D-MAP : A DISTANCE-NORMALIZED MAP ESTIMATION OF SPEAKER MODELS FOR AUTOMATIC SPEAKER VERIFICATION

Mathieu BEN, Frédéric BIMBOT

IRISA/METISS, Campus Universitaire de Beaulieu
35042 Rennes Cedex, FRANCE, European Union
mben,bimbot@irisa.fr

ABSTRACT

In this paper we introduce a MAP estimation of speaker models in Automatic Speaker Verification with a distance constraint: the D-MAP adaptation. The D-MAP is based on the Kullback-Leibler distances and provides an easy way to automatically compute a speaker-dependent adaptation of the model parameters. We formulate a distance constrained MAP criterion and we show an equivalence between the D-MAP adaptation and the score normalization called D-Norm. From the results obtained with the D-MAP technique, we show that this method gives better performance than a classical speaker-independent MAP adaptation. It is also found that the D-MAP based system without score normalization performs similarly to a classical MAP system with a model-based score normalization.

1 Introduction

In text-independent Automatic Speaker Verification (ASV), the main state-of-the-art approach for speaker modeling consists in using Gaussian Mixture Models (GMM) [1]. This family of models are suited to describe multi-variate real densities when choosing an appropriate number of gaussians in the mixture, and they provide a good way to model statistical behavior of speaker acoustic features. Bayesian adaptation of the speaker models, with a Maximum A Posteriori (MAP) criterion, have shown to be more efficient that the Maximum Likelihood (ML) estimation, in particular when the amount of training data available is limited. In the MAP estimation scheme, the speaker models are adapted from an *a priori* model, called the background model or world model, using the data of the training utterance for each speaker. This method gives more robust estimates of the model parameters because it limits over-adaptation on the training data (as opposed to the ML estimation) by assuming a prior distribution for these parameters. The background model, which is used as an initial model for adaptation, is built using a large amount of speech data from various speakers, with ML estimation.

Crucial factors in MAP adaptation are the weighting coefficients between the *a priori* parameters and the parameters derived directly from the training data (i.e. the ML parameters). These coefficients balance the weight between *a priori* knowledge taken from the background model, and new knowledge from the training utterance. Thus, in a general way, it should be given a high contribution, in the adaptation process, to training utterances with high information content, in term of variety and amount of acoustic features, because in these cases the ML estimates of model parameters should be good. On the contrary, training utterances with poor information content would lead to bad ML estimated parameters and then should have a low weight in MAP adaptation, letting prior parameters to dominate. In the GMM MAP theory [2], the weighting coefficients are data-dependent, gaussian index-dependent, and can be parameter-dependent. However, in practice, more simple determinations of these coefficients are usually used by considering coefficients that are gaussian-independent and parameter-independent, and which only depend on the training utterance duration or are simply fixed to a constant value.

In this paper, we propose an alternative way to automatically determine a speaker dependent weighting coefficient using information content considerations on the training data. We use a relative information measure from the field of information theory which is the Kullback-Leibler (KL) distance between an estimated speaker model and the world model. Recent works [3] have shown that the KL distances between the speaker models and the world model are strongly correlated with the mean scores delivered by these models for impostor accesses. We propose a technique to normalize the speaker models at the estimation level such that every speaker models are at a fixed distance from the world model, by computing the appropriate weighting coefficient. This should imply a concentration of the impostor scores and thus a better separation of the distributions of client scores and impostor scores. This technique can be formulated as a distance constrained MAP adaptation, the D-MAP, which is exposed in section 2 of this paper. We first recall the classical MAP adaptation formulae for GMM and the various approaches used in practice to determine the weighting coefficient for adaptation. We then expose a constrained MAP criterion based on the KL distances and we explain the links between the D-MAP adaptation and the score normalization called D-Norm (distance normalization). The experiments and results of the D-MAP technique are reported in section 3

where we show that this technique, which determines an individual weighting coefficient for each speaker, outperforms the classical MAP technique with a global optimization of a fixed weighting coefficient. The section 4 concludes this paper and exposes the perspectives of this work.

