BECARS : a free software for speaker verification

Raphaël Blouet[†]*, Chafic Mokbel*[‡]*, Hoda Mokbel*[‡]*, Eduardo Sánchez Soto*[†]*, Gérard Chollet*[†] *and Hanna Greige*[‡]

[†]ENST, dépt. TSI
46 rue Barrault, 75634 Paris
France
{blouet,esanchez,chollet}@tsi.enst.fr

[‡]University of Balamand
El-Koura,BP 100 Tripoli
Lebanon
{chafic.mokbel,hanna.greige}@balamand.edu.lb

Abstract

The aim of Automatic Speaker Verification (ASV) is to detect whether a speech segment has been uttered by the claimed identity or by an impostor. Our contribution includes the distribution of BECARS , a free software based on Gaussian Mixture Models (GMM) for Automatic Speaker Verification (ASV), and the design of a new methodology to estimate the decision score in an ASV system. BECARS in available at `http://www.tsi.enst.fr/~blouet/Becars/`. The main characteristic of this software is to allow the use of several adaptation techniques including the most common ones such as *Maximum A Posteriori* (MAP) and *Maximum Likelihood Linear Regression* (MLLR). The proposed method for score computation is based on the use of a hierarchical Gaussian clusterization method that we describe in details in this paper.
We introduce this work with a general summary of Automatic Speaker Verification (ASV), followed by a description of the adaptation technique available in BECARS used in this work. We then present and evaluate our score computation scheme before concluding the paper.

1. Introduction

Given a speech segment $\underline{\underline{Y}} = \{\underline{Y}_1, \cdots, \underline{Y}_N\}$ and a hypothesized (or claimed) identity X, the aim of Automatic Speaker Verification is to determine whether $\underline{\underline{Y}}$ has been uttered by X or not.

Automatic Speaker Verification is often formulated as a classical hypothesis test between two hypotheses H_X and $H_{\bar{X}}$ with:

H_X : $\underline{\underline{Y}}$ has been uttered by X
$H_{\bar{X}}$: $\underline{\underline{Y}}$ has been uttered by another speaker

The optimal test to decide between these two hypotheses is the likelihood ratio test:

$$S_X(\underline{\underline{Y}}) = \log \frac{p_X(\underline{\underline{Y}})}{p_{\bar{X}}(\underline{\underline{Y}})} \underset{H_{\bar{X}}}{\overset{H_X}{\gtrless}} \theta \qquad (1)$$

where θ is the decision threshold.

This approach relies on the hypothesis of the existence of both probability density functions p_X and $p_{\bar{X}}$ on the whole observation space $\mathbf{Y}$ of frames $\underline{Y}_t$.

For a decade, the state of the art approach consists in using Gaussian Mixture Models [9] to modelize both probability density functions. Moreover, training of each client model is mostly done by adaptation of $p_{\bar{X}}$ parameters [10]. BECARS allows the use of several kinds of adaptation criterions. In the next section, we describe the one that we used for this work. More details on adaptation techniques available in BECARS can be found in [2] or in the software documentation.

2. Adaptation techniques for client models determination

Hidden Markov Model (HMM) and GMM adaptation techniques have largely been studied in the last decade. In [6] a unified theoretical framework has been proposed in which the two major classes of techniques, *Maximum A Posteriori* (MAP) and transformation based adaptation (like *Maximum Likelihood Linear Regression* (MLLR)), appear as particular cases. Several adaptation techniques have been derived within this framework and applied in BECARS in order to determine client models.

Model adaptation can be seen as a particular case of training where a small amount of uncontrolled data is used to estimate new values for the parameters. In such cases, the training must be controlled in order to ensure that the estimated parameters are not specific to the training data. In order to incorporate this idea, adaptation is seen as a function with a variable degree of freedom that transforms the values of the parameters. The degree of freedom must be chosen as a function of the available training data. In order to achieve this variable degree of freedom, the general adaptation theory proposed in [6] matches a binary tree with a GMM. As shown in Figure 1 (with a 4 Gaussian components GMM), each component of the GMM stands for one leaf of the tree. From the root of the tree to the leaves, different cuts can be defined allowing different possible distribution classifications. For every possible classification, a number of classes is obtained. In each of these classes, an adaptation function

may be associated.

