SPEAKER IDENTIFICATION USING DIFFUSION MAPS

Yan Michalevsky, Ronen Talmon, Israel Cohen

Department of Electrical Engineering, Technion, Israel Institute of Technology
Technion Campus, Haifa 32000, Israel
crcat@tx.technion.ac.il, ronenta2@tx.technion.ac.il, icohen@ee.technion.ac.il

ABSTRACT

In this paper we propose a data-driven approach for speaker identification without assuming any particular speaker model. The goal in speaker identification task is to determine which one of a group of known speakers best matches a given voice sample. Here we focus on text-independent speaker identification, i.e. no assumption is made regarding the spoken text. Our approach is based on a recently developed manifold learning technique, named *diffusion maps*. Diffusion maps enable embedding of the recording into a new space, which is likely to capture the speech intrinsic structure. The algorithm is tested and compared to common identification algorithms. Experimental results show that the proposed algorithm obtains improved results when few labeled samples are available.

1. INTRODUCTION

Automatic speaker recognition draws a growing interest as modern communication systems develop. The goal in speaker identification task is to determine which one of a group of known speakers best matches a given voice sample. It has a variety of applications in forensics, customer service over telephone, biometric access control and more. In this paper, we focus on text-independent speaker identification

Common identification algorithms have a typical structure, which consists of feature extraction, model training given labeled data, and finally classification of unlabeled samples. Usually, the first stage of these methods is of a key importance, where various features are extracted to distinguish between different speakers. For instance, simple and commonly-used features for speaker identification are the Mel-Frequency Cepstral Coefficients (MFCC) and their first and second derivatives, pitch, and a variety of vocal tract parameters [1]. In this work, we put less emphasis on the feature extraction phase, and adaptively build an intrinsic representation of the data.

In [2], Reynolds and Rose proposed a speaker identification technique based on a Gaussian Mixture Model. This model is considered one of the most successful likelihood functions for stochastic modeling of speakers for text-independent speaker identification [3]. A combination of multiple Gaussian distributions is used to approximate the shape of the spectral features of a speaker, allowing speaker modeling using few statistical parameters. By increasing the number of Gaussians, it is possible to increase the accuracy of the model.

To date, many algorithms are based on the GMM model. These methods usually obtain satisfactory results, however they suffer from several drawbacks. First, the assumed GMM model requires calibration of the model parameters. Second, setting these parameters usually requires a large set of labeled samples for training. Third, the performance and complexity of these methods are heavily dependent upon the number of speakers, since the GMM-based classification involves computation of a score per every speaker model.

In this paper, we propose a data-driven approach for speaker identification without assuming any particular speaker model. Our approach is based on a recently developed manifold learning technique, termed *diffusion maps* [4]. We assume that the input speech recordings are sampled from a low-dimensional manifold lying in a high-dimensional space. The samples are embedded into a new space which parametrizes the manifold. We show that the manifold parametrization, emerged based on the data, provides improved features. The embedding enables visualization of the speech data in two or three dimensions, which provides a notion of the similarity or disparity of the vocal features of different speakers. In addition, the classification is independent upon the number of speakers and requires a small training set. Our experimental results demonstrate high identification rate of speakers compared to alternative methods.

This paper is organized as follows: In Section 2 the identification algorithm is described. In Section 3 experimental results are presented and analyzed, followed by conclusions in Section 4.

2. SPEAKER IDENTIFICATION

The proposed identification is carried out in three steps:

1. *Feature extraction* - given training data of labeled speech samples, a variety of auditory features are computed for each sample in the training set. These features constitute a feature vector that captures temporal or spectral properties of a speech sample.
2. *Manifold learning* - the feature vectors of the training data are embedded into a new Euclidean space. The embedded space conveys parametrization in a manifold using *diffusion maps* [4]. The embedded samples are divided into clusters of speakers according to their labels.
3. *Classification* - based on the parametrization of the training set, we find the embedding of new unlabeled samples using *geometric harmonics* [5]. Using k-NN algorithm we associate the new embedded samples with one of the clusters corresponding to a certain speaker.

