Topological representation of speech for speaker recognition

Gabriel H. Sierra[1,2]*, Jean-F. Bonastre*[2]*, Driss Matrouf*[2]*,Jose R. Calvo*[1]

[1]Advanced Technologies Application Center, Havana, Cuba
[2]University of Avignon, LIA, Avignon, France
{gsierra,jcalvo}@cenatav.co.cu, {jean-francois.bonastre,driss.matrouf}@univ-avignon.fr

Abstract

During last decade, researchers in speaker recognition have been working over the same acoustic space, regardless of whether the data lie on a linear space or not. Our proposal is to take into account the inner geometric structure of the speech in order to obtain a new space with a better representation of the speech data. A topological approach based on manifolds obtained thanks to Laplacian and Isomap algorithms is proposed. In this first work, the proposal is evaluated in terms of dimension reduction of the supervector space, known to have a high redundancy. The experiments are done in the NIST-SRE framework. It appears that the proposed approach allows to reduce by a factor four the dimension of the supervector space without losses in terms of EER. This first result highlights the potential of topological approaches for speaker recognition.

Index Terms: speaker recognition, topological information, dimension reduction

1. Introduction

During the last decade researchers in the field of text independent speaker recognition have developed statistical classifiers, discriminatory classifiers and mixtures of them. If a large spectra of classification approaches was explored, it is interesting to notice that few works was proposed on the data representation, the main part of the works are using the same representation of the acoustic space, assuming that this space is linear.
Jansen and Niyogi [1] shown recently that acoustic features lie in a low-dimension manifold that is embedded in an acoustic space of high-dimension. A low-dimension manifold can have a highly non-linear structure that linear methods would not be able to discover. Based on this, we assume in this work that the acoustic sounds lie on a high-dimension manifold, so if the range of the acoustic sounds of the human voice is a subset of all sounds, therefore it lies on a low-dimensional submanifold of the high-dimension space of all possible sounds.

In the field of speaker recognition, the assumption described above related to the space of acoustic features, have been poorly used. This work aims to investigate this avenue. It is developed in the space described by mean vectors gathered from GMMs-UBM [2], the so-called GMM supervector space. This space corresponds to the acoustic features space as these mean vectors are the centers of the acoustic classes.

The main objective of the presented work is to evaluate the effectiveness of an non linear topological representation of the supervector space, in terms of dimensionality reduction. The supervector space has usually a very high dimensions (thousands of coefficients) and a large redundancy between these dimensions. We propose to extract the topological structure of the data, using manifolds obtained thanks to Laplacian and Isomap algorithms, and to use this representation in order to propose a dimensionality reduction of the supervector space.

This paper is organized as follows. In section 2 a description of GMM-UBM-SVM speaker recognition method is done. Section 3 presents two classical manifold learning algorithms: Isomap and Laplacian. Section 4 describes the nature of the supervectors for speaker recognition. In section 5 the experimental protocols and the corresponding results are shown. Finally section 6 presents some conclusions.

2. Text-independent speaker verification method

Gausian Mixture Models (GMM) method for text-independent speaker verification [3], is based on statistical theory and has become a reference and is the most widely used method. A Universal Background Model (UBM) is usually trained from a very large set of development data comprised of various speakers, channel types and number of sessions. Target speaker models can be obtained by adapting the UBM according to their corresponding enrollment data. Support vector machine (SVM), were first proposed in the early 1990s as optimal margin classifiers in the context of Vapnik's statistical learning theory [4]. Recently SVM has become an alternative of GMM and is widely used as a discriminative classification technique in the field of speaker recognition. SVM systems can obtain comparable performances to GMM-UBM systems with relatively moderate computation complexity [5]. A new SVM-based speaker verification strategy using GMM-UBM supervectors has been proposed by Campbell [6] to combine the advantages of the two systems. These *CD*-dimensional supervectors are obtained by stacking the D-dimensional mean vectors of a *C*-component adapted GMM, which are used as inputs to SVM.

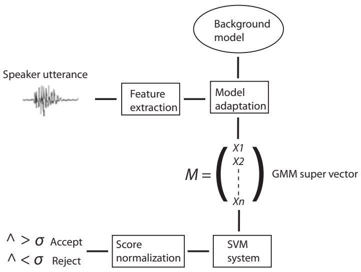

Figure 1: *Structure of a GMM-SVM system.*

A GMM-SVM system is a combination of a GMM-UBM and a SVM system. The GMM-UBM system serves as a feature extractor for the attached SVM system. SVM is trained for each target speaker using the GMM supervector of the speaker's enrollment utterances as positive samples, and GMM supervectors of all utterances from background speakers as negative samples. For classification, a class decision is based upon whether the score value is above or below a threshold.