In order to build the tree, Gaussian distributions are grouped two by two up to the root. At each grouping step, we chose the two closest distributions given the distance $d(\cdot,\cdot)$. The distance $d(\mathcal{N}_1,\mathcal{N}_2)$ between two Gaussian $\mathcal{N}_1$ and $\mathcal{N}_2$ is defined as the likelihood loss on the training frames if $\mathcal{N}_1$ and $\mathcal{N}_2$ are replaced by a single distribution $\mathcal{N}_3$. In the tree construction, $\mathcal{N}_3$ is associated with the $\mathcal{N}_1$ and $\mathcal{N}_2$ parent node.

$$d(\mathcal{N}_1,\mathcal{N}_2) = \log \frac{|\underline{\underline{\Sigma_3}}|^{\frac{\alpha_1+\alpha_2}{2}}}{|\underline{\underline{\Sigma_1}}|^{\frac{\alpha_1}{2}} |\underline{\underline{\Sigma_2}}|^{\frac{\alpha_2}{2}}} \quad (2)$$

On equation 2, $|\underline{\underline{\Sigma_1}}|$ and $|\underline{\underline{\Sigma_2}}|$ are the covariance matrices of the two nodes and $|\underline{\underline{\Sigma_3}}|$ is the covariance matrix of an equivalent node. α_1 and α_2 are the training factors for the two nodes respectively.

In order to perform adaptation, the standard **EM** algorithm [3] is modified.

At the end of the **E** step, the weights associated with Gaussian distributions forming the leaves of the tree are propragated up to its root. Then, starting from the root, the tree is processed and we stop at nodes whose children have weights less than a predefined threshold. This predefined threshold represents the minimal amount of data necessary to perform the adaptation. This allows the determination of a classification that is a function of the amount of available data. This process is described in the Figure 1.

At the **M** step of the **EM** algorithm, a regression function is associated with every class of Gaussian distributions and every dimension of the acoustic feature space. All the Gaussian distributions in the class will have their mean and variance adapted as following:

$$\begin{aligned} \mu_i &= a_i m_i + b_i \\ \sigma_i^2 &= a_i^2 s_i^2 \end{aligned} \quad (3)$$

with:

- m_i and s_i^2 respectively the mean and the variance of the i^{th} *prior* distribution of the *prior* GMM,
- μ_i and σ_i^2 respectively the mean and variance of the adapted Gaussian distribution,
- a_i and b_i, parameters of the regression function.

General equations for the estimation of regression parameters in the framework of the unified adaptation theory are given in [6]. Here, we only consider the case of one particular adaptation called MLLR_MAP in BECARS . Equations 4 and 5 allow us to obtain the regression coefficients and equation 6 presents reestimation formulae obtained in this case. In equations

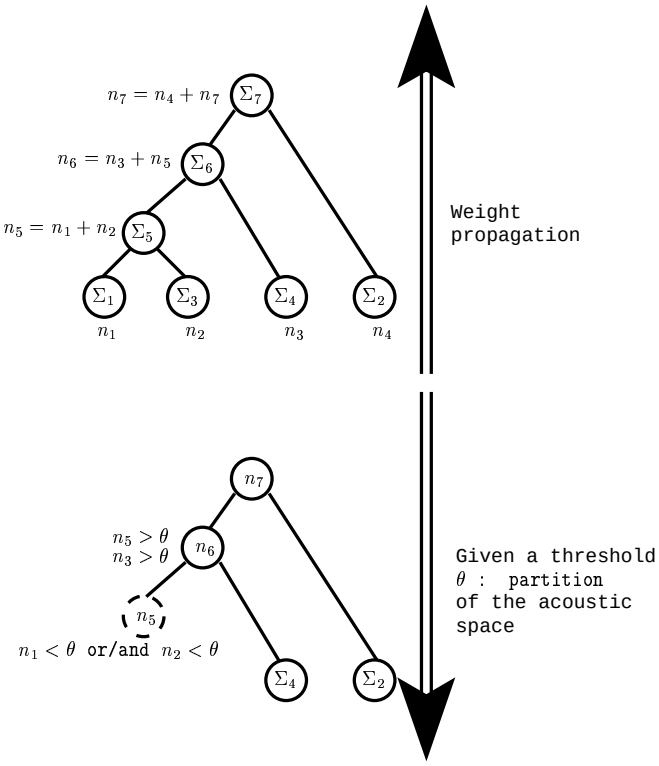

Figure 1: Description of the two steps that permit the determination of the number of degrees of freedom used for adaptation

