based normalization leads to minimize the amount of data and tuning conditions usually required for the MAP normalization function estimation.

3. THE SPEAKER VERIFICATION SYSTEM

3.1 Speaker models

The speaker verification system is based on EM-trained (Expectation-Maximization [9]) Gaussian Mixture Models (GMM [10]) to represent the 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 constraint $\sum_{i=1}^{M} p_s^i = 1$, and the i-th uni-modal gaussian density, summarized by mean vector μ_s^i and covariance matrix Σ_s^i.

In this experimental context, a 16 gaussian mixture characterized by full covariance matrices is used to estimate speaker and world models.

3.2 Segmental framework (and acoustical parameterization)

The signal is characterized each 10 ms by cepstrum parameters. Cepstral mean subtraction (CMS) is applied in order to operate a blind deconvolution.

The segmental framework relies on two successive steps. First, 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 yielded by computing a geometrical mean over T consecutive frames (T frame long segments) as follows:

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

In this context, the segment length is constant.

3.3 Multi-Recognizer System

The verification system presented here is based on several recognizers working each on a specific frequency band [11]. This multi-band architecture consists of:

- a Full Band (FB), composed of 16 cepstrum components and representing band 300-4000Hz
- three Sub-Bands (SB1, SB2, SB3), each made up of 8 cepstrum components and representing bands 300-1660Hz for SB1, 1100-3100Hz for SB2 and 2500-4000Hz for SB3.

This architecture coupled with the segmental scheme, described in the previous section, allows to merge recognizer scores at the segmental level (Recognizer Fusion step on figure 1). The merging technique applied here is a weighted arithmetical mean.

Therefore, once the recognizer fusion is performed, a temporal fusion (Segment Fusion on figure 1) has to be achieved on resulting merged segment scores in order to yield a final score. This second fusion is based on an arithmetical mean.

Then, the final score is compared to a threshold in order to accept or reject the claimed speaker.

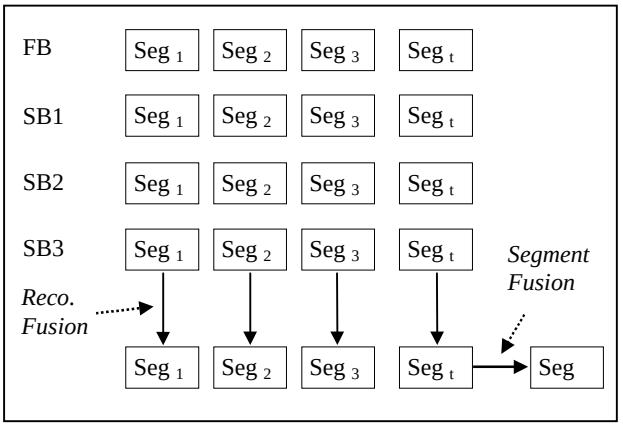

Figure 1: speaker verification system based on segmental and multi-recognizer architecture.

4. EXPERIMENTS

4.1 Data sets

The method proposed in this paper is experimented on a data set extracted from NIST/NSA 1998[3] evaluation campaign of speaker verification systems. 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[4]. 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 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 and 100 female speakers (50 client and 50 impostor speakers for each gender). Each verification test is about 30 seconds long (30s NIST test condition). Finally, the test stage includes about 600 target and 4400 impostor trials.
- A validation data set (denoted as Eva data set) with the same size and structure as the previous one, but on a different speaker population.

NB: It has to be noticed that there is no overlapping between these three data sets.

The 2 sets, Dev and Eva, are split into 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 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 [12].

The a priori probabilities are determined according to expected target and non target test trials. In this context, $P(X = X_s)$ and $P(X \neq X_s)$ are set respectively to 0.1 and 0.9.

[3] http://www.nist.gov/speech/spkrec98.htm

[4] The Elisa consortium is composed of European research laboratories, working on a reference plateform for speaker recognition system evaluation. These laboratories are: ENST (France), EPFL (Switzerland), IDIAP (Switzerland), IRISA (France), LIA (France), VUTBR (Czech Republic), RMA (Belgium).

4.3 Speaker Verification Results

A series of experiments are conducted in order to:

- compare performance of individual recognizers using either a classical world model-based normalization, named Ratio, or the one presented here named Ratio+MAP.
- examine the behavior of this original normalization in the multi-recognizer system.

