SPEAKER IDENTIFICATION BY LOCATION IN AN OPTIMAL SPACE OF ANCHOR MODELS

Yassine Mami and Delphine Charlet

France Télécom R&D
DIH/IPS, 2 av. Pierre Marzin
22307 Lannion Cedex, FRANCE
{yassine.mami, delphine.charlet}@rd.francetelecom.com

ABSTRACT

The process of speaker recognition is generally based on modeling the characteristics of each speaker. An interesting method for modeling consists in representing a new speaker, not in an absolute manner, but relatively to a set of well trained speaker models which constitutes the new representation space. This paper addresses the task of finding a good representation space for speaker identification. It describes a representation space built either by clustering speakers or by selecting an optimal subset of them. In this representation space, speaker location is then performed by the anchor models technique. We present experimental results and compare them with GMM-based results. We show that clustering and subset selection give good representation spaces. With a little amount of training data, identification by location in a space of virtual voices performs much better than GMM.

1. INTRODUCTION

Since the first studies dedicated to speaker recognition, many approaches were proposed in the literature - vectorial approach, connexionist, predictive, statistical, ... The statistical approach gives the best performances in automatic speaker recognition systems. Indeed, gaussian mixture models (GMM) [1] provide the best rates in text-independent recognition. Unfortunately, the performances degrade considerably if the amount of training data is insufficient. However, in most applications, the training phase must be very short (about a few seconds of speech).

To cope with this problem, an interesting method for modeling consists in representing a new speaker, not in an absolute manner, but relatively to a set of well trained speaker models. Each new speaker is represented by its location in an optimal representation space.

This paper describes a speaker recognition system where the representation space is generated either by clustering speakers or by selecting an optimal subset of them and the speaker location is realized by the anchor models technique [2].

2. SPEAKER IDENTIFICATION BY LOCATION IN AN OPTIMAL REPRESENTATION SPACE

Speaker representation by location is a new technique of speaker recognition and adaptation. One of the most interesting ways is based on the eigenvoices [3]. A speaker is represented relatively to other speakers. The principal motivation is that the dimension (number of parameters) of the absolute speaker models is very large compared to the dimension that can be reliably estimated with few training data. Thus, rather than estimating the numerous parameters of an absolute model of the speaker, we estimate a few parameters of a model relatively to other speaker models. Consequently, we don’t model absolutely but relatively to a set of reference speakers or virtual speakers.

We search to build a space of virtual speakers which are supposed to be the most representative of all speakers. Thus, each speaker model $\lambda$ is associated with a characteristic vector $w$:

$$w = \{w_e\}_{e=1, \ldots, E}$$

where $w_e$ is the characteristic of the speaker relatively to the virtual speaker model $\bar{\lambda}_e$ and $E$ is the number of virtual speakers. The speaker identification system by virtual voices is divided into two parts:

Building representation space: This stage consists in creating a virtual speaker base (or a new space of representation). Starting from GMM models of a well trained set of speaker models, we have many possibilities to build a space of reference speakers: we can either build an eigen or orthogonal space [3], cluster the speakers (see 3.1.1) or select an optimal subset of them according to a specific criterion (see 3.1.2).

Speaker recognition process which proceeds in two phases:

Training: This step consists in locating speakers to be later identified in the new space.

Testing: It also consists in locating test speakers in the same representation space. Then a distance between coordinates vectors of the speakers and an unknown test speaker is computed. The recognized speaker is the one whose reference vector is the closest to the test vector.

The speaker location can be made by a simple projection (which makes sense only if the space is orthogonal), or by maximizing the likelihood of the problem (MLED) [3] or by using the anchor models technique [2]. Each speaker is associated with a characteristic vector $w = [w_1, \ldots, w_E]^T$ which represent its coordinates in the virtual speaker space. If the space is built by eigenvoices, each speaker can be located by orthogonal projection or by MLED and his model is approximated by a weighted sum of all eigenvoices $\bar{\lambda}_e$:

$$\bar{\lambda} = \sum_{e=1}^{E} w_e \bar{\lambda}_e$$

(1)
In another approach, the speakers can also be located by the anchor models. The basic concept of anchor modeling is the representation of a target speech utterance with information gained from a defined set of speakers models.

\[
\begin{align*}
    w &= \{ w_e \} \\
    &= \left[ p(x|\lambda_1) \quad p(x|\lambda_2) \quad \ldots \quad p(x|\lambda_E) \right]^T \quad (2)
\end{align*}
\]

where \( w_e = p(x|\lambda_e) \) is the likelihood score of utterance data \( x \).