4 and 5, J is the number of Gaussian distributions in the q^{th} class, n_j is the weight of the j^{th} distribution after the **E** step, $r0_j$ is the *a priori* precision, m_j the *a priori* mean, $\bar{y}_j$ and $\bar{y}_j^2$ are the first and second order moments observed after the **E** step.

$$\begin{aligned} 0 &= |a_q^2| \cdot \left[\sum_{j=1}^{J} n_j\right] + |a_q| \cdot \left[\sum_{j=1}^{J} r0_j \cdot n_j \cdot m_j \cdot \bar{y}_j \right. \\ &\quad \left. - \frac{\left(\sum_{j=1}^{J} r0_j \cdot n_j \cdot \bar{y}_j\right) \cdot \left(\sum_{k=1}^{J} r0_k \cdot n_k \cdot m_k\right)}{\sum_{j=1}^{J} r0_j \cdot n_j}\right] \\ &\quad - \left[\sum_{j=1}^{J} r0_j \cdot n_j \cdot \bar{y^2}_j - \frac{\left(\sum_{j=1}^{J} r0_j \cdot n_j \cdot \bar{y}_j\right)^2}{\sum_{j=1}^{J} r0_j \cdot n_j}\right] \quad (4) \end{aligned}$$

$$b_q = \frac{\sum_{j=1}^{J} [r0_j \cdot n_j \cdot (\bar{y}_j - a_q \cdot m_j)]}{\sum_{j=1}^{J} [r0_j \cdot n_j]} \quad (5)$$

Equation 4 shows that the regression coefficient a_q is the solution of the second degree equation. As shown in [6], this equation always has two solutions of opposite sign. In order to smooth further the adaptation, an empiric *Maximum A Posteriori* (MAP) is applied to the

estimation of the mean :

$$\mu_i = 0.8 \cdot (a_i m_i + b_i) + 0.2 \cdot m_i \quad (6)$$

3. Frame weighting using MMI

GMM modeling of speech frames represented by their corresponding feature vectors does not take into consideration the order of the frames. This means that rearranging the speech signal frames will not affect their likelihood computed using the GMM. Thus, the collection of frames from a speaker's utterance available in a test represents a sample from the speaker population of frames. Moreover, the GMM-based Automatic Speaker Verification system can be viewed as calculating the expected value over this sample of a decision function. In our case, this decision function corresponds to the log likelihood ratio betwenn hypothesis H_X and $H_{\bar{X}}$.

Let $\underline{Y}_t$ be the feature vector representing a speech frame extracted from a test speech utterance; $\underline{Y}_t$ is supposed to be the realization of a random multivariate variable $\underline{Y}$. Let $llr(\underline{Y})$ be the log likelihood ratio function considered as the main argument of the decision function; a theoretical expected value of $llr(\underline{Y})$ is given by:

$$\begin{aligned} LLR &= E[llr(\underline{Y})] \\ &= \int_{\Omega_Y} llr(\underline{Y}) p(\underline{Y}) d\underline{Y} \end{aligned} \quad (7)$$