2.1 Feature Extraction

We use the Mel Frequency Cepstral Coefficients (MFCC) [6] to characterize different speakers. Let $\{x_i\}_{i=1}^{M}$ denote a training data set of M speech samples of several speakers. We assume a sampling frequency of 8 kHz. For each training sample we calculate the MFCC and their first temporal derivative

(delta-MFCC) using 40 subbands for short frames of 64 ms each. We then take the first 13 coefficients and calculate the mean, variance, minimum and maximum for the MFCC and mean and variance for the delta-MFCC across all frames of a recording. Consequently, we obtain a 78-dimensional feature for each recording. Collecting all the vectors constitutes a high-dimensional data set $\{y_i\}_{i=1}^{M}$, where each point y_i is the feature vector corresponding to the speech sample x_i.

2.2 Diffusion maps framework

In order to capture the subtle properties of different speakers we compute a large number of features, which implies a high-dimensional feature vector. To cluster such high-dimensional data efficiently - dimensionality reduction needs to be applied. The feature vectors can be seen as points in a high-dimensional space. We study the intrinsic geometry of the points using diffusion maps.

The common approach when using Diffusion Maps is to define an affinity metric $k(x_i,x_j)$ between pairs of speech samples based on the corresponding features using the following Gaussian kernel

$$k(x_i,x_j) = \exp\left\{-\frac{\|y_i-y_j\|^2}{2\sigma^2}\right\} \quad (1)$$

where σ^2 is the variance of the Gaussian kernel which determines the scale of the affinity metric.

Since the characteristic scale of different features is non-homogeneous, instead of using a uniform scale control σ^2 for all features we use a different value for each element. The kernel has now the form

$$k(x_i,x_j) = \exp\left\{-\frac{1}{2}(y_i-y_j)^T \underline{\varepsilon}^{-1}(y_i-y_j)\right\} \quad (2)$$

where $\underline{\varepsilon}_{N\times N} = diag(\varepsilon_1,...,\varepsilon_N)$ and N is the number of features. Using a different scale for each feature takes into account its range of values which might differ from that of other features. The scale ε_i is chosen as the variance of the i-th feature across all samples, multiplied by some constant C (referred to as a scale control parameter). The constant value is chosen empirically according to an optimization performed during the training stage.

We view the vectors $\{x_i\}_{i=1}^{M}$ as nodes of an undirected symmetric graph. Two nodes x_i and x_j are connected by an edge with weight $k(x_i,x_j)$, that corresponds to the affinity between x_i and x_j. We continue with the construction of a random-walk on the graph nodes by normalizing the kernel k [7]

$$p(x_i,x_j) = k(x_i,x_j)/d(x_i) \quad (3)$$

where $d(x_i) = \sum_j k(x_i,x_j)$. Consequently, $p(x_i,x_j)$ represents the probability of transition in a single step from node x_i to node x_j. Similarly, let $p_t(x_i,x_j)$ be the probability of transition in t steps from node x_i to node x_j. Let K denote the matrix corresponding to the kernel function k, and let $P = D^{-1}K$ be the matrix corresponding to the transition function p, where D is a diagonal matrix with $D_{ii} = d(x_i)$. Accordingly, P^t is the matrix corresponding to the transition function p_t.

Another possible normalization of the kernel k is the symmetric normalization that approximates the Fokker-Planck operator [4]

$$p_{sym}(x_i,x_j) = \frac{k(x_i,x_j)}{\sqrt{d(x_i)}\sqrt{d(x_j)}}. \quad (4)$$

It is called symmetric for being represented by a symmetric affinity matrix

$$P_{sym} = D^{-\frac{1}{2}} K D^{-\frac{1}{2}}. \quad (5)$$

This normalization is less sensitive to the distribution of samples and therefore has advantages in representation of non-uniformly sampled data.

