AUTOMATIC LOW-DIMENSIONAL ANALYSIS OF AUDIO DATABASES

José Anibal Arias, Régine André-Obrecht, Jérôme Farinas

IRIT/SAMOVA, Université Paul Sabatier
118 route de Narbonne, 31062 Toulouse, France
{arias,obrecht,jfarinas}@irit.fr

ABSTRACT

In this paper we present an approach designed to map variable size audio sequences into fixed-length vectors, useful to discover contents of audio databases. First, we model standard audio parameters with Gaussian mixture models (GMM). Then, symmetric Kullback-Leiber divergences between models are approximated with a Monte-Carlo method. We use these statistical dissimilarities to find a low-dimensional representation of each audio sequence through Multidimensional scaling (MDS) algorithm. Vectors in low-dimensional spaces are then easily explored with kernel and clustering methods. Experiments carried out in different kind of audio databases (music, speakers and languages) show good potential of the proposed approach and provide a framework for more challenging applications.

Index Terms— Unsupervised learning, audio databases, machine learning

1. INTRODUCTION

Audio data collections are continuously growing in recent years, and different automatic procedures have been proposed to structure their contents [11]. Unfortunately, due to the complexity of this data type most of the procedures are ad hoc, and parallel architectures as well as exotic features have been explored to improve systems performance. But we still lack of general methods to explore and classify automatically audio contents.

Current research in audio processing focuses in efficient music/speech discrimination [3], automatic music classification [4], language [10] and speaker identification systems [7]. State-of-the-art systems perform supervised tasks and use Hidden Markov Models, Gaussian Mixture Models, and Kernel methods in the classification step, or mixtures of them [5]. We propose an unsupervised framework which utilizes GMM to deal with variable-length features, MDS to visualize distances between GMM and Kernel methods to separate clusters. The objective is to discover clusters in audio databases. The main contribution of the paper is to define a simple and general approach for mapping audio sequences to low dimensional vectors. Only slight modifications are needed in features and parameters to custom specific applications.

We organize the paper as follows: section 2 provides a brief overview of the audio data we are processing. Section 3 describes the system and reviews the main algorithms we use. Section 4 describes the data used for experiments and shows some results of speech-music discrimination, speaker identification and language classification tests. Section 5 discusses conclusions and future work.

2. AUDIO DATA AND MACHINE LEARNING

Confronted with audio data, machine learning procedures have several challenges. For example, the quantity and size of files to be processed (audio databases can store thousands of records). While current similarity measures rely on fixed length vectors, we need very often to compare audio sequences that differ considerably in length.

Feature extraction is an open topic in audio processing. Several specialized extraction methods [11] have shown their performance on specific tasks and datasets, but it is difficult to find an unified framework which links feature extraction with a classification method that works well for general purposes. It is also very difficult to organize audio data issued from different sources or according with some user's preferences. In order to delimitate the problem, we use standard feature extraction techniques and we focus on the mapping of acoustic sequences to fixed low dimensional vectors.

3. PROPOSED SYSTEM AND ALGORITHM DESCRIPTIONS

The proposed framework is described in figure 1. First we compute the cepstra and delta cepstra features of each sequence in the database. These features train a GMM per sequence. When all GMM have been trained, Kullback-Leiber divergence between each pair of pdf's is calculated. All these dissimilarities are grouped in a square matrix, which is submitted to MDS. At the output of MDS we obtain low-dimensional vectors that represent every audio sequence in the database. Although this representation is already interesting, we can further apply Kernel PCA to achieve better class separation. Cluster analysis is performed

using the average linking agglomerative method. It can be visualized (and optimized) by the user, which is an advantage against other state-of-the-art methods, like HMM, Fisher kernel or neural networks.

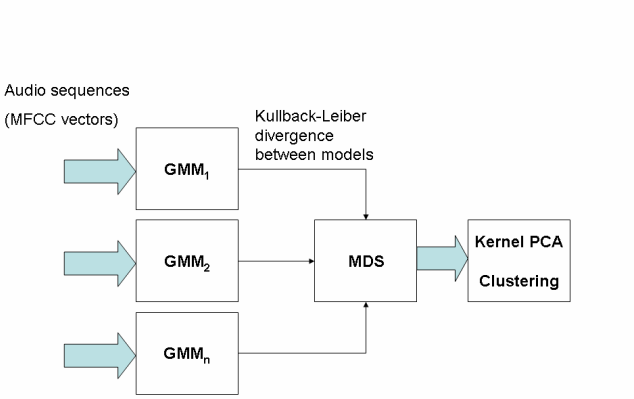

Figure 1. Proposed framework. GMM model the distribution of feature vectors. MDS finds a coordinate system for each audio sequence using the statistical distance among GMM. Class separation properties of Kernel PCA are used for clustering.

