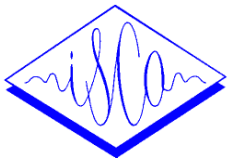

2001: A Speaker Odyssey
The Speaker Recognition Workshop
Crete, Greece
June 18–22, 2001

Bayesian Approach based-Decision in Speaker Verification

Corinne Fredouille, Jean-François Bonastre, and Teva Merlin

LIA/CERI Université d'Avignon, Agroparc,
BP 1228, 84911 Avignon Cedex 9, France
{corinne.fredouille, jean-francois.bonastre, teva.merlin}@lia.univ-avignon.fr

Abstract

Considering Bayesian decision framework applied in the context of speaker verification, this paper presents a new way of handling troublesome anti-speaker model by proposing a redefinition of hypotheses involved in the classical statistical hypothesis test. This new definition of hypotheses is then implemented through a speaker independent normalization technique, named MAP approach. Besides supporting these new hypotheses, MAP approach takes the advantages of projecting likelihood scores into a probabilistic domain and therefore of providing the decision threshold with bounded and meaningful values.
In this paper, different variants of MAP approach are presented which mainly aims at reducing likelihood variability, well-known in speaker verification to degrade system performance. MAP approach is firstly combined with classical normalization techniques (likelihood ratio normalization (world model) and/or Hnorm normalization technique). The second kind of variants consists in redesigning MAP approach to become speaker dependent. Experiments conducted on a subset of Switchboard database involving these different variants have showed that MAP approach is able to perform as well as classical normalization techniques while yielding probabilistic scores suitable for the decision threshold setting or the fusion of recognizer scores in the context of a multi-recognizer architecture.

1. Introduction

In speaker verification, one of the key problems remains the decision strategy and the setting of decision threshold(s). This is especially highlighted in the context of field applications. The Bayesian hypothesis test is usually proposed in the state of the art as the decision framework. If this solution seems well suited to set the decision threshold, different points remain sensitive. Particularly, while the statistical estimation theory provides a good way of estimating the client hypothesis, the situation is different for the anti-speaker hypothesis. This one is commonly estimated empirically by using a world model [1], without theoretical support. Numerous difficulties come from this point such as the treatment of variability and threshold setting, partially taken into account in the literature by combining various normalization processes (like Z/Hnorm [2] or Tnorm [3]). Moreover, using all these approximations and normalizations takes the signification of scores and thresholds away from the Bayesian framework. In the context of field applications, it could be difficult to define and to explain the threshold as well as to combine scores yielded by different recognizers (used for instance in security protocols). To cope with these problems, we propose a new definition of the client and anti-speaker hypotheses and a new formulation of scores, in a posteriori Bayesian framework.

This work was supported by the LIARMA Project (LIA/Avignon and RMA/Brussels).

2. Theoretical aspects

2.1. Hypothesis test

Let X denote the speaker whose identity is claimed and y a speech utterance claimed as belonging to speaker X. Considering the statistical hypothesis test theory, we can denote :

• Hypothesis H_0 : *speech utterance y belongs to speaker X.*

• Hypothesis H_1 : *speech utterance y belongs to another speaker $\overline{X}$*, named the anti-speaker.

and the decision rule is given by :

$$p(H_0) \underset{reject}{\overset{accept}{\gtrless}} p(H_1) \quad (1)$$

Applied in speaker verification, equation 1 becomes:

$$p(X|y) \underset{reject}{\overset{accept}{\gtrless}} p(\overline{X}|y) \quad (2)$$

where $p(X|y)$ is the probability of speaker X to have uttered speech signal y.

2.2. Bayesian decision

After application of Bayesian rule, equation 2 can be formulated in the following way :

$$\frac{p(y|X)}{p(y|\overline{X})} \underset{reject}{\overset{accept}{\gtrless}} \frac{P(\overline{X})}{P(X)} \quad (3)$$

where $P(X)$ (resp. $P(\overline{X})$) is the a priori probability of the claimed speaker to be speaker X – client test – (resp. speaker $\overline{X}$ – impostor test –).