In our implementation we slightly modify the kernel. We enlarge the connections between the points in the same cluster by setting the affinity of the point to itself to 0, formulated as

$$k(x_i,x_j) = \begin{cases} \exp\left[-\frac{1}{2}(x_i-x_j)^T \underline{\varepsilon}^{-1}(x_i-x_j)\right] & i \neq j \\ 0 & i = j \end{cases}. \quad (6)$$

Thus, the normalization (3) is not affected by the trivial affinity between the point to itself, and connections to other points are emphasized. It results in a better embedding and increases the correct identification percentage.

Spectral decomposition [7] is employed to describe P, enabling to study the geometric structure of the data in a compact and efficient way. It can be shown that P has a complete sequence of left and right eigenvectors $\{\varphi_j,\psi_j\}$ and eigenvalues, written in a descending order $1 = \lambda_0 \geq \lambda_1 \geq \lambda_2 \geq \ldots$, satisfying $P\psi_j = \lambda_j\varphi_j$.

The construction of the random walk and the spectral decomposition lead to a definition of a new affinity metric $D_t(x_i,x_j)$ between pairs of samples, given by

$$\begin{aligned} D_t^2(x_i,x_j) &= \left\|p_t(x_i,\cdot) - p_t(x_j,\cdot)\right\|_{\varphi_0}^2 \\ &= \sum_{k=1}^{M} \left(p_t(x_i,x_k) - p_t(x_j,x_k)\right)^2 / \varphi_0(x_k) \end{aligned} \quad (7)$$

for any integer t. This metric is termed *diffusion distance* as it relates to the evolution of the transition probability distribution p_t. It enables to describe the relationship between pairs of samples in terms of their graph connectivity.

We use the right eigenvectors $\{\psi_j\}$ of the transition matrix P to obtain a new data-driven description of the M samples $\{x_i\}$ via a family of mappings that are termed *diffusion maps* [4]. Let $\Psi_t(x_i)$ for some $t > 0$ be the diffusion mappings of the set $\{x_i\}$ into a Euclidean space $\mathbb{R}^\ell$, defined as

$$\Psi_t(x_i) = \left[\lambda_1^t\psi_1(x_i),\ldots,\lambda_\ell^t\psi_\ell(x_i)\right]^T \quad (8)$$

where ℓ is the new space dimensionality of our choice, ranging between 1 and $M-1$. We note that a fast decay of $\{\lambda_j\}$ may enable dimensionality reduction, as coordinates in (8) become negligible for large ℓ.

It can be shown [4] that the diffusion distance (7) equals the Euclidean distance in the diffusion maps space for $\ell = M-1$, i.e.

$$D_t^2(x_i,x_j) = \left\|\Psi_t(x_i) - \Psi_t(x_j)\right\|^2. \quad (9)$$

This result provides a justification for using the Euclidean distance in the new space for clustering and classification tasks. In particular, since the spectrum is fast decaying for a large enough t, the diffusion distance can be well approximated by only the first few ℓ eigenvectors, yielding efficient comparisons. Similar spectral decomposition and embedding are applied to P_{sym} as well.

2.3 Classification

The training stage consists of feature extraction for all samples in the training set and computation of their corresponding embedding. To classify a new sample $\bar{x}$ as a recording of one of the known speakers, we repeat the feature extraction process and obtain a new feature vector $\bar{y}$. In order to find the coordinates of the embedding of the new sample $\Psi_t(\bar{x})$ we use geometric harmonics[5]. It is based on the following Nyström extension of the eigenvectors

$$\overline{\psi}_j(\bar{x}) = \frac{1}{\lambda_j} \sum_{i=1}^{M} p(\bar{x}, x_i)\, \psi_j(x_i). \qquad (10)$$

We note that the Nyström extension of the eigenvectors agree on the training set, i.e. $\overline{\psi}_j(x_i) = \psi_j(x_i), \forall i = 1,\ldots,M$. We refer to $\overline{\psi}_j(\bar{x})$ as *extended eigenvectors*. It is the approximation of the embedding coordinates for a new sample using the eigenvectors of the transition matrix calculated for the training set.

