PROBABILISTIC ANCHOR MODELS APPROACH FOR SPEAKER VERIFICATION

Mikaël Collet(1)(2), Yassine Mami(1), Delphine Charlet(1), Frédéric Bimbot(2)

(1) France Telecom R&D - TECH/SSTP - 2 av. Pierre Marzin - 22307 Lannion Cedex - FRANCE
{mikael.collet, yassine.mami, delphine.charlet}@francetelecom.com
(2) IRISA (CNRS & INRIA) - Campus de Beaulieu - 35042 Rennes Cedex - FRANCE
bimbot@irisa.fr

Abstract

This paper presents a probabilistic approach for representing a speaker using the anchor modelling technique and discusses the relation between this new approach and the deterministic approach. In the first part, the technique of anchor modelling is presented. Then the new approach, which models the various utterances of a speaker by a normal distribution in the anchor models space, is presented. A study of the relation between the two approaches (deterministic and probabilistic) leads to the definition of a combined approach. All these approaches are evaluated on the NIST 2000 database and results show improved performance of the combined approach over the deterministic and the probabilistic one.

1. Introduction

Recently, proposals have been made to model a speaker relatively to a set of other speaker models. This concept was first developed for adaptation in speech recognition, where eigenvoices [1] or speaker clustering [2] constitute promising ways to perform fast speaker adaptation. Each speaker is represented relatively to other speakers.

The use of a relative position of a speaker with respect to a set of reference speakers called anchor models has already been studied in [3], [4] and [5] where a deterministic approach was proposed. The deterministic approach has a major limitation: it does not modelize the intra-speaker variability.

To overcome this limitation, we introduce a probabilistic approach which models the intra-speaker variability and then a combined approach based on the relation between deterministic and probabilistic approaches.

The concept of anchor models is described in section 2 and section 3 details the deterministic metrics used for speaker verification in the reference speakers space. Then section 4 introduces the probabilistic approach for speaker location, section 5 shows the relation between the probabilistic approach and the deterministic approach and section 6 describes the combined approach.

In the last section of the paper, all these approaches are evaluated on the NIST-2000 database.

2. Concept of anchor models

Recent research [3][4] have been oriented on a relative speaker representation. This modelling consists in projecting a speaker utterance into a space of reference speakers. The speaker is not represented in an absolute way but relatively to a set of speakers whose GMM-UBM models are pre-trained. These models are called anchor models.

The speaker is characterized by a vector defined as the set of the likelihood between the speaker data and the anchor models. This vector is called Speaker Characterization Vector (SCV) and denoted \( \tilde{X} \).

\[
\tilde{X} = \begin{bmatrix}
\tilde{s}(X|\lambda_1) \\
\tilde{s}(X|\lambda_2) \\
\vdots \\
\tilde{s}(X|\lambda_N)
\end{bmatrix}
\]

where \( \tilde{s}(X|\lambda) \) is the average log likelihood ratio of the data \( X \) (of \( N \) acoustics feature vectors) for the GMM model of the reference speaker \( \lambda \) relative to a Universal Background Model:

\[
\tilde{s}(X|\lambda) = \frac{1}{N} \log \frac{p(X|\lambda)}{p(X|\lambda_{UBM})}
\]

where \( \lambda_{UBM} \) is the Universal Background Model which has been used to initialize the training of the anchor models associated to reference speakers.

With this modelling, the speaker verification can be viewed as projecting the speaker training data and the test data into the anchor space. Then a measure between the speaker and the test is calculated and finally, the measure is compared to a threshold to decide whether the test has been uttered by the speaker or not.

3. Deterministic approach

The formerly studied metrics for SCV comparison are the Euclidean metric [3], the Angular metric [4] and the Correlation metric [5]. The efficiency of a metric depends on its capacity to be robust against the mismatch (recording condition, intra-speaker variability) between the training data and the testing data.

This section details the metrics and the kind of mismatch they are robust against.

Let \( X \) and \( Y \) two speech segments, \( \tilde{X} \) and \( \tilde{Y} \) their Speaker Characterization Vector.