2 D-MAP estimation of GMMs

2.1 MAP adaptation of GMMs

The probability density function (p.d.f) $p(y)$ of a K-component GMM for d-dimensional acoustic vectors y is defined as:

$$p(y) = \sum_{k=1}^{K} w_k \mathcal{G}_k(y) \quad (1)$$

where $\mathcal{G}_k$ is a gaussian function with a d-dimensional mean vector μ_k and a $d \times d$ covariance matrix Σ_k which is usually assumed to be diagonal, and w_k is the relative weight of $\mathcal{G}_k$ in the mixture ($\sum_{k=1}^{K} w_k = 1$). From a training utterance $\mathcal{Y} = \{y_1,...,y_T\}$ of duration T, the MAP estimation of the parameters $\{w_k,\mu_k,\Sigma_k\}$ of a GMM speaker model can be achieved using the EM algorithm by iteratively adapting the parameters, initially set to the prior model parameters. The new estimates $\{\hat{w}_k,\hat{\mu}_k,\hat{\Sigma}_k\}$ are computed from the old ones $\{w_k,\mu_k,\Sigma_k\}$ and from the training data via the following formulae [1] :

$$\hat{w}_k \propto \alpha_k^w(\gamma_k/T) + (1-\alpha_k^w)w_k \quad (2)$$

$$\hat{\mu}_k = \alpha_k^\mu \overline{y}_k + (1-\alpha_k^\mu)\mu_k \quad (3)$$

$$\hat{\Sigma}_k = \alpha_k^\Sigma S_k + (1-\alpha_k^\Sigma)(\Sigma_k + \mu_k^2) - \hat{\mu}_k^2 \quad (4)$$

The data-derived parameters γ_k, $\overline{y}_k$, and S_k are respectively the global occupation rate of gaussian $\mathcal{G}_k$, and the sample mean and variance calculated under the p.d.f $\mathcal{G}_k$ (see [1] or [2] for the corresponding formulae). The new estimates of the weights $\{\hat{w}_k\}$ are rescaled so that they sum to one.

The weighting coefficients $\{\alpha_k^w,\alpha_k^\mu,\alpha_k^\Sigma\}$ depend on the training data, the gaussian index and the parameter. For a parameter $\theta \in \{w,\mu,\Sigma\}$, the corresponding coefficient is defined as :

$$\alpha_k^\theta = \frac{\gamma_k}{\gamma_k + \rho(\theta)} = \frac{1}{1+\rho(\theta)/\gamma_k} \quad (5)$$

where $\rho(\theta)$ is an a priori fixed relevance factor which can depend on the parameters and which control the balance of the adaptation. This factor, which is homogeneous to γ_k, must be determined using prior experiments or knowledge, and can be seen as an equivalent occupation rate for the a priori data. In practice, some simplifications can be used to compute the weighting coefficients by considering parameter- and gaussian- independent coefficients. For example, these coefficients may only depend on the duration T of the training utterance which corresponds to set $\rho(\theta)$=ρ and to replace γ_k by T in (5). This is a very intuitive determination of the weighting coefficients because it gives a high weight in the adaptation to training utterances with a long duration. An other way, which is the simplest one, is to fix the weighting coefficients to a constant value independently of the training data. This corresponds to $\rho(\theta) \propto \gamma_k$ in equation (5). In that case, it is usually necessary to globally optimize the value of this unique weighting coefficient α on preliminary experiments, which implies the use of a development data set. This technique which does not take into account the information content of the training utterances, leads to non optimal performance because of speaker-dependent biases in the distribution of the verification scores. This is due to the heterogeneous information content of the training utterances and score normalization techniques can be undertood as a means to compensate these biases. In our baseline ASV system, this basic classical MAP adaptation is implemented and we will use it as a reference in the following of this paper.

In the next section, we propose an information-based MAP criterion which lead to a speaker-dependent MAP adaptation, the D-MAP.

