Classification of Unvoiced Fricative Phonemes using Geometric Methods

Michal Genussov
Department of Electrical Engineering,
Technion- Israel Institute of Technology,
Haifa, Israel
e-mail: mgenus@tx.technion.ac.il

Yizhar Lavner
Department of Computer Science,
Tel-Hai Academic College,
Upper Galilee, Israel
e-mail: yizhar.lavner@gmail.com

Israel Cohen
Department of Electrical Engineering,
Technion- Israel Institute of Technology,
Haifa, Israel
e-mail: icohen@ee.technion.ac.il

Abstract— **Phoneme classification is the process of finding the phonetic identity of a short section of a spoken signal. Performances of existing classification techniques are often insufficient, since they rely on Euclidean distances between spectral and temporal features, whereas the relevant features lie on a non-linear manifold. In this work, we propose to integrate into the phoneme classification a non-linear manifold learning technique, namely "Diffusion maps". Diffusion maps builds a graph from the feature vectors and maps the connections in the graph to Euclidean distances, so using Euclidean distances for classification after the non-linear mapping is optimal. We show that Diffusion maps allows dimensionality reduction and improves the classification results.**

Keywords- feature extraction, phoneme classification, unvoiced fricatives, diffusion maps, geometric harmonics.

I. INTRODUCTION

Classification of phonemes is the process of finding the phonetic identity of a short section of a spoken signal [1]. It is a key stage in many speech processing algorithms and applications, such as spoken term detection, continuous speech recognition, and speech coding, but it can also be useful on its own, for example in selective processing of phonemes for the hearing impaired, or in the professional music industry. In existing methods for discrimination between phonemes [1], [2], [3], the first step in the process of classification is extracting relevant features from the signal, and constructing a vector of features (feature-vector) for each phoneme segment. In the second step, a classification algorithm is applied on the feature-vector, such as k-nearest neighbors (K-NN) [4] or support vector machines (SVM) [2].
There are two fundamental problems in these methods:
1. In order to capture optimally the nature of the signal and differ efficiently between phonemes, the feature-vector usually needs to be high-dimensional. As the number of signals increases, the computational complexity increases as well, leading to the need of a dimensionality reduction technique.
2. Usually the data - the feature-vectors, lie on a non-linear manifold, therefore classification techniques which rely on Euclidean distances might yield poor classification results. Moreover, linear and global methods for dimensionality reduction, such as Principal Component Analysis (PCA) [5], will not reveal the true geometry of the manifold.

To overcome these problems, non-linear manifold learning techniques, such as ISOMAP, LLE, Laplacian Eigenmaps or Hessian Eigenmaps, can be applied as an intermediate step of dimensionality reduction, before the classification operation itself. In this work we use such a technique called "Diffusion Maps" [6], [7], which provides a parameterization of the data set on a lower-dimensional manifold, while emphasizing the differences between feature-vectors of different phonemes.
Another task which is dealt with is the out-of-sample extension problem. A method called "Geometric Harmonics" [8] allows reducing the computational complexity by extending the parameterization of diffusion maps from a limited training set to the rest of the training set. Furthermore, it embeds each new phoneme we wish to classify, from a testing set, into the diffusion maps parameterization of the limited training set.

The rest of the paper is organized as follows: Methods and algorithms are presented in section II. The experimental procedure and preprocessing are described in Section II.A, the feature extraction step is detailed in Section II.B and the Diffusion framework is introduced in Section II.C. Experimental results are presented in Section III. Finally, the advantages of using the diffusion maps for phoneme spotting are discussed in Section IV.

II. METHODS AND ALGORITHMS

A. *Experimental Procedure and Preprocessing*

The dataset for this study includes more than 1100 isolated phonemes, excerpted from the TIMIT speech database, of both male and female speakers. The phonemes chosen for the analysis are the unvoiced fricatives /s/, /sh/, /f/ and /th/. These phonemes are specifically important since they tend to be indistinguishable for the hearing impaired [9].
In the preprocessing stage, each phoneme segment is divided into consecutive non-overlapping short frames (8 ms) which are denoted as "analysis frames", and multiplied by a hamming window. The reason for the short length of the frames is twofold: first, the classification of the whole phoneme can be improved by using a majority vote decision, and in addition, it can be used for a real-time application of phoneme spotting. Since the important information of the unvoiced fricatives is contained in the high frequency range, this choice is suitable. A feature extraction stage is then applied, using both time domain and frequency domain parameters to

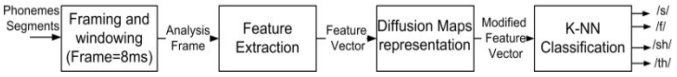

Figure 1: The experimental procedure.

represent each frame. As described in the next section, a total of 15 features are selected according to their discrimination ability.