This procedure transforms variable size sequences of d-dimensional acoustic features **x** into 3-dimensional vectors **y** in Euclidian space.

3.1 Gaussian mixture models

GMM represent multimodal densities [1]. They are formed by summing weighted multivariate Gaussian density functions as,

$$f(x) = \sum_{k=1}^{K} p_k N_k(x) \qquad (1)$$

where K is the mixture order and p_k is the mixture weight for a multivariate Gaussian component density,

$$N_k = \frac{1}{(2\pi)^{D/2}|\Sigma_k|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)'\Sigma_k(x-\mu_k)} \qquad (2)$$

A GMM is defined by the weights, **p**, mean vectors **μ**, and covariance matrices Σ. These parameters are represented by λ, $\lambda = \{p_k, \mu_k, \Sigma_k\}, k = 1, \ldots, K$.

The statistical dissimilarity d_{ij} between two GMMs can be estimated using a Monte-Carlo sampling approach for approximating the Kullback-Leiber symmetric divergence between models, which is defined as the sum of two oriented divergences $KL(GMM_i/GMM_j)$ and $KL(GMM_j/GMM_i)$,

$$d_{ij} = \frac{1}{2}(KL(GMM_i / GMM_j) + KL(GMM_j / GMM_i)) \qquad (3)$$

To calculate $KL(GMM_i/GMM_j)$, first we generate a stochastic sample X from GMM_i, and then we compute the mean of the log rate between the likelihood of X in models GMM_i and GMM_j.

3.2 Multidimensional scaling

MDS states that from a dissimilarity measure or distance between points, we can find a coordinate system preserving the original input distances [2]. MDS computes a dot product in terms of the distance between points. As this relation is valid only when vectors are centered, the dot product expression depends on the distances among all the examples,

$$y_i \cdot y_j = -\frac{1}{2}(d_{ij}^2 - \frac{1}{n}\sum_{i=1}^{n} d_{ij}^2 - \frac{1}{n}\sum_{j=1}^{n} d_{ij}^2 + \frac{1}{n^2}\sum_{i,j=1}^{n} d_{ij}^2) \qquad (4)$$

The distances d_{ij}, *i,j=1, ..., n* are the dissimilarities obtained with eq. (3) for each pair of GMM. If we form a (*n x n*) dot product matrix $Y = U\Lambda U'$ using eq. (4), the q principal eigenvectors and eigenvalues $y_{iq} = \lambda_q^{1/2} \upsilon_{iq}$ of this matrix can be used as an approximation of $y \in \Re^q$ that respect the original distances. In this manner we can find a q-dimensional vector y_i representing the i-th GMM. If q = 3, y_i can be represented visually.

3.3 Kernel PCA

Kernel PCA [8] performs principal component analysis in feature space by the mean of a kernel function; $k(y_i, y_j) = \langle \Phi(y_i) \cdot \Phi(y_j) \rangle$. The implicit non linear transformation $\Phi(y)$ allows Kernel PCA to find a subspace that seems to be the result of an information extraction process. Eigenvectors l^p in feature space lie in the span of $\alpha_j^p \Phi(y_j)$. Weights α_j^p are obtained from the relationship between eigenvalues and eigenvectors of kernel and covariance matrices of centered data. For the dataset Y = ($y_1, \ldots, y_n$), we first compute $\alpha_j^p = \nu_j^p / (\sqrt{\lambda_j^p})$, where ν_j and λ_j represent the p principal eigenvectors and eigenvalues of the kernel matrix $K = k(y_i, y_j)_{ij=1}^{ij=n}$. Projections of test vectors $\Phi(y)$ onto the principal eigenvectors are computed using a kernel function,

$$u_p' \Phi(y) = \left\langle \sum_{i=1}^{n} \alpha_i^p \Phi(y_i), \Phi(y) \right\rangle = \sum_{i=1}^{n} \alpha_i^p k(y_i, y) \qquad (4)$$

4. EXPERIMENTS

Our test platform consists of more than 300 audio files. Speech files are issued from the OGI-MLTS (3 languages) and ANITA (6 speakers) corpus. Music files are extracted

form a personal music database with different styles. The OGI speech corpus [6] consists of spontaneous telephonic speech sequences (~40 seconds each) sampled at 8 kHz. Speaker database ANITA consists of short (~7 seconds) 'phonetically balanced' speech sequences sampled at 16 kHz. Music corpus contains singing voice, rock and jazz sampled at 16 kHz presented in sequences of around 1 minute.

We perform MFCC analysis on these files using 20ms frames of signal with 10ms overlap across frames; the features consist of 15 cepstra coefficients and their respective derivatives.

