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

A Tree-Based Approach for Score Computation in Speaker Verification

Raphaël BLOUET and Frédéric BIMBOT

IRISA-METISS (CNRS & INRIA)
Campus Universitaire de Beaulieu
35042 RENNES cedex - France - European Union
{rblouet,bimbot}@irisa.fr

Abstract

This paper proposes an original approach to the task of speaker verification, in which the training process consists in a direct modeling of the score function. It divides the parameter space in disjoint regions where a score can be obtained as a simple function of the vector position in the region. The aim of this approach is, on the one hand to overcome some undesirable properties of the Gaussian Mixture Models (GMMs), and on the other hand, to speed up the decision process.

First, we present the formalism of probabilistic speaker verification and we discuss some motivations for exploring alternative approaches. We then describe a method currently under investigation, which is based on a binary recursive partition of the acoustic parameter space into regions to which an elementary scoring function is associated. Finally, we provide illustrations and preliminary results of the method, together with conclusions and perspectives.

1. Introduction

The problem of speaker verification can be formulated in a probabilistic framework as follows :

Given a speech segment $Y = \{y_1, \ldots, y_N\}$ and a hypothesized (or *claimed*) speaker X, the aim of speaker verification is to determine whether Y has been uttered by X, or not.

As described for instance in [6], and using the same notations, the speaker verification task is a classical hypothesis test between two hypotheses H_X and $H_{\overline{X}}$ with :
H_X : Y has been uttered by X
$H_{\overline{X}}$: Y has been uttered by another speaker

Theoretically, the optimal test to decide between these two hypotheses is a likelihood ratio test :

$$S_X(Y) = \frac{p_{H_X}(Y)}{p_{H_{\overline{X}}}(Y)} \begin{cases} \geq \theta & \text{accept } H_X \\ < \theta & \text{reject } H_X \end{cases}$$

where θ is a decision threshold.
This approach relies on the hypothesis of the existence of both probability density functions p_{H_X} and $p_{H_{\overline{X}}}$ on the whole observation space $\mathbf{B}$ of frames y_t.

The state of the art approach consists in using GMM's [5],[6] to estimate both probability density functions. Yet, despite its good performance, the GMM approach has two main disadvantages:

- high computational complexity,
- uncontroled behaviour for low-likelihood frames.

High computational complexity is caused by the large number of exponentials that have to be computed for calculating the likelihood functions. The second point is due to the fact that GMMs model the whole distribution of acoustic parameters, but sometimes the tail of the distribution is poorly modeled, which may cause uncontroled behaviour of the likelihood *ratio* for outlier frames in the test utterance, and thus bias the decision.

We propose an alternative approach which consists in modeling directly the likelihood ratio, rather

than both likelihood functions independently. We also investigate on the use of a particular family of functions, relying on a tree-based partition of the acoustic parameter space and a constant score in each region.
The aim of this work is mainly to highly reduce the CPU execution time and the amount of memory required to process a verification access.

2. Description Of The Method

2.1. Tree-based partition of the acoustic space

Let $R = \{R_k\}_{1 \leq k \leq K}$ be a partition of the acoustic space $\mathbf{B}$:

$$\bigcup R_k = \mathbf{B} \quad \text{and} \quad R_i \cap R_j = \emptyset$$

Let's define a criterion $C(R)$ measuring some property of the partition R and let's denote R^* the *best* partition of $\mathbf{B}$ with respect to criterion C :

$$R^* = \arg \min_R C(R)$$

For instance, C can be a measure of the heterogeneousness of the partition in terms of claimed speaker frames *vs* non-speaker frames in each region R_k.

In practice, finding R^* is generally untractable. However, sub-optimal solutions can be obtained by specific algorithms, such as the CART approach (Classification and Regression Trees) [1].

Under this approach, the acoustic space $\mathbf{B}$ is split recursively. At each iteration, the algorithm selects the optimal 2-region split among all possible ways of splitting existing regions (resulting from the previous iteration).

In the classical CART approach, criterion C is expressed as a weighted average of a local criterion c_k (called region impurity), which can be the entropy (1) or the Gini diversity index (2) of the data within each region.

$$c_k = -\sum_{j=1}^{J} p_j \log p_j \qquad (1)$$

$$c_k = 1 - \sum_{j=1}^{J} p_j^2 \qquad (2)$$

where p_j is the probability of the class j in region R_k.

2.2. Score computation

Once the partition has been obtained, it is possible to assign, to any frame y, the region $R(y)$ to which it belongs. Then, a local score $s_X(y)$ can be attributed to frame y, depending on the region $R(y)$. It can for instance be derived from the probability $P(X|R(y))$ of the claimed speaker conditionally to $R(y)$. Other estimators can also be used.