Now consider the total cost of a speaker verification system defined as :

$$C = C_{\overline{X}|X}.P(X).p(\overline{X}|X) + C_{X|\overline{X}}.P(\overline{X}).p(X|\overline{X}) \quad (4)$$

where $P(X)$ (resp. $P(\overline{X})$) is the a priori probability of a client test (resp. an impostor test), $C_{\overline{X}|X}$ (resp. $C_{X|\overline{X}}$) is the cost of a false rejection (resp. a false acceptance) and $p(\overline{X}|X)$ (resp. $p(X|\overline{X})$) is the probability of a false rejection (resp. a false acceptance) yielded by the system.

Minimization of this total cost leads to the transformation of equation 3 into:

$$\frac{p(y|X)}{p(y|\overline{X})} \underset{reject}{\overset{accept}{\gtrless}} \frac{C_{\overline{X}|X}}{C_{X|\overline{X}}}.\frac{P(\overline{X})}{P(X)} \quad (5)$$

3. Redefinition of hypotheses

3.1. Hypothesis definition

Since speaker verification systems classically rely on statistical approach, probabilities $p(y|X)$ and $p(y|\overline{X})$, defined in equation 5, are actually approximated by likelihood measures denoted $L_{\mathcal{X}}(y)$ and $L_{\overline{\mathcal{X}}}(y)$ (where $\mathcal{X}$ (resp. $\overline{\mathcal{X}}$) is the model of speaker X (resp. the model of anti-speaker $\overline{X}$)).
In this context, the estimation of anti-speaker model $\overline{\mathcal{X}}$ is troublesome. Indeed, as anti-speaker $\overline{X}$ represents all the speakers different from speaker X, $\overline{\mathcal{X}}$ is not directly computable.

As we cannot very well approximate anti-speaker model $\overline{\mathcal{X}}$, we investigate an approach which considers the hypothesis test in a different context. Indeed, we now consider hypotheses H_0 and H_1 defined as follows :

- Hypothesis H_0 : *The claimed identity is correct given likelihood $L_{\mathcal{X}}(y)$ (client test).*
- Hypothesis H_1 : *The claimed identity is wrong given likelihood $L_{\mathcal{X}}(y)$ (impostor test).*

Under these new hypotheses, equation 5 can be reformulated as :

$$p(X = Y|L_{\mathcal{X}}(y)) \underset{reject}{\overset{accept}{\gtrless}} p(X \neq Y|L_{\mathcal{X}}(y)) \quad (6)$$

or as (after application of Bayes rule):

$$\frac{p(L_{\mathcal{X}}(y)|X = Y)}{p(L_{\mathcal{X}}(y)|X \neq Y)} \underset{reject}{\overset{accept}{\gtrless}} \frac{P(X \neq Y)}{P(X = Y)} \quad (7)$$

where $P(X = Y)$ (resp. $P(X \neq Y)$) is the a priori probability of a client test (resp. an impostor test).

3.2. Pdf estimation

One of the main concern regarding equation 7 is related to the estimation of both probabilities $p(L_{\mathcal{X}}(y)|X = Y)$ and $p(L_{\mathcal{X}}(y)|X \neq Y)$. In this paper, it has been decided to estimate these probabilities empirically from probability density functions (pdf), learned from a set of client (for $p(L_{\mathcal{X}}(y)|X = Y)$) and impostor (for $p(L_{\mathcal{X}}(y)|X \neq Y)$) verification tests. By this way, probabilities take the behavior of the speaker verification system into account.

3.3. Advantages

To summarize, the main advantages of the proposed approach are:

- to provide a new way of handling anti-speaker $\overline{X}$;
- to take the intrinsic quality of a given recognizer into account.

4. Score and threshold signification

4.1. MAP approach

Given a state of the art speaker verification system, it is currently difficult to provide decision threshold with sense-based value. Nevertheless, it would be very appreciated, in the context of field applications, to be able to set the decision threshold with directly interpretable values.
Similarly, in the context of a multi-recognizer architecture, sense-based scores would make the fusion of various recognizer scores easier.

To deal with these two aspects, the MAP normalization technique is proposed. This technique relies on the new hypotheses presented in the previous section and proposes, according to Bayesian rule, to define probability $p(X = Y|L_{\mathcal{X}}(y))$ as follows:

$$p(X = Y|s) = \frac{p(s|X = Y).P(X = Y)}{p(s|X = Y).P(X = Y) + p(s|X \neq Y).P(X \neq Y)} \quad (8)$$

where s refers to $L_{\mathcal{X}}(y)$.