3.1. Euclidean metric

\[
d(\tilde{X}, \tilde{Y}) = \sqrt{\|\tilde{X} - \tilde{Y}\|^2}
\]

with

\[
d(\tilde{X}, \tilde{Y}) = 0 \iff \tilde{Y} = \tilde{X}
\]

This metric is efficient when there is no mismatch between the training data and the testing data.
3.2. Angular metric

\[ \delta(\mathbf{e}_X, \mathbf{e}_Y) = \arccos \frac{\mathbf{X}^T \mathbf{Y}}{\| \mathbf{X} \| \| \mathbf{Y} \| } \]  

with

\[ \delta(\mathbf{e}_X, \mathbf{e}_Y) = 0 \iff \mathbf{Y} = a \mathbf{X}, \forall a \in \mathbb{R}, a \neq 0 \]  

Thanks to this property, the angular metric is robust against a mismatch modelized by a multiplicative coefficient \( a \) between the two SCV.

3.3. Correlation metric

\[ \rho(\mathbf{X}, \mathbf{Y}) = 1 - R(x, y) \]  

where \( R(x, y) \) is the correlation coefficient between the components of the two SCV (there are considered as the realization of two random variables \( x \) and \( y \)):

\[ R(x, y) = \frac{C_{x,y}}{\sigma_x \sigma_y} \]  

where \( C_{x,y} \) is the covariance between the two variables and \( \sigma_x, \sigma_y \) are respectively the standard deviation of \( x \) and \( y \).

\[ \rho(\mathbf{X}, \mathbf{Y}) = 0 \iff \mathbf{Y} = a \mathbf{X} + b, \forall (a, b) \in \mathbb{R}^2, a \neq 0 \]  

This metric is robust against a mismatch that is modelized by a positive multiplicative coefficient and an additive constant.

4. Probabilistic approach

In the preceding sections, speaker verification using anchor models by a deterministic approach was described. Unfortunately, this technique has a major limitation: it does not modelize the intra-speaker variability.

In this section, a new speaker representation based on a distribution of SCV is described. The idea is to model the various utterances of a speaker SCV by a normal distribution.

4.1. Principle

The aim of this approach is to represent a speaker by a probability density, which models its distances, to a set of reference speakers. In other words, instead of representing the location of a speaker by only one point in the representation space, he is located by a distribution.

Doing this, the compact representation of anchor models is maintained and a probability density is introduced. It means that a priori information can be used for speaker modeling and a probabilistic metric can be applied between the test occurrence and speakers models (instead of a deterministic metric).

In practice, the proposed approach consists in representing a speaker \( \lambda \) who has pronounced one or several utterances by:

\[ \lambda = \mathcal{N}(\mu, \Sigma) \]  

where \( \mathcal{N} \) is a Gaussian distribution, \( \mu \) is the mean vector and \( \Sigma \) is the covariance matrix. These parameters are estimated in the anchor space.

4.2. Estimation of speaker model parameters

Let \( \lambda \) be a speaker model which is estimated from a set of \( M \) speech segments. These segments are represented by \( M \) SCV:

\[ W = (w_1^\lambda \ldots w_M^\lambda) \]  

4.2.1. Estimation by Maximum A Posteriori

A first possibility to estimate the parameters \((\mu, \Sigma)\) consists in maximum likelihood estimation:

\[ \mu_s = \frac{1}{M} \sum_{j=1}^{M} W_{ij} \]  

and

\[ \Sigma_{ij'} = \frac{1}{M} \sum_{j=1}^{M} (W_{ij} - \mu_i) (W_{ij'} - \mu_{i'}) \]  

where \( i, i' = 1, \ldots, E \) and \( (W_{ij}) \) is the distance from speech segment \( j \) (of speaker \( \lambda \)) to reference speakers \( \lambda_i \).

However, the maximum of likelihood estimator is efficient only when enough training data is available. If this is not the case, the number of segments is not sufficient and the maximum likelihood estimate is not reliable.

To cope with this problem, a priori information is used. The speaker parameters are adapted starting from the initial parameters with a simplified version of MAP (Maximum A Posteriori) [6].

The estimation formulae of mean vector for each speaker is:

\[ \mu_s = \alpha \mu_0 + (1 - \alpha) \mu \]  

where \( \alpha \) is a parameter which controls the weight of the a priori distribution. The covariance matrix of speaker \( \lambda \) can also be estimated by MAP. However, it is more robust and simpler to choose a covariance matrix common to all speakers:

\[ \Sigma_s = \Sigma_0 \]  