2.2 Kullback-Leibler distance constrained MAP criterion

2.2.1 General formulation

It has been experimentally shown in [3] that the mean impostor score $\overline{S}_{imp}$ obtained with a speaker model is strongly correlated with the symmetric KL distance $KL2$ between this model and the world model. These two quantities are linked with an approximative linear relation :

$$\overline{S}_{imp} \approx -c.KL2 \quad (6)$$

where c is a positive constant. It means that varied and sparse KL distances between the speaker models and the world model would lead to sparse impostor scores with speaker-dependent biases in their distribution. On the contrary, concentrated KL distances would lead to more concentrated impostor scores with reduced speaker dependent biases. Starting from these considerations, we want to introduce an equi-distance constraint at the estimation level, to force every speaker models to be at a common reference distance from the world model.

The symmetric KL distance $KL2$ between a speaker model p_{X_l} and the world model p_W is the sum of the two oriented dual KL divergences, defined as relative entropies between these models:

$$KL2 = E_{p_{X_l}}[\log \frac{p_{X_l}}{p_W}] + E_{p_W}[\log \frac{p_W}{p_{X_l}}] \quad (7)$$

Given that in the MAP adaptation scheme the world model is used as the a priori model, the $KL2$ distance for a given speaker model is linked to the weighting coefficient α used for the adaptation :
- for α=0, no adaptation at all is performed and then the speaker model and the world model are strictly identical. In this case we have: $KL2 = 0$.
- for α=1, the maximum adaptation is performed and the

estimated speaker model is the ML estimate. In this case the KL distance has its ML value : $KL2 = D_{ML}$.

We can then express $KL2$ as a function of α and D_{ML}:

$$KL2 = D_{ML}.f(\alpha) \tag{8}$$

where $f(.)$ is a function of α with the constraint $f(0) = 0$ and $f(1) = 1$. If we want to obtain a constant KL distance equal to a reference distance D_{ref} for each speaker X_l, the corresponding speaker-dependent α coefficient is :

$$\alpha^{(X_l)} = f^{-1}(D_{ref}/D_{ML}^{(X_l)}) \tag{9}$$

This relation defines the distance constrained MAP criterion for the D-MAP adaptation scheme. Note that it implies to know the ML distance $D_{ML}^{(X_l)}$ for every speaker. So an ML estimation of every speaker model has to be done before performing the D-MAP adaptation. Furthermore, we assume in relation (9) that the function $f(.)$ is known and invertible. Also, as the KL distance increases (possibly not monotonically) from 0 to D_{ML} when α varies from 0 to 1, the reference distance should be chosen such that $\forall X_l, D_{ref} < D_{ML}^{(X_l)}$ to ensure that $\alpha^{(X_l)} \in [0,1]$ for all X_l.

2.2.2 GMM mean-only adaptation with D-MAP

In our ASV system, we only adapt the means $\{\mu_k\}$ of the speaker GMMs which corresponds to set α_k^w=0 in (2) and α_k^Σ=0 in (4). The weights and variances of every speaker models are set to the values of the world model ones. In practice, this method has shown no loss of performance versus the more sophisticated method of a complete adaptation of the parameters (weight, means and variances) [1]. In this section we expose the determination of the D-MAP coefficient $\alpha^{(X_l)}$ in the case of mono-gaussian models and multi-gausian models, with mean-only adaptation.

The gaussian case:

The gaussian case corresponds to have K=1 in (1). We study this basic case to understand the behavior of $KL2$ versus α and to introduce the multi-gaussian case. The MAP estimate of the mean $\hat{\mu}^{X_l}$ of a gaussian speaker model p_{X_l} has the following form :

$$\hat{\mu}^{(X_l)} = \alpha\overline{y} + (1-\alpha)\mu^{(W)} \tag{10}$$

We can then show that, in the case of mean only adaptation (i.e. $\Sigma^{(X_l)} = \Sigma^{(W)} = \Sigma$) the KL distance between a gaussian speaker model p_{X_l} and the gaussian world model p_W has the quadratic form :