4.3.1 Individual Recognizer performance

Table 1 shows Equal Error rate (EER) obtained on each individual recognizers (FB, SB1, SB2 and SB3) with Ratio and Ratio+MAP normalization. Results are provided for tests conducted on Dev and Eva data sets.

	EER on Dev data set		**EER on Eva data set**	
Recognizer	Ratio	Ratio+MAP	Ratio	Ratio+MAP
FB	0.16	0.15	0.17	0.16
SB1	0.28	0.265	0.27	0.27
SB2	0.20	0.20	0.22	0.23
SB3	0.30	0.29	0.30	0.29

Table 1: EER according to Dev data set (MAP normalization learnt on and applied to Dev data set) and Eva (MAP normalization learnt on Dev and applied to Eva data set).

It can be observed that individual recognizers are not disturbed by Ratio+MAP normalization compared to results obtained by Ratio one. EER across Ratio and Ratio+MAP normalization remain similar on both Dev and Eva data sets.

Independently of the normalization applied, the significant difference, in terms of EER, between the recognizers has to be noticed. This is an important issue to consider in the case of the multi-recognizer framework whose results are given in the next section.

4.3.2 Multi-recognizer system results

Table 2 gives results of experiments conducted with the multi-recognizer system (on Eva data set only). This system integrates the two fusion steps described in section 3.3. In this experimental context, the first fusion step, involving recognizer merging, is performed using different kinds of weighting:

- equally distributed weights across the different classifiers (1,1,1,1)
- performance dependent weights according to individual recognizers (EER).

As in the previous section, Ratio and Ratio+MAP normalization results are provided for comparison.

NB: the normalization is applied, at the segmental level, on each individual recognizer before the fusion step. Obviously, in the case of Ratio+MAP approach, the normalization function is recognizer dependent.

Results (illustrated by Table2) show that:

- With equally distributed weighting, Ratio+MAP leads to a 10% EER gain (a decrease from 0.18 to 0.158) if compared to Ratio normalization technique.
- With EER weighting, a gain is observed for Ratio normalization whereas Ratio+MAP performance remains constant.

As expected, Ratio+MAP normalization seems to be able to take the quality of the recognizers, in terms of performance, into account. This point is borne out by:

- the results obtained with an equally distributed weighting across the different recognizers although their own performances are very dissimilar (16 to 29% of EER)
- the constant behavior of Ratio+MAP normalization through the two kinds of weighting functions.
- the performances of the recognizer fusion and the best individual recognizer, FB, in the case of (1,1,1,1) weighting, which remain similar for Ratio+MAP normalization (0.158 against 0.16) whereas they are degraded for Ratio-based recognizer fusion (0.18 against 0.17)

	Recognizer Weights (FB,SB1,SB2,SB3)	
Normalization	(1,1,1,1)	(EER)
Ratio	0.18	0.167
Ratio+MAP	0.158	0.158

Table 2: EER obtained by the multi-recognizer system on Eva data set, according to normalization methods and different recognizer weights.

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 performance.

In the multi-recognizer system framework, this approach is able to integrate the difference in performance across classifiers, involving an EER gain without recognizer weighting.

6. FURTHER WORK

This normalization allows to "shift" scores from the similarity domain into the probability one. Secondly, this normalization systematically integrates both the behavior of the recognizer and environment data conditions. Accordingly, these properties should be fully explored in order to examine the potentiality of this original method to make the tuning step of the decision threshold for speaker verification easier.

7. 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] Heck L., Weintraub M., "Handset-dependent background models for robust text-independent speaker recognition", *Proc. International Conference on Acoustics, Speech and Signal Processing*, 1997.

[6] 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.

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

[8] 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.

[9] 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.

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

[11] Besacier L., Bonastre J.-F., Fredouille C., "Localization and Selection of Speaker Specific Information with Statistical Modeling", *Speech Communication*, 1999 (to be published)

[12] Fredouille C., Bonastre J.-F., Merlin T., "Segmental normalization for robust speaker verification ", *Proc. Workshop on robust methods for speech recognition in adverse conditions Cost 249*, May, 1998.