4.2.2. A Priori model estimate

It consists in estimating a speaker-independent distribution \( \mathcal{N}(\mu_0, \Sigma_0) \) of SCV in the anchor models space. It is estimated from a development corpus of \( S \) speakers. These data are represented in the anchor models space: each speaker segment is represented by a SCV. So, the initial matrix \( \mathcal{W} \) corresponds to the following expression:

\[ \mathcal{W} = (w_1^{loc1} \ldots w_M^{loc1} w_1^{loc2} \ldots w_M^{loc2} \ldots) \]  

and

\[ M_0 = \sum_{s=1}^{S} M_s \]  

where the SCV \( w \) are of dimension \( E \) and \( M_1, M_2, \ldots, M_S \) are the number of segments of speaker 1, speaker 2, . . . speaker \( S \). A scatter of points is obtained which is characterized by a mean vector \( \mu_0 \) and a covariance matrix \( \Sigma_0 \). The estimate by maximum of likelihood of the mean vector is given by:

\[ \mu_{0s} = \frac{1}{M_0} \sum_{j=1}^{M} W_{ij} \]  

The covariance matrix \( \Sigma_0 \) is estimated by maximum of likelihood and it corresponds to an intra-class covariance matrix of
the initial data:

\[ \Sigma_{0i'j} = \frac{1}{M_i} \sum_{s=1}^{S} (W_{ij} - \overline{W}_{is})(W_{ij} - \overline{W}_{is}) \]  (19)

where \( S \) is the total number of initial speakers. Each class of speaker characterizes a set \( I_s \) of \( M_i \) speech segments of mean:

\[ \overline{W}_{is} = \frac{1}{M_i} \sum_{j \in I_s} W_{ij} \]  (20)

and \( i, i' = 1, \ldots, E \).

### 4.3. Application to speaker verification

A likelihood score between the test segment SCV \( \overline{X} \) and the claimed speaker model is computed. The likelihood is normalized by the likelihood between the test and the a priori distribution and then the normalized score is compared to a threshold \( \theta \):

\[ \text{score} = \log \frac{p(\overline{X} | \mu_0, \Sigma_0)}{p(\overline{Y} | \mu_0, \Sigma_0)} \leq \theta \]  (21)

and \( X = \{ \text{clients, impostors} \} \).

### 5. Relation between probabilistic and deterministic approach

The log-likelihood of a test utterance for a normal distribution is defined as:

\[ \log(p(\overline{X} | \mu_a, \Sigma_a)) = \log \frac{1}{(2\pi)^{\frac{D}{2}} | \Sigma_a |^{\frac{1}{2}}} - \frac{1}{2}(\overline{X} - \mu_a)^T \Sigma_a^{-1}(\overline{X} - \mu_a) \]  (22)

The covariance matrix is the same for all speaker models so the log-likelihood can be reduced to:

\[ D(\overline{X} | \mu_a, \Sigma_a) = (\overline{X} - \mu_a)^T \Sigma_a^{-1}(\overline{X} - \mu_a) \]  (23)

#### 5.1. Space transformation

The last equation is also equivalent to:

\[ D(\overline{X} | \mu_a, \Sigma_a) = (\overline{X}^* - \mu_a^*)^T (\overline{X}^* - \mu_a^*) \]  (24)

where \( \overline{X}^* = \Sigma_a^{-\frac{1}{2}} \overline{X} \) and \( \mu_a^* = \Sigma_a^{-\frac{1}{2}} \mu_a \).

This is a space transformation where the transformation matrix is \( \Sigma_a^{-\frac{1}{2}} \).

So the log-likelihood is the square of an Euclidean metric between the SCV of the test and the mean of the speaker in the transformed space.

#### 5.2. Adaptation

The SCV of the speaker in the new space is given by \( \mu_a^* = \Sigma_a^{-\frac{1}{2}} \mu_a \). If \( \mu_a^* \) is replaced by its expression defined by equation (16), the new mean vector becomes:

\[ \mu_a^* = \alpha \mu_0^* + (1 - \alpha) \mu^* \]  (25)

where \( \mu_0^* = \Sigma_a^{-\frac{1}{2}} \mu_0 \) and \( \mu^* = \Sigma_a^{-\frac{1}{2}} \mu \). The mean of the speaker model is adapted from the a priori mean in the transformed space.

#### 5.3. Normalization

Finally, the normalized score defined by equation (23) becomes:

\[ D = d_2(\overline{X}^*, \mu_a^*) - d_2(\overline{X}^*, \mu_0^*) \]  (26)

where \( d_2 \) is the square of the Euclidean metric and can be replaced by other metrics.

### 6. Combined approach

This section proposes a combination of the deterministic and probabilistic approaches using the metrics of the first one with the elements of the second one.

Let \( X \) and \( Y \) two speech segments, \( \overline{X} \) and \( \overline{Y} \) their Speaker Characterization Vector, \( \mu_0 \) the a priori mean and \( \Sigma_a^{-\frac{1}{2}} \) the transformation matrix.