If we assume that the process underlying the production of $\underline{Y}$ is ergodic, it is equivalent to estimating $llr(\underline{Y})$ over the parameter space than estimating it with a time average. This explains why the score for a given utterance of T frames is calculated as:

$$L\hat{L}R = \sum_{t=1}^{N} llr(\underline{Y}_t) \quad (8)$$

However, a signal frame carries different information about the underlying speech message such as the speaker's identity, prosody, the communication channel, etc. Let us define the binary random variable I_S representing the fact that a signal frame carries information about the speaker identity. To illustrate this idea we simply cite the obvious example of silence signal frames which, in general, carry little information about the speaker identity. Using this random variable, a better estimation of the LLR may be derived from 7:

$$\begin{aligned} LLR &= E_{I_s}[llr(\underline{Y})] \\ &= \int_{\Omega_Y} llr(\underline{Y}) p(I_s|\underline{Y}) P(\underline{Y}) d\underline{Y} \\ &\propto \int_{\Omega_Y} llr(\underline{Y}) p(\underline{Y}|I_s) d\underline{Y} \end{aligned} \quad (9)$$

Therefore, and even if the process of generating the feature vectors is supposed to be ergodic, the average in equation 9 is not equivalent to an average over time. The choice of the sample for estimating the LLR value should be done with care. An alternative consists in weighting the instantaneous LLR measurement by a factor that depends on the relevance of the corresponding frame regarding the characterization the speaker identity.

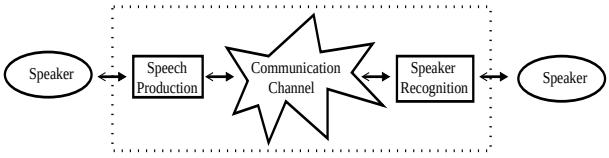

Figure 2: *View of the global chain of a speaker recognition system.*

Several approaches exist in order to calculate these weights. In the present work, we propose to perform a vector quantization and to associate a weight with each feature subspace defining a class. If $C(\underline{Y})$ defines the class of a feature vector $\underline{Y}$, we approximate $p(I_s|\underline{Y})$ in the equation 9 by a discrete distribution defined by $p(I_s|C(\underline{Y}))$.

In this paper and in order to determine parameters of the discrete probability distribution $\{p(I_s|C(\underline{Y}))\}$, the maximization of the Mean Mutual Information (MMI) is used. To illustrate the principle of using the MMI criterion, the complete chain of an Automatic Speaker Recognition system is provided in Figure 2. A given speaker utters a speech signal which goes across a communication channel to reach the ASV system that is used to determine the identity of the speaker. In this model, we suppose that a communication channel is defined going from the speech production module to the speaker recognition module included. In the development phase, we add the true speaker identity to the input of the ASV system and we optimize $\{p(I_s|C(\underline{Y}))\}$ to have the output identity as close as possible to the one provided in the input. To summarize, we want to obtain a maximum of the information that has been put into the communication channel or to maximizes the Mean Mutual Information. In the developement phase of our ASV system, we chose the weights of the discrete probability distribution $\{p(I_s|C(\underline{Y}))\}$ that maximise this information. Figure 3 summarises the process of the weights estimation.

4. Evaluation protocol and result

Evaluations related in this section are made using the NIST 2003 data set [8].

Acoustic parametrisation consists in 20 mel cepstra filter bank coefficients with their delta. Channel equalisation is performed through Feature Warping [7].

We use a 128 components GMM to model p_X and $p_{\bar{X}}$. For each speaker p_X is obtained by adaptation of $p_{\bar{X}}$ following equation 6. $p_{\bar{X}}$ is trained using 2 hours