$$KL2 = D_{ML}.\alpha^2 \tag{11}$$

where D_{ML} is a function of the sample mean $\overline{y}$, the world model mean $\mu^{(W)}$ and the covariance matrix Σ. The value for $\alpha^{(X_l)}$ can easily be computed by inverting this relation :

$$\alpha^{(X_l)} = (D_{ref}/D_{ML}^{(X_l)})^{\frac{1}{2}} \tag{12}$$

This relation may be used for every speaker in the case of gaussian models. We will see that we can not find such a relation in the case of multi-gaussian models.

The multi-gaussian case:

In the multi-gaussian case, we can not find a closed form relation for the $KL2$ distance as in the mono-gaussian case. So we can't express the D-MAP coefficient $\alpha^{(X_l)}$ as a simple function of $D_{ML}^{(X_l)}$ and D_{ref}. We choose to use an iterative procedure to determine the appropriate value of $\alpha^{(X_l)}$. We make a first estimation $\alpha_{(0)}^{(X_l)}$ of this coefficient by determining an easily-inversible approximative relation between $KL2$ and $\alpha^{(X_l)}$ of the following form :

$$KL2 \approx D_{ML}.\alpha^\beta \tag{13}$$

We then have:

$$\alpha^{(X_l)} \approx (D_{ref}/D_{ML}^{(X_l)})^{1/\beta} \tag{14}$$

With a given β, we compute a first approximation $\alpha_{(0)}^{(X_l)}$, we estimate the corresponding model and we compute its KL distance with a Monte-Carlo method as in [3]. We then refine the value of $\alpha^{(X_l)}$ with a dichotomous procedure until $KL2^{(X_l)}$ approximates D_{ref} with a 5% accuracy.

The experiments have shown that for the majority of the speakers, only a few iterations (1 or 2) of this procedure are necessary to achieve the estimation of $\alpha^{(X_l)}$. A development on a little subset of speakers has shown that $\beta = 4$ is a good compromise to approximate the average behaviour of $KL2^{(X_l)}$ versus $\alpha^{(X_l)}$.

2.3 Link between D-MAP and D-Norm

The D-MAP process leads to speaker models that all are at a constant distance D_{ref} from the world model. Given the relation (6) the impostor mean scores for every speaker should then be approximately constant :

$$\overline{S}_{imp}^{D-MAP}(X_l) \approx -c.D_{ref} = constant \tag{15}$$

The D-Norm is a score normalization based on the use of the Kullback-Leibler distance between the speaker models and the world model. A score $S(X_l)$ for an access with claimed identity X_l is D-Normalized as follows :

$$S^{D-Norm}(X_l) = \frac{S(X_l)}{KL2^{(X_l)}} \tag{16}$$

By the relation (6), this leads to impostor mean scores that are quasi-constant, as in the case of D-MAP :

$$\overline{S}_{imp}^{D-Norm}(X_l) = \frac{\overline{S}_{imp}(X_l)}{KL2^{(X_l)}} \approx -c = constant \tag{17}$$

Thus, the D-MAP adaptation and the D-Norm score normalization have a similar effect on the impostor scores. They only differ by the scaling factor D_{ref} and they should therefore lead to comparable performance. This indicates that there is an equivalence between a model-based score normalization and an appropriate adaptation scheme and that such score normalizations are not always necessary.

3 Experiments and results

3.1 Description of the ASV system and database

The ASV system that we use is derived from the IRISA/ELISA baseline system for the NIST'01 evaluation [4]. The acoustic analysis of this system gives 32-dimensional acoustic vectors with the first 16 cepstral coefficients and their respective deltas. The statistical models are 128-components GMMs with diagonal covariance matrices and we use gender- and handset-type-dependent world models. We adapt the speaker models from the world model using the EM algorithm with a MAP criterion. The database that we use is a subset of the NIST'01 evaluation set [5], which contains telephone conversations of american students and where the training utterances are about 2 minutes long.