This combination consists in 4 successive steps:

- **Space transformation**
  \[
  \overline{X}^* = \Sigma_a^{-\frac{1}{2}} \overline{X}, \overline{Y}^* = \Sigma_a^{-\frac{1}{2}} \overline{Y}, \mu_0 = \Sigma_a^{-\frac{1}{2}} \mu_0
  \]

- **Adaptation**
  \[
  \overline{X}_a^* = \alpha \overline{X}_0^* + (1 - \alpha) \overline{X}^*, \overline{Y}_a^* = \alpha \overline{Y}_0^* + (1 - \alpha) \overline{Y}^*
  \]

- **Normalization**
  \[
  D(\overline{X}|\overline{Y}) = d_2(\overline{X}_a^*, \overline{Y}_a^*) - d_2(\overline{X}^*, \mu_0^*)
  \]

- **Symmetrisation**
  \[
  D(\overline{X}, \overline{Y}) = \frac{\text{score}(\overline{X}|\overline{Y}) + \text{score}(\overline{Y}|\overline{X})}{2}
  \]

Note that the combined euclidean approach is just the symmetrisation of the probabilistic approach.

### 7. Experiments and results

#### 7.1. Evaluation database

The evaluation is done on a subset of the switchboard database, which was used for the NIST-1998, NIST-1999 and NIST-2000 evaluations restricted to the male speakers recorded with an electret microphone. The NIST database used is extracted from Switchboard-2 Phase-3 (this corpus was collected by LDC (Linguistic Data Consortium)).

This database is divided into three subsets:

- **Subset \( L_1 \)** of 215 reference speakers from NIST-1998 and NIST-1999 having more than 160 seconds of speech. This subset is used to build the anchor model space.
- **Subset \( L_2 \)** of 166 development speakers from NIST-1998 and NIST-1999 having test sentences of 30 seconds. This subset is used to estimate the a priori distribution \( N(\mu_0, \Sigma_0) \).
- **Subset \( L_3 \)** of NIST-2000. It is composed of 457 test speakers. This subset corresponds exactly to one of the test sets used in NIST-2000 evaluations (one session training (110-130 seconds), 30 seconds testing, male, electret). This is also to the test set used in [3].
7.2. System configuration

In all experiments, 13 Mel-frequency cepstral coefficients with their first and second derivatives plus $\Delta E$ and $\Delta \Delta E$ are used and the statistical models are 256-component GMMs. CMS is applied. The 215 models of the anchor models space are adapted from a UBM (trained on $L_2$) model with a MAP criterion.

7.3. Results

The performance in speaker verification for the deterministic, probabilistic and combined approaches are reported on Figure 1.

This figure can be divided in three parts corresponding to the performance of each approach. The performance with the deterministic approach (part 1) are similar to those reported in [3] with an EER in the range of [24-28]. The probabilistic approach (part 2) gives far better results (EER = 12.9%) than the deterministic approach. The combined approach (part 3) still improves performance compared with the probabilistic approach. In the case of the euclidean combined approach, the improvement is due to the symetrisation. Figure 1 also shows that performance of the combined approach are the same independently of the metric.

These results are promising, notably for speaker tracking tasks, because the improvement is more important in the low false alarm zone which is the operating zone of such a system.

Table 1 shows the influence of each step of the combined euclidean approach on the performance (0 : step not performed, 1 : step performed).

The major improvement (EER=27.9 to EER=15.9) is due to the combination of the space transformation and the normalisation. It is important to note that the transformation performed alone does not improve the performance and that the normalisation alone only improves slightly the performance.

Table 1 also shows that the performance are better when the adaptation and the symetrisation are performed.

8. Conclusions

In this article, we have investigated the concept of the relative speaker representation for speaker recognition. It consists in representing a speaker by comparing him to a set of well trained speaker models. Each speaker is represented by his location in a space of reference speakers called anchor models. We also have introduced the deterministic metrics to compare SCV in the reference space.

Then a new speaker representation based on a distribution of SCV, which modelizes the intra-speaker variability, was proposed. This allows the introduction of a priori knowledge into the parameter estimation and enables an efficient normalization step in the scoring process. A combined approach, based on the deterministic and probabilistic approaches, was also proposed. The probabilistic approach yields a visible improvement of the speaker verification performance compared with the deterministic approach and performance are further improved by the combined approach.

The probabilistic approach on anchor models is a good compromise between the speaker model complexity and the amount of training data available. This compact representation can be used in many tasks of indexation, segmentation or speaker tracking.

9. References