4.1. Speech-music classification

For the first test we mix 10 singing voice files, 30 rock/jazz music sequences and 30 OGI speech files (15 in English and 15 in German). For speech-music discrimination, expect to find two clusters, so after the sequence-to-vector map we let the clustering algorithm to automatically discover two classes (this information is the stopping criteria in average linkage clustering).

The result is two well defined clusters in three dimensions, as we show in figure 2. One consists of music and singing voice and the other is composed only of speech files. The fact that singing voice is considered music instead of speech is not a surprise; it has been already stated in different applications.

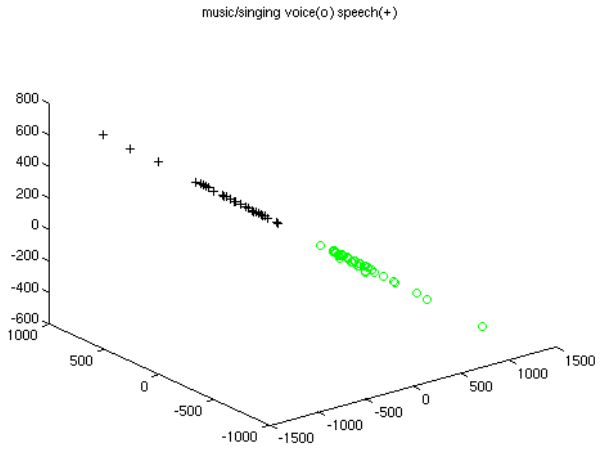

Figure 2. Speech-music discrimination. Two well defined clusters are automatically discovered, one corresponds to music and singing voice and the other to speech. Each point y in the figure represents an audio sequence in R^3.

4.2. Speaker identification

For speaker identification tests we use 180 files recorded by 6 speakers (3 men and 3 women) obtained from ANITA database.

The MDS mapping gives the low-dimensional sequence representation displayed in figure 3. Using original labels we can observe the natural distribution of 6 consistent clusters corresponding to each speaker.

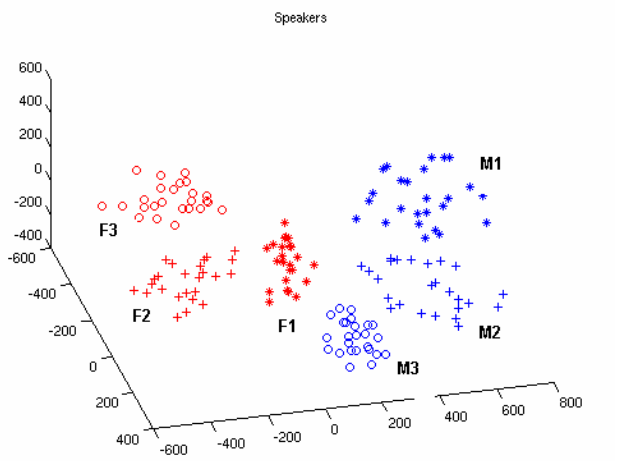

Figure 3. Supervised representation of 180 sequences issued from 6 speakers. Original labels are used to present 3 male and 3 female speakers.

To perform automatic clustering on this data, we use the mean distance between vectors as stop criteria in the agglomerative clustering. In this manner we detect 5 classes, which represent 83.33% of accuracy (figure 4). Kernel PCA compact the clusters and facilitates the stop criteria, but do not increases accuracy (figure 5).

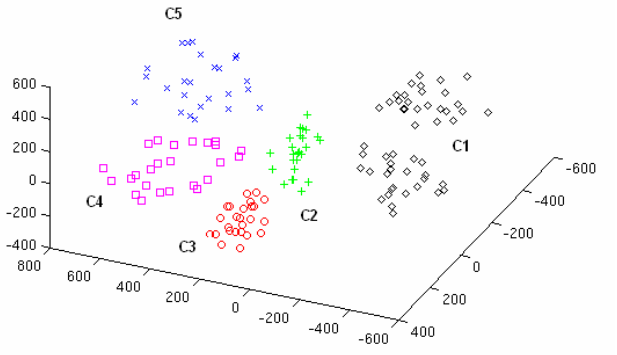

Figure 4. Automatic clustering of 180 speech sequences. 5 clusters are detected.