Based on the extended eigenvectors $\{\overline{\psi}_j\}$, the new sample embedding $\Psi_t(\bar{x}_i)$ is computed. Then, we apply k-NN classification to associate $\Psi_t(\bar{x})$ with a certain speaker. In our implementation we use $k = 10$.

3. EXPERIMENTAL RESULTS

The experimental study was conducted with YOHO speaker verification database [8]. The corpus consists of 106 male and 32 female speakers pronouncing "combination lock" phrases. There are 24 enrollment and 10 verification sessions per speaker. For all measurements we have repeatedly chosen random subsets of speakers from the whole set, and averaged the obtained values to present results that are independent of specific speakers choice.

Our empirical tests show that using our modification of the kernel scale yields better results. Moreover, based on these tests, we calibrated the scale control parameter.

In Fig. 1 and 2 we show the diffusion map embedding of the feature vectors of five different speakers, using both the symmetric normalization and random-walk normalization. We observe a clear clustering of speakers and distinct separation between points associated with different speakers. In particular, the presented 3-dimensional embedding enables visualization of the speakers voice similarity. Therefore such an embedding might be useful for finding the "sheep" and "wolves"[1] among many speakers.

We compare the correct identification rate of the proposed classifier with two other classifiers, one based on a GMM implementation described in [9], configured to 5 Gaussians, and the other classifier based on classification using k-NN in the original feature space, with $k = 10$. All the

[1] "Sheep" - the default speakers for which recognition systems perform well. "Wolves" - speakers whose voice features are exceptionally similar to features of other speakers.

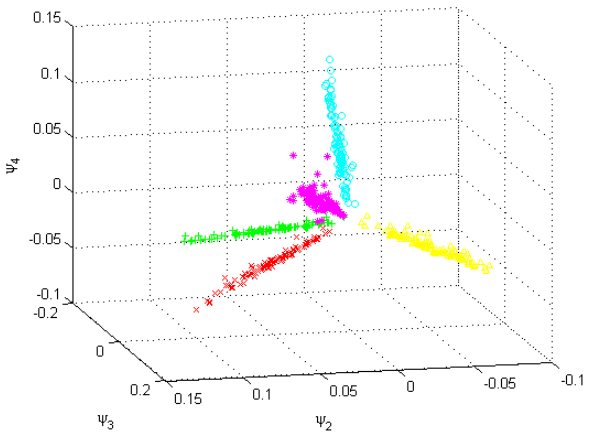

Figure 1: Embedding into $\mathbb{R}^3$ of features of 5 speakers using symmetric normalization. First three nontrivial eigenvectors were used for the embedding, at time t=3.

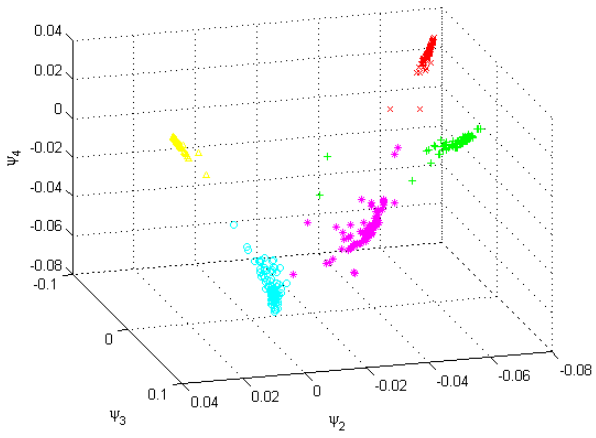

Figure 2: Embedding using Markovian random-walk matrix. First three nontrivial eigenvectors were used for the embedding, at time t=3.