3. Classical Manifold learning algorithms

A series of manifold learning algorithms (also known as non-linear dimensionality reduction algorithms) have been proposed [7], which go beyond the limitations of linear methods. The steps followed by the manifolds learning algorithms (Isomap and Laplacian Eigenmaps) begin creating a graph and then they use the information contained in the graph to reduce the dimensionality. Let us assume that we have a dataset in a *N-by-D* matrix *X* consisting of *N* data vectors $\overline{X}_i$, *(i = 1, 2, . . . , N)* with dimensionality *D*, besides this dataset has intrinsic dimensionality *d* (where $d < D$, and often d $\ll$ D). Here, in mathematical terms, intrinsic dimensionality means that the points in dataset *X* are lying on or near a manifold with dimensionality *d* that is embedded in the *D*-dimensional space.

3.1. Isometric feature mapping: "Isomap"

The key assumption made by Isomap [7] is that the distance along the curve between two points is not the straight line that connects them, but the shortest path through the points on the curve that connect them. The basic idea is to construct a graph whose nodes are the data points, where a pair of nodes are adjacent only if the two points are close in R^D (*D* is the dimension of the data), and then to take the geodesic distance along the manifold between any two points as the shortest path in the graph and finally to use multidimensional scaling "MDS" -a classical method for embedding dissimilarity information into metric space [7]- to extract the low dimensional representation (as vectors in R^d, $d \ll D$).

3.2. Laplacian Eigenmaps

Laplacian Eigenmap algorithm, proposed by Belkin and Niyogi [8], is based on ideas from spectral graph theory. This work establishes both a unified approach to dimension reduction and a new connection to spectral theory. We describe the Laplacian Eigenmap for discrete data.

Again, we consider *N* vectors, $\overline{X}_i$, (*i = 1, 2, ..., N*), in the *D*-dimensional data space. For each vector $\overline{X}_i$, let us suppose a neighbor vector set N_i is computed. A graph identical to the graph in ISOMAP can be defined. We defined for any pair of connected points $\overline{X}_i$ and $\overline{X}_j$, a "local similarity" matrix *W* which reflects the degree to which points are near to one another. There are two choices for *W*:

1. $W_{i,j} = 1, if \overline{X}_j$ is one of the *k*-nearest neighbors of $\overline{X}_i$; *0* otherwise (simple weight scheme).
2. $W_{i,j} = e^{\frac{-\|\overline{X}_i - \overline{X}_j\|^2}{2\sigma^2}}$, for neighboring nodes; *0* otherwise. This is the Gaussian heat kernel. On the other hand, use of heat kernel requires σ manual setting, so is considerably less convenient than the simple weight scheme.

Let *D* denote a diagonal matrix such that $D_{i,i} = \sum_j W_{i,j}$. Let *W* denotes the symmetric matrix with entries, $1 \leq i, j \leq N$. Finally, given a graph and a matrix of edge weights, *W*, the Laplacian graph is defined as $L = D - W$. Consider the solutions to the problem:

$$Lf = \lambda Df \tag{1}$$

where $f \in \Re^N$. Let $f_0, f_1, , f_{k-1}$ be the eigenvectors with corresponding eigenvalues $0 = \lambda_0 \leq \lambda_1 \leq \lambda_2 ... \leq \lambda_{k-1}$;

$$Lf_0 = \lambda_0 Df_0$$
$$Lf_1 = \lambda_1 Df_1$$
$$\vdots$$
$$Lf_{k-1} = \lambda_{k-1} Df_{k-1}$$

The eigenvector associated with zero eigenvalue (λ_0) is left out and the next *m* eigenvectors are used for the embedding in an *m*-dimensional Euclidean space.

The eigenvalues and eigenvectors of the Laplacian reveal a wealth of information about the graph such as whether it is complete or connected. Here, the Laplacian will be exploited to capture local information about the manifold.

4. Nature of the supervector for speaker recognition

The supervectors are built from the concatenation of the mean of Gaussian components for each speaker model, represented as a means matrix M_i, which is concatenated for columns, as follows:

$$M_i = \left\{ \begin{array}{cccc} \overline{x_1} & \overline{x_2} & \cdots & \overline{x_C} \\ \downarrow & \downarrow & & \downarrow \\ x_{1,1} & x_{1,2} & \cdots & x_{1,C} \\ \cdots & \cdots & \cdots & \cdots \\ x_{D,1} & \cdots & & x_{D,C} \end{array} \right\}$$

$$SV_i = \{x_{1,1}, x_{2,1}, ..., x_{D,1}, x_{1,2}, ..., x_{D,2}, x_{1,C}, ..., x_{D,C}\}$$

where *SV* is the mean supervector, *i* represents the index of each speaker, *D* is the dimension of each Gaussian component and *C* is the number of Gaussian components of GMM.