In this context, final scores are defined in a probabilistic domain where the decision threshold can be set and interpreted easily. The projection of likelihood scores into a probabilistic domain is quite useful for the required fusion of recognizer scores in the case of a multi-recognizer architecture [4].
The second advantage of this new definition of scores relies on the integration of a priori probabilities – $P(X = Y)$ and $P(X \neq Y)$ – directly into the computation of scores. In this way, the condition of use of the system is directly conveyed by final scores.

4.2. Practical aspects

A well-known issue affecting speaker verification systems is the variation of speech signal between training and testing phases. This variation comes from the speaker himself/herself (intra-speaker variability) as well as from differences in recordings and/or in transmission conditions. This variability is classically handled by applying normalization techniques such as likelihood ratio based on a world model [1], Z/Hnorm normalization [2] or Tnorm normalization [3].

WMAP and Hnorm
To minimize the variability of likelihood scores, we propose to couple MAP approach with different normalization techniques such as :

- a classical world model. In this context, likelihood scores $L_{\mathcal{X}}(y)$ are now normalized and score s, defined in equation 8, consequently refers to : $\frac{L_{\mathcal{X}}(y)}{L_{\mathcal{W}}(y)}$ (where $\mathcal{W}$ denotes the world model). This combination will be referred to as WMAP approach in this paper.

- a classical world model as well as Hnorm normalization technique. Here, score s resulting from the world model application is then involved in the specific Hnorm formulation : $\frac{s-\mu_{\mathcal{X}}}{\sigma_{\mathcal{X}}}$ where $\mu_{\mathcal{X}}$ and $\sigma_{\mathcal{X}}$ parameters respectively refer to the mean and the standard deviation of the impostor distribution related to speaker model $\mathcal{X}$.

In both situations, it has to be noticed that the client and impostor pdfs (required for MAP approach as seen in section 3.1) are learned on a separate data set. During test, the probability scores resulting from the MAP application are therefore independent from the claimed speaker as opposed to Hnorm normalization technique.

Speaker dependent WMAP
A variant of WMAP approach is also proposed in this paper to deal with the likelihood score variability. Here, we investigate the way of introducing speaker dependency into WMAP approach. As not enough data may be available to estimate reliable speaker dependent client pdfs, speaker dependent impostor pdfs are considered in this paper. In a similar way to Hnorm normalization technique, each speaker model is involved in a series of impostor tests to provide impostor scores and to estimate a speaker dependent impostor pdf. Finally, these speaker dependent impostor pdfs are involved in equation 8 to provide speaker dependent probability scores. This approach will be referred to as speaker dependent WMAP approach in this paper.

5. Experiments

5.1. Database

The method proposed in this paper is validated by using a data set extracted from the NIST/NSA 1999 evaluation campaign. This subset is composed of speech recordings issued from Switchboard database and built from concatenated telephone conversation segments. Experiments are conducted on four different data subsets defined by the ELISA consortium [5][6]:

- The first subset is used to compute gender and handset dependent world models for likelihood ratio normalization. It is composed of 120 male speakers (102 males using electret handsets and 18 males using carbon handsets) and 196 female speakers (186 females using electret handsets and 10 females using carbon handsets), which corresponds to about 6 hours of speech.

- The second subset is called *Dev*. It is composed of 100 speakers (50 males and 50 females). About two minutes of speech recorded over only one session are available for each speaker to train speaker models. On the other hand, 519 client tests and 5190 impostor tests are available for the test set. Duration of test speech segments varies from 3 to 60s. In this paper, this subset will be used to estimate speaker independent client and impostor pdfs.

- Concerning the third subset, called *Eva*, its structure is similar to the previous one. Only the population of speakers used is different as well as the number of tests (499 client tests and 4990 impostor tests).

- The final subset, called *Norm*, is used for Hnorm normalization technique and for speaker dependent WMAP approach. It is composed of 159 speakers (50 females using electret handsets, 24 females using carbon handsets, 50 males using electret handsets and 35 males using carbon handsets).

5.2. Baseline system

The signal is characterized each 10 ms by 16 cepstrum coefficients with their delta. Cepstral Mean Subtraction (CMS) is applied to operate a blind deconvolution. Finally, a specific algorithm based on a bi-Gaussian distribution of energy is used to perform a frame selection [6].