The classification algorithm is applied in two steps:

1. Dimensionality reduction – embedding the data into a lower dimensional manifold using diffusion maps (section II.D).
2. Classification of the data, with and without the dimensionality reduction, using the K-NN algorithm.

A schematic block diagram of the experimental procedure is shown in Figure 1.

B. *Feature Extraction*

The features used to characterize the phonemes are mostly based on the spectral shape, but also on the time domain parameters. The features are computed for each analysis frame, and include:

1. *Spectal Peak Locations* [3]: These include frequency locations of the peaks of the spectral envelope.
2. *Spectral Rolloff:* The spectrum rolloff point is defined as the boundary frequency f_r, below which p percent of the magnitude distribution is concentrated
$$\sum_{k=0}^{f_r} M_t(k) = p \cdot \sum_{k=0}^{K-1} M_t(k) \qquad (1)$$
where $M_t(k)$ is the magnitude of the Fourier transform at frame t and frequency bin k. In this study, p values of 25%, 50% and 75% are used.
3. *Spectral Centroid*: The spectrum centroid is defined as the center of gravity of the magnitude spectrum of the STFT
$$S_t = \frac{\sum_{k=0}^{K-1} M_t(k) \cdot k}{\sum_{k=0}^{K-1} M_t(k)}$$
4. *Band Energy Ratio*: Band Energy ratio is defined as the ratio of the spectral energies of two bands
$$\hat{E}_t = 10\log_{10}\left(\frac{E_{B_1}}{E_{B_2}}\right)$$
where E_{B_1} and E_{B_2} are the spectral energies of two frequency bands, (here $B_1 = 4-8kHz$ and $B_2 = 2-4kHz$).
5. *Zero Crossing Rate (ZCR):* The zero-crossing rate of a frame is defined as the number of times the audio waveform changes its sign in the duration of the frame:
$$ZCR = \frac{1}{2}\sum_{n=1}^{N-1}\left|\mathrm{sgn}\left(x[n]\right) - \mathrm{sgn}\left(x[n-1]\right)\right|$$
where $x(n)$ is the time domain signal for frame t.
6. *Time Domain Zero Crossings Standard Deviation, Skewness and Kurtosis:* These moments of the ZCR are computed using the statistics of the time intervals between consecutive zero crossings.
7. *Mel Frequency Cepstral Coefficients*: Mel-Frequency Cepstral Coefficients (MFCC) are also based on the STFT. After computing the logarithm of the magnitude spectrum, and grouping the DFT bins according to a Mel-frequency scale, a discrete cosine transform is performed on the result. Only the first three coefficients were found to be discriminative and are used for the feature vector.
8. *Lacunarity β parameter*, as described in [10].

C. *Diffusion Framework*

1) Embedding Into a Lower Dimensional Manifold

Let $X = \{\mathbf{x}_i\}_{i=1}^{M}$ be a high-dimensional data set of M samples. Let $k : X \times X \to \mathbb{R}$ be a kernel representing a notion of similarity between two data samples. Based on the relation defined by the kernel, we form a weighted graph or a Euclidean manifold, where the data samples are the vertices and the kernel sets the weights of the edges connecting the data points. The kernel, for $(\mathbf{x}_i, \mathbf{x}_j) \in X$, is

- symmetric : $k(\mathbf{x}_i, \mathbf{x}_j) = k(\mathbf{x}_j, \mathbf{x}_i)$
- positive semi-definite: $k(\mathbf{x}_i, \mathbf{x}_j) \geq 0$.