The intuitive representation of a speaker by its location in the representation space assumes that the more similar the speakers are, the closer their projection points are. Thus, to evaluate the proximity in representation space, we use a metric between the speaker coordinates. Let an unknown speaker \( X \) be represented by

\[
    w_X = \{ w_e^X \}_{e=1, \ldots, E}
\]

The recognized speaker \( \hat{R} \) is the one whose reference model

\[
    w_R = \{ w_e^R \}_{e=1, \ldots, E}
\]

gives the smallest distance i.e:

\[
    \hat{R} = \arg \min_R d(w_X, w_R) \quad (3)
\]

Alternatively, the speakers location coefficients can also lead to a better estimate of the models.

3. SPEAKER RECOGNITION SYSTEM

We propose to work on a speaker recognition system where the representation space is generated either by clustering or by subset selection and the speakers are located by the anchor models technique. Thus, our speaker identification system by virtual voices proceeds in two phases.

1. Building the representation space. Two techniques are investigated and compared in this paper: clustering and subset selection.

2. Training and testing.
   (a) Training: Location of the speakers to be recognized relatively to the anchor models.
   (b) Identification test: Location of each new speaker in the representation space. The closest reference speaker model determines the identity of the speaker. In this paper two distances will be compared.

Let us note that the first phase is an off-line processing. The following paragraphs describe each stage of our system.

3.1. Building representation space

3.1.1. Clustering speakers

The aim is to find a set of virtual speakers which is the most representative of all the speakers. The clustering technique is generally used, on the one hand, to segment a speech signal utterance so that each final cluster contains only the features of one speaker and, on the other hand, to gather the same speaker features in a cluster [4] [5]. In this paper, the speaker clustering is only used to merge two by two the closest speakers. Thus, the clustering algorithm is:

1. Initialize the distances matrix between each pair of speakers.
2. Pick the two closest speakers and merge them in a new element.
3. Build a new distance matrix and go to (2). Reiterate the process until having only \( E \) virtual speakers.

Distance computation: To calculate the distance \( d(i,j) \) between a speaker \( i \) and a speaker \( j \), we evaluate the likelihood \( p(x_i|\lambda_j) \) of acoustic features \( x_i \) from speaker \( i \) over the model \( \lambda_j \) of the speaker \( j \). Each speaker is modeled by a weighted sum of gaussians [1]. The distance is given by (after normalization):

\[
    d(i,j) = \frac{1}{N_i} \log \frac{p(x_i|\lambda_{UBM})}{p(x_i|\lambda_j)} + \frac{1}{N_j} \log \frac{p(x_j|\lambda_{UBM})}{p(x_j|\lambda_i)} \quad (4)
\]

where \( N_i \) and \( N_j \) are, respectively, the number of speech frames for speaker \( i \) and \( j \). \( \lambda_{UBM} \) is the Universal Background Model which was used to train all speakers models.

This distance, called cross likelihood ratio, is symmetric which was used to train all speakers models.

This distance, called cross likelihood ratio, is symmetric which was used to train all speakers models. The distance is given by (after normalization):

\[
    d(i,j) = \frac{1}{N_i} \log \frac{p(x_i|\lambda_{UBM})}{p(x_i|\lambda_j)} + \frac{1}{N_j} \log \frac{p(x_j|\lambda_{UBM})}{p(x_j|\lambda_i)} \quad (4)
\]

where \( N_i \) and \( N_j \) are, respectively, the number of speech frames for speaker \( i \) and \( j \). \( \lambda_{UBM} \) is the Universal Background Model which was used to train all speakers models.

The intuitive representation of a speaker by its location in the representation space assumes that the more similar the speakers are, the closer their projection points are. Thus, to evaluate the proximity in representation space, we use a metric between the speaker coordinates. Let an unknown speaker \( X \) be represented by

\[
    w_X = \{ w_e^X \}_{e=1, \ldots, E}
\]

The recognized speaker \( \hat{R} \) is the one whose reference model

\[
    w_R = \{ w_e^R \}_{e=1, \ldots, E}
\]

gives the smallest distance i.e:

\[
    \hat{R} = \arg \min_R d(w_X, w_R) \quad (3)
\]

Alternatively, the speakers location coefficients can also lead to a better estimate of the models.

3. SPEAKER RECOGNITION SYSTEM

We propose to work on a speaker recognition system where the representation space is generated either by clustering or by subset selection and the speakers are located by the anchor models technique. Thus, our speaker identification system by virtual voices proceeds in two phases.

1. Building the representation space. Two techniques are investigated and compared in this paper: clustering and subset selection.