3.2 Results

The Equal Error Rates (EERs) of the IRISA/ELISA system with the classical MAP adaptation are given in TAB.1 with several values of α. The performance is globally optimized by choosing α=0.4, which gives an EER of 11.2%. TAB.1 also reports the EER of the system

MAP					
α	0	0.25	0.4	0.6	0.8
EER(%)	13.7	12.05	11.2	11.25	11.45

D-MAP					**MAP0.4 +D-Norm**
D_{ref}	0.25	0.5	0.8		
EER(%)	10.45	10.4	10.4		10.75

TAB. 1 –. *EER of the IRISA/ELISA system with classical MAP adaptation, D-MAP and MAP0.4+D-Norm*

with a D-MAP adaptation with three different values for D_{ref}. The results show that the choice of D_{ref} does not seem to be decisive for the performance. The D-MAP procedure, which gives an EER of about 10.4%, outperforms the classical MAP with the optimal value for α. Figure 1 plots the Detection Error Tradeoff (DET) curves of the IRISA/ELISA system for an ML estimation (MAP0) as a reference, an optimal classical MAP adaptation (MAP0.4) and the D-MAP adaptation (D-MAP). Through these curves we show that the D-MAP system performs better than the MAP0.4 system for functioning points going from low miss rates to the EER point, and that these systems perform comparably for points with a low false alarm rate.

In TAB.1 we also give the EER of the MAP0.4 system with D-Norm (10.75%), which is slightly outperformed by the D-MAP system. Furthermore, through additionnal experiments we have noticed that model-based normalizations like D-Norm or Z-Norm have no influence on the performance of the D-MAP system. Thus, such score normalizations are not necessary when the adaptation process is appropriate, and they are only useful to compensate speaker-dependent biases due to non-optimal adaptation.

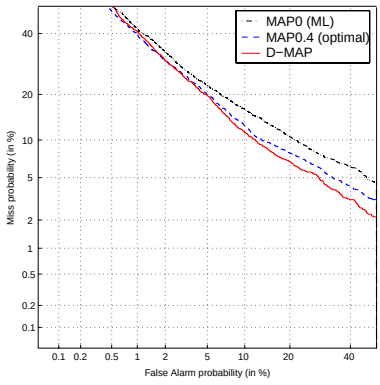

FIG. 1 –. *DET curves. Performance of the IRISA/ELISA system with various estimation schemes, without score normalization.*

4 Conclusion

We have proposed a distance-normalized MAP estimation of the speaker models in ASV: The D-MAP adaptation. From the results we obtained, it has been shown that the D-MAP performs better than a classical MAP which globally optimizes a fixed adaptation coefficient. Furthermore we have illustrade that model-based score normalizations are not so essential (even may be unnecessary) when the D-MAP is performed because a model normalization is directly applied at the estimation level. However, the D-MAP does not have the effect of score normalizations based on the test data like the T-Norm. Nevertheless, we think that information content considerations can also be used for the test data, in a similar way than in the D-MAP, for example by learning distance-normalized models of the tests. These test models could be used in the scoring process and could lead to test-normalized scores similar to those obtained with the T-Norm.

5 References

[1] A. Reynolds, T.F. Quatieri, and R.B. Dunn. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10(1–3), 2000.

[2] J. L. Gauvain and C. H. Lee. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Trans. on Speech and Audio Proc.*, 2(2), April 1994.

[3] M. Ben, R. Blouet, and F. Bimbot. A Monte-Carlo method for score normalization in automatic speaker verififcation using Kullback-Leibler distances. In *Proceedings of ICASSP2002*, 2002.

[4] I. Magrin-Chagnolleau, G. Gravier, and R. Blouet. Overview of the 2000-2001 ELISA consortium research activities. In *Proceedings of 2001: A Speaker Odyssey-The Speaker Recognition Workshop*, 2001.

[5] National Institute of Standards and Technology. The 2001 NIST speaker recognition evaluation. http://www.nist.gov/speech/tests/spk/2001/.