The specific kernel function is application oriented, chosen to yield meaningful connections, and it constitutes our prior definition of the local geometry of X. Following classical construction in spectral graph theory [11], a Markov random walk on the data set is defined
$$p(\mathbf{x}_i, \mathbf{x}_j) = \frac{k(\mathbf{x}_i, \mathbf{x}_j)}{d(\mathbf{x}_i)}, \qquad (2)$$
where $d(\mathbf{x}_i) = \sum_{j=1}^{M} k(\mathbf{x}_i, \mathbf{x}_j)$. The function p can be considered as the transition probability function of a Markov chain on $\{\mathbf{x}_i\}$, since $\sum_{j=1}^{M} p(\mathbf{x}_i, \mathbf{x}_j) = 1$. Specifically, $p(\mathbf{x}_i, \mathbf{x}_j)$ represents the probability of transition in a single random walk step from node $\mathbf{x}_i$ to node $\mathbf{x}_j$. Let $p_t(\mathbf{x}_i, \mathbf{x}_j)$ be the probability of transition from $\mathbf{x}_i$ to $\mathbf{x}_j$ in t steps. Let K denote the matrix corresponding to the kernel function $k(\cdot,\cdot)$,

where its $(i,j)th$ element is $k(\mathbf{x}_i,\mathbf{x}_j)$, and let $P = D^{-1}K$ be the matrix corresponding to the function $p(\cdot,\cdot)$ on the data set $\{\mathbf{x}_i\}$, where D is a diagonal matrix with $D_{ii} = d(\mathbf{x}_i)$. Let $\mathbf{X}$ be a matrix consisting of the data samples

$$\mathbf{X} = \left[\mathbf{x}_1, \mathbf{x}_2 ..., \mathbf{x}_M\right]^T. \quad (3)$$

Advancing the random walk on the data set a single step forward can be written as $P\mathbf{X}$. Running the random walk t steps forward corresponds to applying the kernel function $p_t(\mathbf{x}_i,\mathbf{x}_j)$, and is equivalent to $P^t\mathbf{X}$. With a smart choice of parameters this process results in revealing the relevant geometric structure of $\{\mathbf{x}_i\}$, and taking larger powers of P enlarges the scale. The matrix P has a discrete sequence of non-negative eigenvalues $\{\lambda_l\}_{l\geq 0}$ and right eigenvectors $\{\psi_l\}_{l\geq 0}$ such that $1 = \lambda_0 > \lambda_1 \geq \lambda_2 \geq ...$ and $P\psi_l = \lambda_l\psi_l$. The distances on the set $\{\mathbf{x}_i\}$ which represent the connectivity in the graph in scale t are called *diffusion distances* and are notated as $\{D_t\}_{t\in\mathbb{N}}$. We define *the family of diffusion maps* [7] $\Psi_t(\mathbf{x}_i)$ as:

$$\Psi_t(\mathbf{x}_i) \triangleq \begin{bmatrix} \lambda_1^t\psi_1(\mathbf{x}_i) \\ \lambda_2^t\psi_2(\mathbf{x}_i) \\ \vdots \\ \lambda_{s(\delta,t)}^t\psi_{s(\delta,t)}(\mathbf{x}_i) \end{bmatrix}, \quad (4)$$

where $s(\delta,t) = \max\left\{l \in \mathbb{N} \text{ such that } |\lambda_l|^t > \delta|\lambda_1|^t\right\}$, and $\delta > 0$ is the relative accuracy (explained later). Each component of $\Psi_t(\mathbf{x}_i)$ is termed a *diffusion coordinate*. Results from spectral theory show that the diffusion map $\Psi_t : X \to \mathbb{R}^{s(\delta,t)}$ embeds the data set into a Euclidean space of $s(\delta,t)$ dimensions, and in this space Euclidean distance is equal to the diffusion distance up to the relative accuracy δ, or equivalently,

$$\left\|\Psi_t(\mathbf{x}_i) - \Psi_t(\mathbf{x}_j)\right\|_2 \cong D_t(\mathbf{x}_i,\mathbf{x}_j). \quad (5)$$

Therefore, if the dimensionality of the data can be reduced to $s(\delta,t)$, then the equation above is an *exact* equality, and the Euclidean norm captures the exact distance between nodes $\mathbf{x}_i$ and $\mathbf{x}_j$ in the manifold of dimension $s(\delta,t)$. As t increases, the spectrum decay is faster, and $s(\delta,t)$ is smaller.