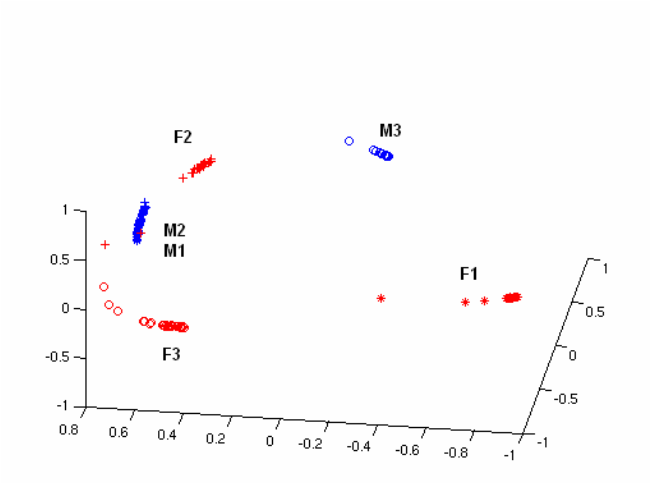

Figure 5. Applying Kernel PCA to the mapping of figure 3 facilitates automatic clustering.

4.3. Language projection

Language identification is a very difficult task using only acoustic space modeling. Currently it is supervised, but we apply our framework to check some results. We use 60 files from the OGI database: 20 sequences are in English, 20 in Italian and 20 in Mandarin.

In relation to precedent tests, we modify the features modeled by GMM, incorporating shifted delta cepstra (SDC) parameters [9] (with configuration 7-1-3-3) as well as a speech detection algorithm for silence elimination.

After low-dimensional mapping we apply Kernel PCA. The resulting points are displayed with the original label of the sequence in the database (figure 6), and we can identify a certain language grouping.

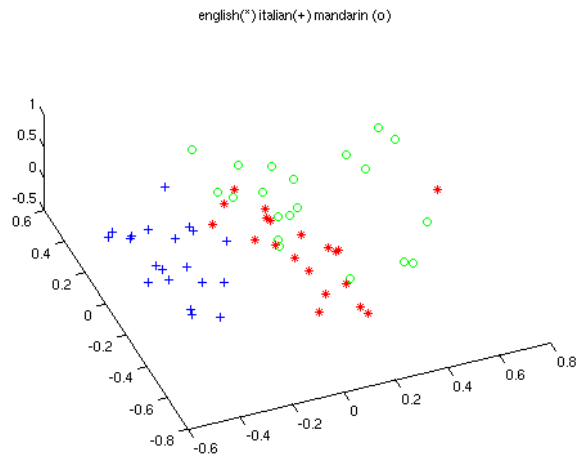

Figure 6. Language projection. Three languages are represented, English (*), Italian (+) and Mandarin (o).

5. CONCLUSIONS

We propose a system which maps variable length audio signals into fixed low-dimensional vectors in Euclidian space. Three applications are explored: speech-music classification, speaker identification and language projection. As low dimensional representations of the sequences offer the possibility of being visually represented, kernel and clustering procedures can be easily customized. The proposed framework can be used to test several algorithms for multimedia processing, like different dissimilarity measures between statistical models or other kernel and clustering strategies. We show encouraging results.

6. ACKNOWLEDGEMENTS

José Arias work is supported by a SFERE-CONACYT scholarship of the Mexican National Council of Science and Technology.

11. REFERENCES

[1] C. Bishop, *Pattern recognition and machine learning*, Springer, 2006.

[2] I. Borg and P. Groenen, *Modern multidimensional scaling: Theory and applications*, Springer, 1997.

[3] M. Carey, E. Parris and H. Lloyd-Thomas, "A comparison of features for speech, music discrimination", *ICASSP*, pp. 149-152, 1999.

[4] P. Knees, M. Schedl, T. Pohle and G. Widmer, "Exploring music collections in virtual landscapes", *IEEE Multimedia*, pp. 46-54, vol. 14 num. 3, 2007.

[5] P. Moreno, P. Ho and N. Vasconcelos, "A Kullback-Leiber divergence based kernel for SVM classification in Multimedia applications", NIPS 16, 2004.

[6] Y. Muthusamy, R. Cole and B. Oshika, "The OGI multilanguage telephone speech corpus", *ICSLP*, pp. 895-898, vol. 2, 1992.

[7] S. Kajarekar and A. Stolcke, "NAP and WCCP: Comparison of approaches using MLLR-SVM speaker verification system", *ICASSP*, 2007.

[8] B. Scholkopf, A. Smola and K. Muller, "Nonlinear component analysis as a kernel eigenvalue problem", *Neural Computation*, pp. 1299-1319, num. 10, 1998.

[9] P. A. Torres-Carrasquillo, "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features", *ICSLP*, 2002.

[10] C. White, I. Shafran and J. L. Gauvain, "Discriminative classifiers for language recognition", *ICASSP*, 2006.

[11] T. Zhang and C. Kuo, "Hierarchical system for content-based audio classification and retrieval", *Conference on Multimedia storage and Archiving systems*, 1998.