Tree Based Score Computation for Speaker Verification

Raphaël Blouet and Frédéric Bimbot

IRISA (CNRS & INRIA) - METISS
Campus Universitaire de Beaulieu
35042 Rennes, 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 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.

The speaker verification task is a classical hypothesis test between two hypotheses \( H_X \) and \( H_Y \) with:

- \( H_X \) : \( Y \) has been uttered by \( X \)
- \( H_Y \) : \( Y \) has been uttered by another speaker

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

\[
S_X(Y) = \frac{p_{H_X}(Y)}{p_{H_Y}(Y)} \begin{cases} 
\geq \theta & \text{accept } H_X \\
< \theta & \text{reject } H_X 
\end{cases}
\]

where \( \theta \) is a decision threshold.

This approach relies on the hypothesis of the existence of both probability density functions \( p_{H_X} \) and \( p_{H_Y} \) on the whole observation space \( \mathcal{B} \) of frames \( Y \).

The State of the Art approach consists in using Gaussian Mixture Models (GMM) [3] to estimate both probability density functions. Yet, despite its good performance, the GMM approach has two main disadvantages:

- high computational complexity,
- uncontrolled 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 uncontrolled 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.

In the conventional approach (Eq.1), each frame \( y_t \) of the test utterance contributes to the decision score \( S_X(Y) \) via the log-likelihood of the two hypotheses \( H_X \) and \( H_Y \) through the use of 2 GMM’s. In the proposed approach (Eq.2) the decision score is computed from a direct likelihood data estimation associated to \( y_t \). The direct score modelization allows to considerably speed up the decision.

In equation 1 and 2, \( \bar{p} \) and \( \bar{s} \) mean an estimation of \( p \) and \( s \)

\[
S_X(Y) = \sum_{t=1}^{N} \log(\bar{p}_X(y_t)) - \log(\bar{p}_Y(y_t)) \quad (1)
\]

\[
S_X(Y) = \sum_{t=1}^{N} \bar{s}_X(y_t) \quad (2)
\]

Figure 1 explains the principle of the method. In this figure, the parameter space is split in regions to which a constant score (here a color) has been associated. During the test process, each vector \( y \) is assigned to a region, the score of which is used to compute the whole utterance decision score.
2. Description of the method

2.1. Acoustic parameter partition

Let \( R = \{ R_i \}_{1 \leq i \leq K} \) be a partition of \( K \) disjoint regions of the acoustic space \( \mathbb{B} \):

\[
\bigcup_{i=1}^{K} R_i = \mathbb{B} \quad \text{and} \quad R_i \cap R_j = \emptyset
\]

Let’s define a criterion \( C(R) \) measuring some properties of the partition \( R \).

- \( C(R) \) can be a measure associated to the observation class. An example of this kind of criterion is given by the intra-region heterogeneity in terms of claimed speaker frames vs non-speaker frames, such as the mean of the entropy over each partition. Let \( c_b \) be the entropy of the partition \( R_b \):

\[
c_b = -p_b(X) \log(p_b(X)) - p_b(\overline{X}) \log(p_b(\overline{X}))
\]

with \( p_b(\cdot) \) the prior probability of the class \( \cdot \) given the region \( R_b \). Then \( C(R) \) is defined as the mean entropy over each region:

\[
C(R) = \sum_{i=1}^{K} p_b \cdot c_b
\]

with \( p_b \) the prior probability of an observation to be put in the region \( R_b \).

- \( C(R) \) can be a measure directly associated to the observation vectors. An example of this kind of criterion is given by the minimization of the intra-region data dispersion. Let \( V_b \) and \( \mu_b \) be the intra-region covariance matrix and mean vector of data that belong to the region \( R_b \):

\[
\mu_b = \frac{1}{N_b} \sum_{y \in R_b} y
\]

\[
V_b = \frac{1}{N_b} \sum_{y \in R_b} (y - \mu_b)(y - \mu_b)^T
\]

\[
c_b = |V_b|
\]

where \( | \cdot | \) is the determinant of the matrix \( \cdot \) and \( N_b \) is the number of observation put in \( R_b \). Then \( C(R) \) is defined as the mean intra-region covariance matrix determinant over each partition:

\[
C(R) = \sum_{i=1}^{K} p_b \cdot c_b
\]

Figure (2,b) illustrates partition that this criterion allows to obtain. Note that this second criterion has not been tested.

Let’s denote \( R^* \) the best partition of \( B \) with respect to criterion \( C \):

\[
R^* = \arg \min_R C(R)
\]

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) [2].

Under this approach, the acoustic space \( \mathbb{B} \) 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 \( c \), which is usually the entropy of the data within each region. This is the criterion that we have used in the results that are presented in section 3.

2.2. Partition 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:

\[
S_X(Y) = \frac{1}{N} \sum_{y} s_X(y)
\]

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 1999 NIST evaluation data.

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 \( p_{H_S} \) have been used:

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

Figure 3 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:

\[
s_X(y) = \log p(y | H_X) - \log p(y | H_S)
\]

The third curve corresponds to the frame score obtained with the tree-based method described in the previous section. 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 controlled 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 4 shows preliminary results obtained on a subset of the NIST 1999 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 used is MAP.

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 50 regions in average (across speakers).

In this experiment, 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 non-client 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 performant.

We are currently working on the improvement of the CART method along these two directions. However, two other methods have been tested in order to obtain a better representation of the non-client hypothesis:

- build 10 trees per client speaker and use the average score over the 10 trees as decision score. The aim is on the one hand to use more non-client data and to smooth the decision score.
- select for each client training frame \( y_X \) the nearest non-client frame \( y_T \) possible,

\[
y_T = \min_{y \in Y_S} ||y_X - y||^2
\]

The aim is to adapt the non-client frame representation to client data.

We have unfortunately observed a performance decrease with all of these attempts to obtain a better non-client hypothesis modeling. Other methods in order to improve the performance can also be found [1].

4. Conclusion 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 fast computation of the frame-based score.

Although, none of our reported attempts to obtain a better non-client representation in the parameter space have improved the performance, this point seems to be crucial and work remains to be done in this direction. Several tracks are likely to improve the performance of the technique in the short-term. Among them we will primary focus on the estimation of the score associated to a region and on the definition of a more adapted split criterion.

5. References

Figure 3: Example of score functions with different approaches (GMM, CART, Histogram) in a 1-dimensional space.

Figure 4: Comparaison of GMM and CART based performances