In this work, the set $\{\mathbf{x}_i\}$ represents the set of feature-vectors of the unvoiced fricative phonemes. In order to convey the *local* geometry of the data, we used a Gaussian kernel

$$k(\mathbf{x}_i,\mathbf{x}_j) = \exp\left(-\left\|(\mathbf{x}_i - \mathbf{x}_j)./\sigma\right\|\right), \quad (6)$$

where σ is a vector that consists of elements proportional to the standard deviations of each of the features, and the division is element-wise, leading to a *multi scale* embedding. We do not change the scaling of the manifold learning (i.e. $t = 1$), and we set $s(\delta,t) = 10$, which means that we use only the top 10 diffusion coordinates (which correspond to the largest eigenvalues that do not equal 1). Therefore, the family of diffusion maps is reduced to:

$$\Psi(\mathbf{x}_i) \triangleq \begin{bmatrix} \lambda_1\psi_1(\mathbf{x}_i) \\ \lambda_2\psi_2(\mathbf{x}_i) \\ \vdots \\ \lambda_{10}\psi_{10}(\mathbf{x}_i) \end{bmatrix}. \quad (7)$$

2) Out-of-Sample Extension

The parameterization described in the previous subsection is conducted over a limited set from the training set, to maintain a limited computational complexity. In order to extend the family of diffusion maps to the rest of the data (which includes both the rest of the training set and the testing set), we use a method called "Geometric Harmonics" [8]. If we denote the limited data set, which was used to build the matrix P as X, and the rest of the training set as $\bar{X}$ $(X \subset \bar{X})$, then the extended eigenvectors which belong to the feature vector $\bar{\mathbf{x}}_i \in \bar{X}$ can be calculated as a weighted sum of the eigenvectors of the limited data set:

$$\bar{\psi}_l(\bar{\mathbf{x}}_i) = \frac{1}{\lambda_l}\sum_{\{\mathbf{x}_j\}\in X} p(\bar{\mathbf{x}}_i,\mathbf{x}_j)\psi_l(\mathbf{x}_j) \quad (8)$$

and the new family of diffusion maps for each vector $\bar{\mathbf{x}}_i$ is:

$$\bar{\Psi}(\bar{\mathbf{x}}_i) \triangleq \begin{bmatrix} \lambda_1\bar{\psi}_1(\bar{\mathbf{x}}_i) \\ \lambda_2\bar{\psi}_2(\bar{\mathbf{x}}_i) \\ \vdots \\ \lambda_{10}\bar{\psi}_{10}(\bar{\mathbf{x}}_i) \end{bmatrix}. \quad (9)$$

In order to classify new data from the testing set, which will be denoted as $\tilde{X}$, the Geometric Harmonics method is applied again for every feature-vector $\tilde{\mathbf{x}}_i \in \tilde{X}$, and the extended eigenvectors $\tilde{\psi}_l(\tilde{\mathbf{x}}_i)$ are calculated as in (8). The new family of diffusion maps for each vector $\tilde{\mathbf{x}}_i$, is, similarly, given by:

$$\tilde{\Psi}(\tilde{\mathbf{x}}_i) \triangleq \begin{bmatrix} \lambda_1\tilde{\psi}_1(\tilde{\mathbf{x}}_i) \\ \lambda_2\tilde{\psi}_2(\tilde{\mathbf{x}}_i) \\ \vdots \\ \lambda_{10}\tilde{\psi}_{10}(\tilde{\mathbf{x}}_i) \end{bmatrix}. \quad (10)$$

A new phoneme is classified by applying K-nearest neighbors (K-NN) algorithm (with K=5) with Euclidean distance. The classification is applied on the family of diffusion maps of the testing set $\tilde{\Psi}\left(\tilde{\mathbf{x}}_i\right)$ according to that of the training set $\bar{\Psi}\left(\bar{\mathbf{x}}_i\right)$.

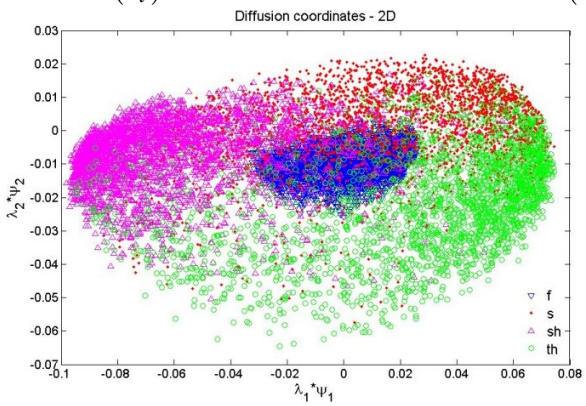

Figure 2: Diffusion map (2D) of the unvoiced fricatives feature vectors. The phonemes are marked by: /f/ - blue, /s/ - red, /sh/ - magenta, /th/ - green.