Note that as a result of this construction, each mean vector of the Gaussian components defines a specific set of dimensions in the SV and the union of all mean vectors defines the place of the point in a high-dimensional space.

This process is only a transformation of the model mean matrix to a SV with *DC*-dimension, converting the matrix into a point on a high-dimensional space which will be used later to classify the speaker through a SVM classifier. Note that the order in which mean vectors of the Gaussian component are chosen in the construction of the SV is not important if the same order is always used for all speakers.

To better address future work, Fig. 2 illustrate, from another point of view a representation of the construction of the SV in a two dimensional space, taking into account that each column *c = (1,2,...,C)* in the M_i matrix corresponds to each Gaussian component $\overline{X}_c$.

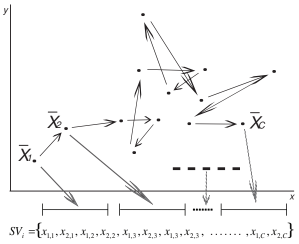
y
$\overline{X}_2$
$\overline{X}_c$
$\overline{X}_1$
x
$SV_i = \{x_{1,1}, x_{2,1}, x_{1,2}, x_{2,2}, x_{1,3}, x_{2,3}, x_{1,3}, x_{2,3}, \ldots\ldots, x_{1,C}, x_{2,C}\}$

Figure 2: *Each point represents the firsts two coefficients of a Gaussian component of a speaker, which defines an acoustic class.*

Fig. 3 illustrate, in the same two dimensional representation, the distribution of Gaussian components of several speakers, the difference between colors representing different acoustic classes, each dot is an acoustic class in each speaker.

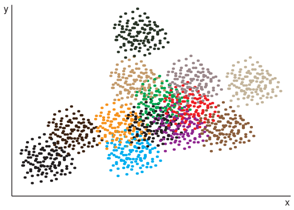

Figure 3: *Firsts two coefficients of the Gaussian components of several speakers.*

Researchers in the field know that there is a lot of redundant information in the speaker acoustic models obtained using the adapted GMM-UBM, which can be observed easily increasing the amount of mixtures in the Universal Background Model and as result we will obtain an improvement in the EER. This is adversely not proportional to the big size reached by the SV of each speaker, which often generates more problems than the small benefits in percentage of EER improvement achieved. If we analyze the Fig. 3 which represents a similar nature to the space where acoustics classes lie, we can see that at the centre of the cloud of dots is where we have the highest overlapping among acoustic classes and the bigger the number of acoustic classes the smallest the distance at the centre. The distance between points in the center of the cloud decreases, increasing overlap in the Gaussian components of these points, in addition to this analysis we also rely on our earlier work, see [9]. So if we use the topologic information of acoustic classes we can reduce the dimension of these classes and the supervector size, without greatly affecting the outcome of the classifier.

5. Algorithms and Experiments

Experiments using NIST'04 [10] for background data (UBM), 1348 male speakers of NIST'05 for development data and 380 speakers of Fisher'04 for impostor data were used for training the SVM classifier, in order to evaluate the algorithm. The male section of the NIST'05 primary task is used for development, for this condition, one side of a 5-min conversation is given for testing and the same amount for training.

The realization has taken place in the context of the open-source ALIZE toolkit [11] and the algorithms Isomap y Laplacian on Matlab. Performances are assessed using DET plots and measured in terms of equal error rate (EER). The cost function is calculated following NIST criteria [10].

Baseline experiment: speaker recognition experiment trained and tested with spontaneous sentences. Models were obtained from target data, impostors data and test data using GMM-UBM adaptation. *DC*-dimensional input supervectors are obtained stacking means of the Gaussian components. SVM classifier is trained with target supervectors and used to score the test supervectors.

The experiment consists in the performance evaluation of speaker recognition using a new speaker representation in a low dimensional space, obtained by the Isomap and Laplacian Eigenmaps algorithms, of the mean matrix of Gaussian component in a closed set of speakers. These mean matrixes M_i of target, impostor and test data were obtained by GMM-UBM adaptation, as baseline. Input supervectors are obtained stacking means of the mixture components in a new space. SVM algorithm was used for classification, as baseline.