	Diffusion Maps		GMM	k-NN
Num. of speakers	**Random Walk**	**Symmetric**		
2	100%	100%	100%	99.2%
3	99.5%	99.3%	99.5%	98.2%
5	99.4%	99.1%	99.5%	97.7%
10	97.8%	96.9%	98.1%	95.1%
20	94.4%	93.2%	97.5%	92.0%

Table 1: Evaluation of Diffusion Maps, k-NN and GMM classifiers for different numbers of speakers. The table presents the average correct classification percentage of each algorithm.

	Diffusion Maps		GMM	k-NN
Num. of speakers	**Random Walk**	**Symmetric**		
2	98.2%	99.1%	97.4%	97.9%
3	97.8%	98.7%	95.2%	97.8%
5	96.5%	96.0%	92.1%	93.4%
10	92.5%	91.9%	87.4%	89.0%
20	86.4%	84.6%	83.3%	84.5%

Table 2: Evaluation of Diffusion Maps, k-NN and GMM classifiers for the case of few labeled samples. The table presents the average correct classification percentage of each algorithm.

classifiers rely on the same features. We compare the three classifiers for different numbers of speakers, averaging the results of multiple tests. The results are presented in Table 1. In this experiment the size of the training set containing labeled samples is 9 times larger than the size of the test set. We can see that diffusion maps perform better than k-NN for every number of speakers, and is comparable to GMM for a small number of speakers. When the number of speakers increases there is a degradation in the performance of diffusion maps compared to GMM.

In a second experiment, we compare the correct classification rate for a scenario in which we have few labeled samples. The samples are partitioned such that the training set containing labeled samples is 4 times *smaller* than the size of the test set. The classification results are presented in Table 2. We see that Diffusion Maps yield better results than GMM and k-NN for each number of speakers in the test.

In both experiments the random-walk normalization outperforms the symmetric normalization for every number of speakers. However, for visualization purposes, we might prefer to use the symmetric normalization and count the "rays" formed by the embedding of feature vectors associated with those speakers (as emerged in Fig. 1).

4. CONCLUSIONS

We have applied a manifold learning method to the speaker identification task. Our method utilizes relatively simple features and relies on the embedding to capture the perceptual variability of the data. We have demonstrated that classification using diffusion maps can outperform nearest neighbors in the original feature space. It can also outperform the Gaussian mixture speaker model, which is used in common speaker-identification algorithms to-date. Our modification to the commonly used Gaussian kernel can be further tested in other applications. We have also demonstrated diffusion maps as a helpful visualization method for speech data.

REFERENCES

[1] J.P. Campbell et al., "Speaker recognition: A tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.

[2] D.A. Reynolds and R.C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[3] F. Bimbot, J.F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, and D.A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing*, vol. 2004, pp. 430–451, 2004.

[4] R.R. Coifman and S. Lafon, "Diffusion maps," *Applied and Computational Harmonic Analysis*, vol. 21, no. 1, pp. 5–30, 2006.

[5] S.S. Lafon, *Diffusion maps and geometric harmonics*, Ph.D. thesis, Yale University, 2004.

[6] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 2003.

[7] F.R.K. Chung, "Spectral graph theory (cbms regional conference series in mathematics, no. 92)," *American Mathematical Society*, vol. 3, pp. 8, 1997.

[8] J.P. Campbell Jr, "Testing with the YOHO CD-ROM voice verification corpus," in *Proc. IEEE Internat. Conf. Acoust., Speech Signal Process., ICASSP-95*, 2002, vol. 1, pp. 341–344.

[9] Anil Alexander and Andrzej Drygajlo, "Speaker identification: A demonstration using matlab," http://scgwww.epfl.ch/courses/Biometrics-Lectures-2005-2006-pdf/03-Biometrics-Exercise-3-2005/03-Biometrics-Exercise-3-2005.pdf, April 2005.