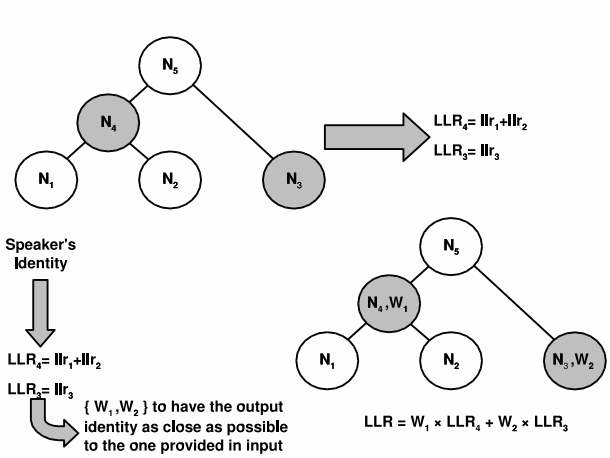

Figure 3: *Description of the MMI based score computation strategy.*

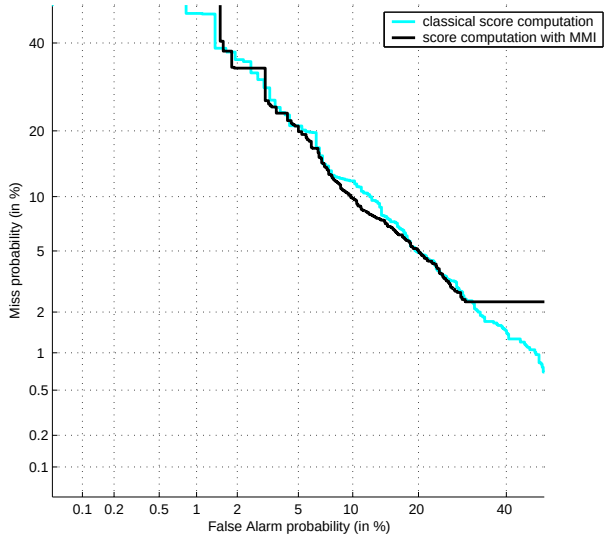

Figure 4: *DET curves associated to the use or not of the MMI criterion.*

of speech uttered from 100 speakers of the NIST 1999 evaluation campaign. The discrete probability distribution $\{p(I_s|C(\underline{Y}))\}$ is estimated using 500 test files and 50 speakers of the NIST 2003 evaluation data. In the results presented here, we used 32 differents classes. The two Det curves [5] obtained with and whithout the use of the frame weighting are plotted in Figure 4. Unless score computation using the MMI approach performs slightly better than the classic score computation strategy, both systems have very close level of performance. We believe that this can be explained mostly by the lack of data available to estimate the discrete probability distribution $\{p(I_s|C(\underline{Y}))\}$.

5. Discussion and conclusion

As the proposed approach appears theoretically very promising, results obtained on the NIST 2003 data set are not as good as expected. However, we still believe that this approach may improve ASV system robustness and we will run further experiments using differents configurations. The use of several strategies and criterion to estimate the discrete probability distribution $\{p(I_s|C(\underline{Y}))\}$ will also be investigated.

6. References

[1] R. Auckenthaler, M. Carey and H. Lloyd-Thomas, *Score Normalization for Text-Independant Speaker Verification Systems*, Digital Signal Processing Vol 10., Nos 1-3, Janvier 2000.

[2] R. Blouet, C. Mokbel and Gérard Chollet, BECARS *: un logiciel libre pour la vrification du locuteur*, JEP 2004, Fèz, April 2004.

[3] A.P. Dempster, N.M. Laird, D.B. Rubin, *Maximum likelihood from incomplete data using the EM algorithm*, Journal of the Royal Statistical Society, 39(B), 1977.

[4] Y. Linde, A. Buzo et R. Gray,*An Algorithm for Vector Quantizer Design*, IEEE Transactions on Communications,1980.

[5] A. Martin, G. Doddington, T. Kamm, M. Ordowski and M. Przybocki, *The DET Curve in Assessment of Detection Task Performance*, Proceedings of EuroSpeech 1997, Volume 4, pp. 1895-1898.

[6] C. Mokbel, *Online adaptation of hmms to real life conditions: A unified framework*, IEEE Transaction on Speech and Audio Processing, 2001.

[7] J. Pelecanos and S. Sridharan, *Feature Warping for Robust Speaker Verification*, Workshop Odyssey, 2001.

[8] M. Przybocki and A. Martin, *The NIST Year 2003 Speaker Recognition Evaluation Plan*, 2003.

[9] D.A. Reynolds, *A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification*, Georgia Institute of Technology, 1992.

[10] D.A. Reynolds, *Comparison of background normalization methods for text independent speaker verification*, Eurospeech'97, 1997.

Acknowledgement:
This work has partly been funded by CEDRE, a French-Lebanese cooperation framework (involving ENST and UOB). It has also benefited from the scientific support of the ELISA consortium.