The experiment involves a global analysis of the initial space where the acoustic classes lie, but at the same time a reduction of dimension. Otherwise, we assume that all the acoustic classes lie in a manifold and we want to find the geometric information of the topological structure of these acoustic classes that better characterizes the speaker, but we will work assuming that each point in this space will be defined by the number of Gaussian components.

The experiment consists of the following steps:

1. Models are obtained from target data, impostors data and test data using GMM-UBM adaptation.

$$M_i = \left\{ \begin{array}{cccc} \underline{\overline{x}_1} & \underline{\overline{x}_2} & \cdots & \underline{\overline{x}_C} \\ \downarrow & \downarrow & & \downarrow \\ x_{1,1} & x_{1,2} & \cdots & x_{1,C} \\ \cdots & \cdots & \cdots & \cdots \\ x_{D,1} & \cdots & & x_{D,C} \end{array} \right\}$$

   $S = \{M_1, M_2, ..., M_N\}$ Speakers Mean matrixes.
   where M_i, *i=1,2,...,N* are the mean matrices of *N* speakers, *D* is the dimension of each Gaussian component and *C* is the number of Gaussian components of GMM.

2. Then, we construct an initial space for *C* Gaussian components of all speakers for its *D* dimension.

$$A^d_{N,C} = \left\{ \begin{array}{cccc} \underline{1} & \underline{2} & \cdots & \underline{C} \\ \downarrow & \downarrow & & \downarrow \\ M_1(d,1) & M_1(d,2) & \ldots & M_1(d,C) \\ \ldots & \ldots & \ldots & \ldots \\ M_N(d,1) & \ldots & \ldots & M_N(d,C) \end{array} \right\}$$

   $A^d_{N,C}$ will have a *NC*-dimension (number of speaker by number of Gaussian components), and *d = 1,2,...,D* will be the amount of spaces constructed. For each *d* we will make a reduction taking into account the topologic information in each space.

3. For each submanifold $A^d_{N,C}$ we propose to obtain a new representation in one linear space F:$R^C \rightarrow R^G$ where the topologic information will intervene in the description of this means. For this, we use Isomap and Laplacian Eigenmaps algorithms to obtain a new projection for each matrix, $F(A^d_{N,C}) = A^d_{N,G}$, where *F* is the algorithm used and *G* represents the new smaller dimension, leading to a reduction in the number of Gaussian components for each model $G \ll C$.

4. Later, each speaker models matrix is reassembled from the new space, and *DG*-dimensional input SV is obtained stacking the vectors for each new speaker matrix. These SV will participate in the SVM training as well on the SVM test, for each corresponding speaker.

As a result, dimensionality reduction facilitates not only classification, but also compression of high-dimensional data.

The DET curve are shown in Fig. 4 and 5, the parameters of the first experiment are:

Baseline, *C* = 512 Gaussian components, *D* = 50 dimension and *DC* = 25600 dimensions for each speaker supervector.

Isomap reduction, number of neighbors *k* = 12, number of dimension *D* = 50, number of Gaussian components *C* = 512, as a result we will have *G* = 128 Gaussian components with the same *D*-dimension and *DG* = 6400-dimensions for each speaker supervector.

The parameters of the second experiment are:

Baseline, *C* = 128 Gaussian components, *D* = 50 dimension and *DC* = 6400 dimensions for each speaker supervector.

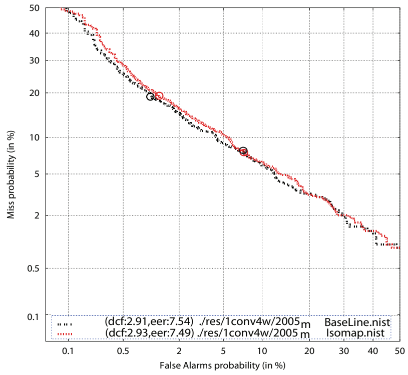

Figure 4: *EER and minDCF performances for BaseLine and using the Isomap algorithm.*

PCA reduction, number of dimension D = 50, number of Gaussian components C = 128, as a result we will have C = 64 Gaussian components with the same dimension and DG = 3200-dimensions for each speaker supervector.

Laplacian reduction, number of neighbors k = 12, number of dimension D = 50, number of Gaussian components C = 128, as a result we will have G = 64 Gaussian components with the same D-dimension and DG = 3200-dimensions for each speaker supervector. Fig. 5 shows DET curve of experiment.