In the experiments reported here, the score $S_X(y)$ is constant and estimated in the maximum likelihood sense, i.e :

$$S_X(y) = \log P(X|R(y)) - \log P(\overline{X}|R(y))$$

More elaborate scoring functions can be used, provided that they need a limited number of operations.

The overall utterance score $S_X(Y)$ is obtained as the average of the frame-based likelihood :

$$S_X(Y) = \frac{1}{N} \sum s_X(y_t)$$

The tree structure and the low computational cost of the local scoring function allow a fast calculation of the utterance score.

3. Results

This section is divided in two parts : the first one is an illustration of the method based on a simplified example. The second one presents preliminary results obtained on a subset of the 2001 NIST Speaker Recognition Evaluation data [3] and [4].

3.1. Illustration of the method

For this illustration, a single client speech segment and a whole set of background speakers have been parametrized into 1-dimensional frames, corresponding to the second cepstral coefficient.

Two types of probability density function for p_{H_X} and $p_{H_{\overline{X}}}$ have been used:

- 16-components gaussian mixture models,
- a 25-bin histogram.

Figure 1 plots three curves. Two of them correspond to the likelihood ratio functions associated to the histogram (dashed line) and to the GMM (solid), and defined for both types of probability density function as :

$$s_X(y_t) = log\ p(y_t|H_X) - log\ p(y_t|H_{\overline{X}})$$

The third curve corresponds to the frame score obtained with the tree-based method described in the previous section. The region impurity function that we use is the entropy. The tree was learned with approximately 2 mn of speech (i.e $\approx$ 12000 frames) from a speaker X and the same number of frames from a background population of 150 speakers. The CART tree learning procedure has been controled so that each region of the tree contains a minimum of 350 frames.

It can be seen that the three curves have consistent behaviours. The CART approach models the log likelihood ratio as a piecewise constant function.

3.2. Preliminary performance evaluation

Figure 2 shows preliminary results obtained on a subset of the NIST 2001 evaluation data [4]. For this experiment, the speaker training data are composed of 2 mn of speech extracted from a single conversation (i.e approximately 12000 frames). The NIST condition used here is the primary condition.

The GMM approach uses 128 gaussian components for both the speaker and non-speaker (world) models. The world population is composed of 150 speakers $\times$ 2 mn of speech. The training criterion is MAP. This result have been obtained with the NIST 2001 ELISA Speaker Verification platform. More details on this platform can be found in [2].

For the CART approach, the non-speaker data are composed of 12000 frames randomly taken in the set of non-speaker frames, in order to have the same number of speaker and non-speaker data. The CART training is controlled in order to force a minimum number of 100 frames per region. This yields trees with 57 regions in average (across speakers).

Figure 2 plots three curves and enables to compare the performance of :

- the classical GMM approach describe above,
- the CART approach with entropy as impurity function,
- the CART approach with the Gini diversity index as impurity function.

In this experiment, the Gini diversity index performs slightly better than the entropy. This is consistent with what can be theoretically expected [1]. The CART approach yields significantly worse results than the GMM approach.

At least two reasons can explain the discrepancy of performance :

- a different amount of data were used, in the two experiments for modeling the nonclient population, as we use $\approx$ 600000 frames to learn the GMM associated to the non-client model, but only $\approx$ 12000 frames for the CART approach.
- the score function used in the CART approach is based on a ML estimation whereas the GMM uses a MAP criterion, which is classically more efficient.

We are currently working on the improvement of the CART method along these two directions. The method is very efficient in terms of quantity of memory needed to store a speaker template, and as concerns the simplicity of the verification process. Indeed as the 128 components gaussian mixture models require more than 30 kO, trees require less than 4 kO. Moreover as the calculation of the gaussian mixture score requires hundreds of exponentials, mutiplications and logarithms our approach only requires a few basic comparisons to affect a score to a frame. It permits to considerably reduce the required CPU-time.

4. Conclusions And Perspectives

A novelty of the approach to speaker verification proposed in this work relies on the direct modeling of the score function using local densities of the training data.

The tree-based representation of the acoustic space allows a very efficient computation of the frame-based score.

Several tracks are likely to improve the performance of the technique in the short-term.