2. Training and testing.
   (a) Training: Location of the speakers to be recognized relatively to the anchor models.
   (b) Identification test: Location of each new speaker in the representation space. The closest reference speaker model determines the identity of the speaker. In this paper two distances will be compared.

Let us note that the first phase is an off-line processing. The following paragraphs describe each stage of our system.

3.1. Building representation space

3.1.1. Clustering speakers

The aim is to find a set of virtual speakers which is the most representative of all the speakers. The clustering technique is generally used, on the one hand, to segment a speech signal utterance so that each final cluster contains only the features of one speaker and, on the other hand, to gather the same speaker features in a cluster [4] [5]. In this paper, the speaker clustering is only used to merge two by two the closest speakers. Thus, the clustering algorithm is:

1. Initialize the distances matrix between each pair of speakers.
2. Pick the two closest speakers and merge them in a new element.
3. Build a new distance matrix and go to (2). Reiterate the process until having only \( E \) virtual speakers.

Distance computation: To calculate the distance \( d(i,j) \) between a speaker \( i \) and a speaker \( j \), we evaluate the likelihood \( p(x_i|\lambda_j) \) of acoustic features \( x_i \) from speaker \( i \) over the model \( \lambda_j \) of the speaker \( j \). Each speaker is modeled by a weighted sum of gaussians [1]. The distance is given by (after normalization):

\[
    d(i,j) = \frac{1}{N_i} \log \frac{p(x_i|\lambda_{UBM})}{p(x_i|\lambda_j)} + \frac{1}{N_j} \log \frac{p(x_j|\lambda_{UBM})}{p(x_j|\lambda_i)} \quad (4)
\]

where \( N_i \) and \( N_j \) are, respectively, the number of speech frames for speaker \( i \) and \( j \). \( \lambda_{UBM} \) is the Universal Background Model which was used to train all speakers models.

This distance, called cross likelihood ratio, is symmetric which was used to train all speakers models.

This distance, called cross likelihood ratio, is symmetric which was used to train all speakers models. The distance is given by (after normalization):

\[
    d(i,j) = \frac{1}{N_i} \log \frac{p(x_i|\lambda_{UBM})}{p(x_i|\lambda_j)} + \frac{1}{N_j} \log \frac{p(x_j|\lambda_{UBM})}{p(x_j|\lambda_i)} \quad (4)
\]

where \( N_i \) and \( N_j \) are, respectively, the number of speech frames for speaker \( i \) and \( j \). \( \lambda_{UBM} \) is the Universal Background Model which was used to train all speakers models.

The intuitive representation of a speaker by its location in the representation space assumes that the more similar the speakers are, the closer their projection points are. Thus, to evaluate the proximity in representation space, we use a metric between the speaker coordinates. Let an unknown speaker \( X \) be represented by

\[
    w_X = \{ w_e^X \}_{e=1, \ldots, E}
\]

The recognized speaker \( \hat{R} \) is the one whose reference model

\[
    w_R = \{ w_e^R \}_{e=1, \ldots, E}
\]

gives the smallest distance i.e:

\[
    \hat{R} = \arg \min_R d(w_X, w_R) \quad (3)
\]

Alternatively, the speakers location coefficients can also lead to a better estimate of the models.
1. Initially we have \((n = S)\) speakers with their distance matrix (between each pair of speakers).
2. Evaluate scattering measure of all subset with \((n-1)\) speakers.
3. Eliminate (or “knocked-out”) speaker not included in the most scattered subset.
4. Return with most scattered subset of size \((n-1)\).
5. Set \(n = n-1\) and go to (2). Iterate the process until having only \(E\) speakers.

This procedure requires \(\frac{1}{2}S(S+1)\) evaluations instead of \(2^S\) when evaluating all the possible subsets.

### 3.2. Training and testing

#### 3.2.1. Anchor Models

Speakers location in the space is done by the technique of anchor models. The key idea of anchor modeling is to project a speaker into a space defined by the anchor models. In this paper, these models correspond to the space of the \(E\) virtual speakers obtained by clustering or by subset selection. Each speaker is modeled by the following vector:

\[
w = \begin{bmatrix}
\tilde{p}(x|\lambda_1) \\
\tilde{p}(x|\lambda_2) \\
\vdots \\
\tilde{p}(x|\lambda_E)
\end{bmatrix}
\]  

(7)

where \(\tilde{p}(x|\lambda_e)\) is the normalized log-likelihood of the utterance data \(x\) (of \(N\) acoustic feature vectors) knowing the GMM model of the virtual or the selected speaker \(\lambda_e\).

\[
\tilde{p}(x|\lambda_e) = \frac{1}{N} \log \frac{p(x|\lambda_e)}{p(x|\text{UBM})}
\]