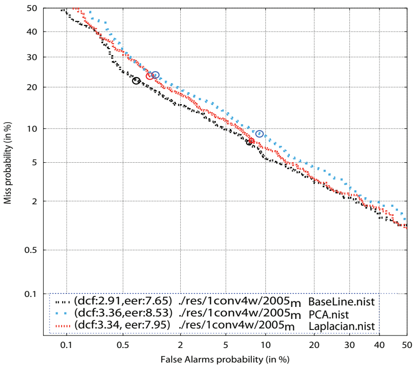

Figure 5: *EER and minDCF performances for BaseLine and using the PCA and Laplacian algorithm.*

First experiment using the Isomap, reflects very similar results as the baseline, 7.54% ERR (baseline) vs 7.49% EER (Isomap) and 2.91% DCF (baseline) vs 2.93% DCF (Isomap).

Second experiment using the Laplacian algorithm, reflects similar results as the baseline, 7.65% ERR (baseline) vs 7.95% EER (Laplacian) and 2.91% DCF (baseline) vs 3.34% DCF (Laplacian). The PCA shows the worst results of the experiment, 8.53% EER and 3.36% DCF.

In EER and DCF we did not obtain an improvement, but in this experiment we focus on reducing the dimension of the supervectors, which were managed successfully showing the large amount of redundant information that exists in the original supervectors. If we compare the SV dimensions in the first experiment, the baseline has 25600 dimensions and in our proposal it has 6400 dimensions that is 1/4 of the original size, which is a significant reduction keeping the same EER. In the second, the baseline has 6400 dimensions and in our proposal it has 3200 dimensions that are half the size, which is a significant reduction too, although in this experiment, the EER is 0.3 % worse.

6. Conclusions

This paper presents one of the first steps in the use of topological approaches for speaker recognition. In order to evaluate the effectivness of such an approach in this field, we focused on the supervector space and tried to reduce the dimension of this space. For that, the topological information present in the data is represented by manifolds gathered thanks to Laplacian and Isomap algorithms. The topological information is used in order to define a new space for the GMM supervectors, with a lower dimensionality than the initial one.
The experimental validation took place in the NIST-SRE framework, using the MISTRAL/ALIZE open source system and MatLab for Isomap and Laplacian algorithms. The results showed clearly the interest of topological approaches for speaker recognition, as a dimensionality reduction of a factor four (from 25600 to 6400) was possible without any loss in terms of speaker recognition performance. Moreover, the Laplacian non-linear approach outperformed a classical linear PCA solution, demonstrating that the non-linear inner nature of the data is well handled by the proposed topological approach. These results showed also the large amount of redundant information embedded in the GMM supervectors. Looking at these encouraging results, we aim to investigate more deeply the use of such an approach for speaker recognition in the next months. It seems particularly interesting to us to take into account the geometric structural information in order to characterize the speaker voices.

7. References

[1] Aren J. and Partha N., ”A Geometric Perspective on Speech Sounds”, Tech. Report TR-2005-08. Computer Science Dept., Univ. of Chicago, 2005.

[2] Bimbot, F., Bonastre, J.-F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-Garcia, J., Petrovska-Delacretaz, D., Reynolds, D., 2004. A tutorial on text-independent speaker verification. EURASIP J. Appl. Signal Process. 4, 430451.

[3] Reynolds, D., Quatieri, T., Dunn, R., ”Speaker verification using adapted gaussian mixture models”, Digital Signal Process. 10 (1), 19-41, 2000.

[4] Vapnik V N., ”Statistical Learning Theory”, New York, USA: Wiley, 1998.

[5] Campbell W M., ”Generalized linear discriminant sequence kernels for speaker recognition”, In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Orlando, USA, 161-164, 2002.

[6] Campbell, W., Sturim, D., Reynolds, D., ”Support vector machines using GMM supervectors for speaker verification” IEEE Signal Process. Lett. 13 (5), 308-311, 2006.

[7] J. B. Tenenbaum, V. de Silva, and J. C. Langford, ”A global geometric framework for nonlinear dimensionality reduction”. Science, vol. 290, pp. 2319-2323, 2000.

[8] M. Belkin and P. Niyogi, ”Laplacian eigenmaps and spectral techniques for embedding and clustering”, Advances in Neural Information Processing Systems. vol. 14, pp. 585-591, MIT Press, 2002.

[9] G. Hernandez, J. R. Calvo, F. J. Reyes and R. Fernndez, ”Simple Noise robust feature vector selection method for speaker recognition”, LNCS vol 5856 pp 313-320 ISBN: 978-3-642-10267-7, 2009.

[10] A. Martin and M. Przybocki, ”NIST speaker recognition evaluation chronicles,” in Proc. Odyssey, pp. 15-22, 2004.

[11] J.-F. Bonastre, F. Wils, and S. Meignier, ”ALIZE, a free toolkit for speaker recognition,” in Proc. ICASSP, pp. 737-740, 2005.