Both speaker and world models are based on Gaussian Mixture Model (GMM) approach. World models are estimated by using EM (Expectation-Maximization) algorithm based on Maximum Likelihood estimation. On the other hand, speaker models are derived from a world model, by using a variant of Maximum a Posteriori estimation [6]. In this paper, a 128 Gaussian mixture characterized by diagonal matrices is used to estimate both world and speaker models.

Finally, similarity measure between speech utterance and model is based on likelihood estimation.

5.3. Results

A series of experiments was conducted to compare MAP approach with classical normalization techniques. This comparison aimed at evaluating the behavior of the different variants of MAP approach (described in section 4.2) regarding speaker verification system performance.

Here, it has to be noticed that probability density functions required for MAP approach (to evaluate probabilities $p(s|X = Y)$ and $p(s|X \neq Y)$) are estimated on *Dev* data set and implied to normalize scores computed from *Eva* data set.:

Finally, results of these various experiments are measured by two different DET curves [7]:

- the first one, named ST, represents results obtained from tests involving electret handsets for both training and testing speech segments, but different telephone lines.
- the second one, named DT, refers to results obtained from tests involving electret handsets for training speech segments and either electret or carbon handsets for testing speech segments. Different telephone lines are also used in this case.

Figure 1 presents a comparison between WMAP approach and the simple use of the likelihood ratio based on a classical world model. Regarding both ST and DT DET curves, we can observe that WMAP and the likelihood ratio method lead to quite similar performance. This point demonstrates that WMAP, by proposing probabilistic scores for the decision phase, does not degrade speaker verification system performance.

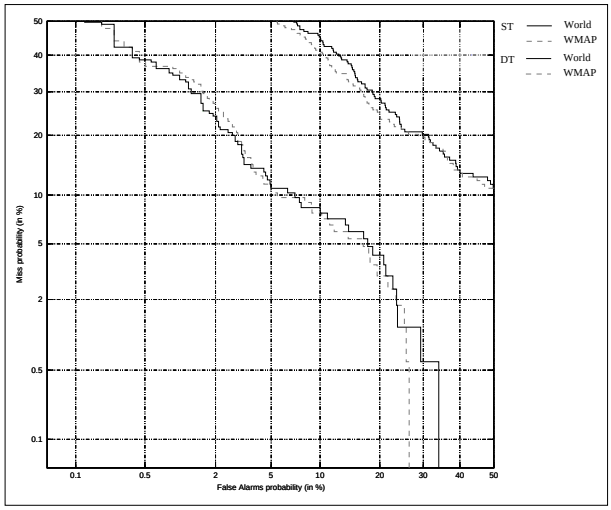

Figure 1: *WMAP approach vs. world model. Comparison between classical likelihood ratio normalization technique and WMAP approach on Eva data set.*

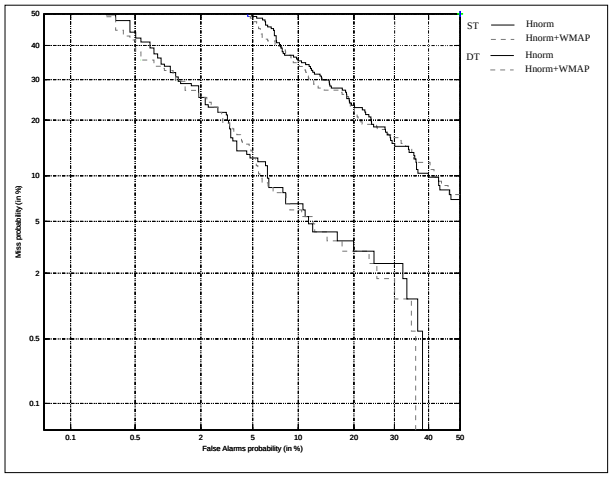

Figure 2: *Hnorm+WMAP approach vs. Hnorm. Comparison between classical Hnorm normalization technique and the combination of Hnorm and WMAP approach on Eva data set.*

Secondly, we compare the WMAP approach coupled with Hnorm normalization technique with Hnorm used alone[1]. This comparison is provided in figure 2. We can observe that both approaches also obtain similar performance. This statement shows that WMAP approach can be easily combined with classical normalization techniques while providing probability scores without loss of performance.

[1]Obviously, Hnorm normalization technique is combined with world model based-normalization.