	"[f]"	"[s]"	"[sh]"	"[th]"
[f]	**0.91**	0.01	0.02	0.06
[s]	0.03	**0.85**	0.11	0.02
[sh]	0.06	0.04	**0.90**	0.00
[th]	0.41	0.16	0.03	**0.40**

	"[f]"	"[s]"	"[sh]"	"[th]"
[f]	**0.90**	0.01	0.05	0.04
[s]	0.01	**0.88**	0.08	0.02
[sh]	0.07	0.05	**0.88**	0.00
[th]	0.51	0.21	0.05	**0.23**

Table 1: Confusion matrix for male data using K-NN– with diffusion maps (up) and without diffusion maps (down), for classification using majority vote

The classification results are compared to classification using K-NN without embedding the features using diffusion maps. We chose K-NN as the classifier because of its simplicity.

For visualization, the embedding of the feature vectors to a 2D mapping using diffusion maps is shown in Figure 2. The clusters that represent different phonemes can be seen.

III. EXPERIMANTAL RESULTS

For the evaluation of the performance of the algorithm the database described in Section II.A is used. The feature vectors are then produced, one for each frame, for male and for female separately (44,148 and 25,531, respectively). A K-NN algorithm is used for the classification, with $K=5$, for the original data (feature vectors of 15-d) and the data after applying the diffusion maps stage (feature vectors of 10-d). The results are obtained using 10-fold cross-validation. The results presented here are the average for 1000 testing sets.

In classification of each analysis window separately, the average correct identification rate of the algorithm for the data with the intermediate step of diffusion maps (feature vectors of 10-d) is $68.23 \pm 1.22\%$, while the classification using the original 15 dimensional feature vectors is $65.34 \pm 1.17\%$. Similar results are obtained for female data.

Applying a majority vote for the feature vectors of the same phoneme segment, an accuracy of $78.44 \pm 5.88\%$ is obtained using the diffusion maps coordinates, compared to an accuracy of only $73.79 \pm 6.08\%$ with the original feature vector.

The results of the evaluation stage of the K-NN with and without the mapping with diffusion maps are summarized in the confusion matrices in Table I.

IV. CONCLUSSIONS

The method of "Diffusion maps" for manifold learning leads to improved classification of unvoiced fricative phonemes when using a simple classification algorithm as K-NN, and to reduction of the dimension of the problem. From this work it seems that the features that distinguish between different phonemes lie on a non-linear, lower-dimensional manifold. The classification should be conducted in this manifold, and not in the original space of features. Moreover, dimensionality reduction leads to lower computational complexity and saves storage space. Future work may include comparison to other methods of nonlinear manifold learning.

REFERENCES

[1] O. Dekel, J. Keshet, and Y. Singer. An online algorithm for hierarchical phoneme classification, Machine Learning for Multimodal Interaction, pp. 146-158, 2005 , Lecture Notes in Computer Science, Springer, Vol. 3361/2005, pp. 146-158, 2005.

[2] J. Salomon. Support Vector Machines for Phoneme Classification, M.Sc Thesis, University of Edinburgh, 2001.

[3] A. M. Abdelatty Ali and J. Van der Spiegel, Acoustic-phonetic features for the automatic classification of fricatives, Journal of Acoustical Society of America, Volume 109, issue 5, 2001

[4] S. Cost and S. Salzberg. A weighted nearest neighbor algorithm for learning with symbolic features, Machine learning, vol. 10, pp. 57-78, 1993.

[5] J. Shlens. A tutorial on principal component analysis, Technical report, Center for Neural Science, New York University, New York City, Apr. 2009.

[6] S. Lafon, Y. Keller, and R. R. Coifman. Data fusion and multicue data matching by diffusion maps. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28(11), pp.1784–1797, Nov. 2006.

[7] R. Coifman and S. Lafon. Diffusion maps. Applied and Computational Harmonic Analysis, vol. 21, pp.5–30, June 2006.

[8] R. Coifman and S. Lafon. Geometric harmonics: A novel tool for multiscale out-of-sample extension of empirical functions. Applied and Computational Harmonic Analysis, vol. 21, pp.31–52, June 2006.

[9] D. Bauer , A. Plinge and M. Finke, Selective Phoneme Spotting for Realization of an /s, z, C, t/ Transposer, Lecture Notes in Computer Science, Springer, Volume 2398-2002, pp. 271-306, 2002

[10] L. J. Hadjileontiadis, A texture-based classification of crackles and squawks using lacunarity, IEEE Transactions on Biomedical Engineering, Vol. 56, No. 3, pp. 718-732, 2009.

[11] F. R. K. Chung, Spectral Graph Theory, American Mathematical Society, 1997.