(8)

#### 3.2.2. Speaker identification

After having represented all the speakers as points in the space, we evaluate the proximity between them. The nearest reference model to the unknown speaker identifies the recognized speaker. It is thus necessary to define a distance between two points representing two speakers. Let \(R\) be a speaker to be recognized and \(T\) a test speaker represented respectively by their coordinates vectors \([r_1, \ldots, r_E]^T\) and \([t_1, \ldots, t_E]^T\). We can evaluate for example the following metric:

- **Hamming distance**: \(d_1(R, T) = \sum_{i=1}^{E} |r_i - t_i|\).
- **Euclidean distance**: \(d_2(R, T) = \sqrt{\sum_{i=1}^{E} (r_i - t_i)^2}\).
- **Max distance**: \(d_{\infty}(R, T) = \max_{i=1,\ldots,E} |r_i - t_i|\).
- **The angle between coordinates vectors of both speakers**: \(\delta(R, T) = \arccos \left[ \frac{r^T_t}{\sqrt{r^T r \cdot t^T t}} \right] \).

In our experiments, we have tested all these metrics, only the Hamming distance and the angle lead to good speaker identification performance. Thus, only results obtained with them will be presented in the next section.

### 4. EVALUATION

#### 4.1. Experimental context

This section presents the experimental evaluation of the text-independent speaker identification by location. The acoustic space vectors are composed of 27 coefficients. The acoustic vectors include energy and the first 8 MFCC, plus their first and second derivatives. The speakers are modeled by 16 gaussians adapted from the UBM model. GMM models are trained using an incremental training procedure [7]. It is an iterative algorithm which adapts a reference model with the training data of the given speaker. In our experiments, we have used a France Telecom R&D telephone speech database which comprises 550 speakers. This database is divided into two subsets:

- Subset \(E_1\) of 500 speakers used for creating the representation space. For each one, we have only one call. Its approximate duration is about 100 seconds.
- Subset \(E_2\) of 50 speakers to be identified composed of 33 female and 17 male speakers. For each speaker of this subset, 25 sentences recorded during one single call are reserved to the training step and 125 sentences recorded during 25 calls (5 sentences/call) spanning over several months are reserved for test evaluation.

We make one test per sentence that makes more than 6000 tests and the 95\% confidence interval is equal to \(\pm 1.26\%\). The UBM model was trained on all speakers of the subset \(E_1\). In our experiments, the representation space is built either by clustering or selection based on knock-out procedure and the proximity between speakers is evaluated by two metrics: the Hamming distance and the angle between the coordinates vectors.

#### 4.2. Results and discussion

##### 4.2.1. Influence of number of virtual speakers

Figure 1 shows the identification performance of 50 speakers of \(E_2\) versus representation space size. The 50 speaker models were trained on about 6 seconds of speech (from 25 sentences). For all curves, the identification performance increases significantly from 10 to 50 space speakers. Then, we notice an area where the correct identification performance is relatively stable until 400 speakers. In this area, where the identification system reaches best performances there is an optimal value of space size. Above this area, the correct identification rates decrease down to a value close to 40\%. These experiments show that the angle evaluated between speaker vectors seems to be the best distance to discriminate between them. Indeed, we reach 60\% of correct speakers identification. This metric is specially interesting if the representation space is generated by the knock-out selection procedure. In these conditions (6 seconds of training data), the GMM modeling leads to only 41.5\% of correct identification.

##### 4.2.2. Influence of the amount of training data

Figure 2 shows the correct identification rate of 50 speakers of \(E_2\) versus amount of training data for an optimal representation space of 200 virtual speakers obtained either with clustering or subset selection. We compare GMM performance to speaker identification by location.

This figure gives an overview of identification performance variations that we can observe when the amount of training data varies.
As we already noticed in the last section, the angle is the best discriminating metric between speakers. This distance used in a space obtained by the knock-out selection procedure leads to good speaker identification performance. In this case, we distinguish two areas: if the amount of training data is lower than 30 seconds, identification by location performs much better than identification by GMM. Above this value, the correct identification rates obtained with the GMM technique increase significantly while those obtained by location stay relatively stable.

5. CONCLUSION

In this paper, we have addressed a speaker identification system based on anchor models location. We proposed two approaches for building the representation space, either by clustering or by selecting a subset of speakers. For selection, we have used a knock-out procedure to find the most scattered speakers. It gives the best representation space. Both approaches show that there is an optimal representation space. To evaluate the proximity between speakers, the angle between their vectors seems be the best distance. With a little amount of training data, our experiments show also that speaker identification based on location on a representation space performs much better than GMM.

6. REFERENCES