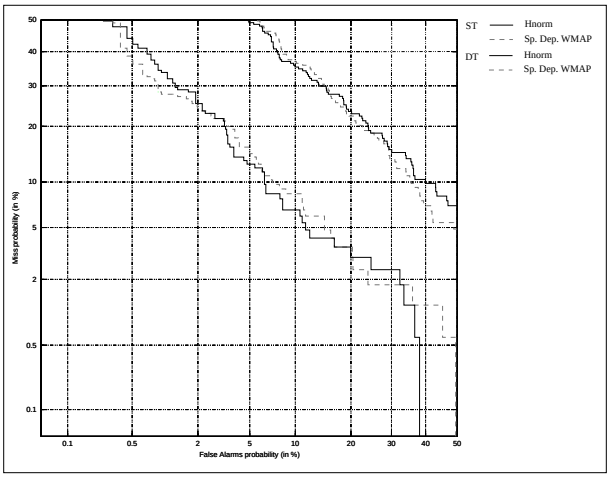

Figure 3: *Speaker dependent WMAP approach vs. Hnorm. Comparison between Hnorm normalization technique and speaker dependent WMAP approach on Eva data set.*

Finally, we present the results obtained by applying speaker dependent WMAP. These ones are reported on figure 3 (Sp. Dep. WMAP) and compared with the DET curve related to the use of Hnorm normalization technique (seen previously on figure 2). Again, speaker dependent WMAP approach performance is quite close to Hnorm normalization technique (Hnorm technique performs slightly better on ST DET curve (8.0 vs. 8.7% for EER), but worse on DT DET curve (22,4 vs. 21.1% for EER)). This statement is quite interesting since speaker dependent WMAP does not take handset information into account as opposed to Hnorm normalization while obtaining very close performance.

6. Conclusion

This paper first proposes a new definition of statistical test hypotheses classically involved for the decision phase in speaker verification. This new definition provides an original way of dealing with anti-speaker $\overline{X}$, which remains troublesome in this context of speaker verification. To handle this new definition of hypotheses, this paper presents a speaker independent normalization technique, called MAP approach. The main advantage of this approach is to project likelihood scores into a probabilistic domain, providing the decision threshold with an easily interpretable value. Finally, different variants of MAP approach are presented to cope with large variability of likelihood scores, which are well-known to degrade speaker verification performance. These variants rely on either the combination of MAP approach with classical world model normalization (referred to as WMAP approach) or Hnorm normalization technique, or on the proposal of a speaker dependent WMAP approach.

Experiments related to these different variants demonstrate that MAP approach, combined with normaization techniques is able to perform quite well compared with those techniques applied only. Especially, it is important to point out that speaker dependent WMAP approach obtains performance close to Hnorm normalization technique while it does not take handset information into account (as opposed to Hnorm technique), which is well-known to be relevant for performance improvement in speaker verification.

Further work will focus on ways of integrating such a handset information into MAP approach while dealing with the probable lack of data available to estimate handset and speaker dependent impostor pdf (required for MAP approach).

7. References

[1] M. J. Carey, E. S. Parris, "Speaker verification using connected words", *Proceedings of Institute of Acoustics*, 1992.

[2] S. A. Reynolds, "The effects of handset variability on speaker recognition performance: experiments on the Switchboard corpus", *ICASSP*, Atlanta (USA), 1996.

[3] R. Auckenthaler, M. Carey, H. Lloyd-Thomas, "Score normalization for text-independent speaker verification system", *Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 speaker recognition workshop*, Vol.10(1-3), 2000.

[4] C. Fredouille, J.-F. Bonastre, T. Merlin, "Similarity normalization method based on world model and a posteriori probability for speaker verification", *In proceedings of Eurospeech'99*, Budapest (Hungary), 1999.

[5] The ELISA consortium, "The ELISA systems for the NIST'99 evaluation in speaker detection and tracking", *Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 speaker recognition workshop*, Vol.10(1-3), 2000.

[6] Magrin-Chagnolleau I. & al, "Overview of the ELISA consortium research activities", *2001 a speaker Odyssey: the speaker recognition workshop*, Chania (Crete), June 2001.

[7] A. Martin, G. Doddington, T. Kamm, M. Ordowski, M. Przybocki, The DET curve in assessment of detection task performance, *Proceedings of EUROSPEECH'97*, Rhodes (Greece), 1997.