5. Acknowledgments

This work is being financially supported by Bull-CP8.
All experiments have been performed using the Edinburgh Speech Tools Library that can be donwloaded from: <http://www.cstr.ed.ac.uk/downloads/>

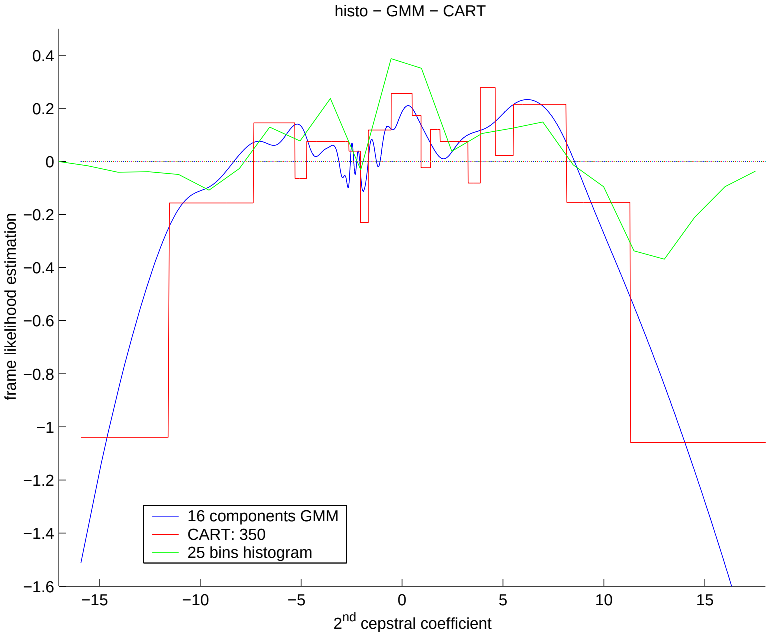

Figure 1: *Illustration of the method*

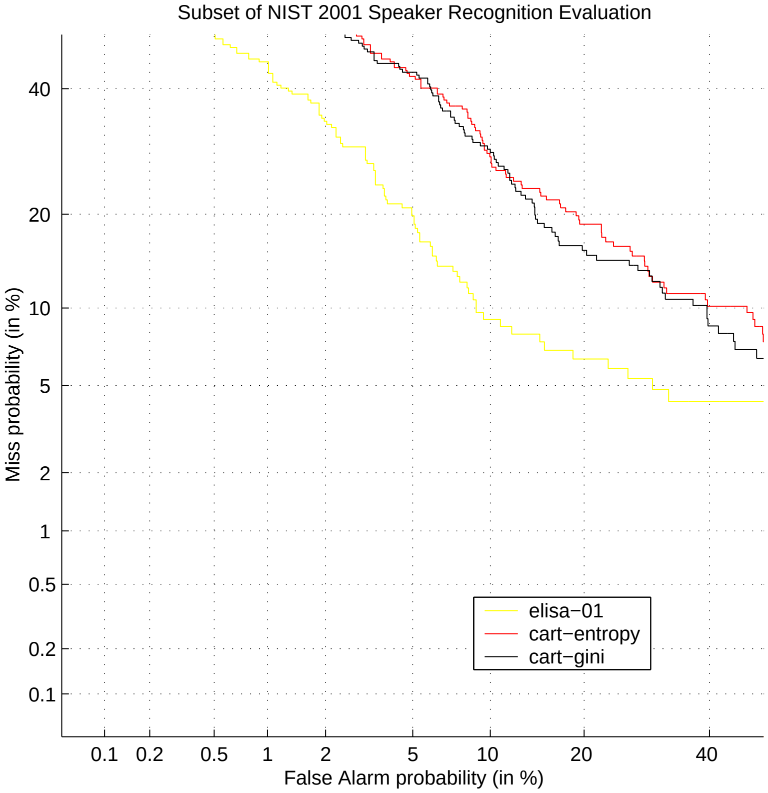

Figure 2: *Preliminary results (DET curves)*

6. References

[1] Breiman, J.H. Friedman, R.A. Olshen and C.J. Stone. *Classification And Regression Trees*, Wadsworth, 1994.

[2] I. Magrin-Chagnolleau, G. Gravier, R. Blouet for the ELISA consortium. *Overview of the 2000-2001 ELISA Consortium Research Activities*. In Proceedings of 2001: A Speaker Odyssey-The Speaker Recognition Workshop.

[3] A. Martin, M. Przybocki. *The NIST 1999 Speaker Recognition Evaluation - An Overview*. Digital Signal Processing Vol 10, Nos 1-3, Janvier-April-July 2000.

[4] National Institute of Standards and Technology. *The 2001 NIST Speaker Recognition Evaluation*. <http://www.nist.gov/speech/tests/spk/2001/index.htm>

[5] Douglas A. Reynolds. *A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification*. *PhD Thesis*, Georgia Institute of Technology, 1992.

[6] A. Reynolds, T.F. Quatieri, and R.B. Dunn. *Speaker Verification Using Adapted Gaussian Mixture Models*. Digital Signal Processing Vol 10, Nos 1-3, Janvier-April-July 2000.