Voiced-Unvoiced-Silence Classification via Hierarchical Dual Geometry Analysis

Maya Harel, David Dov, Israel Cohen, Ronen Talmon and Ron Meir
Andrew and Erna Viterbi Faculty of Electrical Engineering
Technion – Israel Institute of Technology
Technion City, Haifa 32000, Israel

Abstract—**The need for a reliable discrimination among voiced, unvoiced and silence frames arises in a wide variety of speech processing applications. In this paper, we propose an unsupervised algorithm for voiced-unvoiced-silence classification based on a time-frequency representation of the measured signal, which is viewed as a data matrix. The proposed algorithm relies on a hierarchical dual geometry analysis of the data matrix, which exploits the strong coupling between time frames and frequency bins. By gradually learning the coupled geometry in two steps, the algorithm allows for the separation between speech and silent frames, and then between voiced and unvoiced frames. Experimental results demonstrate the improved performance compared to a competing algorithm.**

Index Terms—**Speech classification, dual geometry analysis, partition trees, hierarchical clustering**

I. Introduction

The classification of speech into voiced, unvoiced and silence (V/UV/S) frames is an important and challenging problem which provides a preliminary acoustic segmentation for pitch estimation, automatic speech recognition, speaker recognition, speech analysis, speech enhancement, and speech signal compression. While in this study we consider the classification of the three classes – V/UV/S, some previous studies separately consider the segmentation only between silent and non-silent frames [1-4] or between voiced and unvoiced frames [5-7]. In addition, the V/UV classification was performed using simple features derived from the speech signal, such as the energy of the signal, the zero-crossing rate and the degree of voice periodicity [5-6]. However, the accuracy of such approaches is limited since the range of the values of such features may overlap between categories.

Several recent studies addressed the problem of V/UV/S classification from a supervised learning perspective, using methods such as support vector machines and Gaussian mixture models (GMM) [7-9]. Such methods typically suffer from a lack of labeled training databases and their performance dramatically degrades when training and test statistics mismatch due to variances in speakers, accents, languages, noise types and levels, etc. V/UV/S classification, which is not based on training databases, is therefore desired to overcome these issues.

We address the problem of V/UV/S classification by analysing the geometry of the time-frequency representation of the measured signal, which is viewed as a data matrix whose rows correspond to time frames and columns to frequency bins. The goal of such an analysis is to organize the data matrix in a manner that respects the hidden connectivity structures between both dimensions. The need for matrix co-organization typically arises when correlations exist among both rows and columns of the data matrix. A typical example is the analysis of documents or psychological questionnaires, where there is no particular reason to prefer treating one dimension as independent and the other as dependent.

To address problems of this sort, Gavish et al. [10-11] and Ankenman [12], proposed a methodology for matrix organization based on dual geometry analysis. It begins with learning the hierarchical structure of the data in one dimension via local clustering of the data, using a partition tree. The tree induces a multiscale metric on the dual dimension, which enables to define an affinity measure between data points. Finally, one can derive an intrinsic embedding via manifold learning. The organization of the entire matrix is carried out in an iterative procedure, where each dimension is organized in turn, based on the other one. Namely, given the organization of the columns of the data matrix, the rows of the matrix are organized into groups of related rows, and vice versa, such that this approach introduces a strong coupling between the dimensions.

In this paper, we present an unsupervised learning algorithm for V/UV/S classification. The algorithm is based on the representation of a speech signal in the time-frequency domain, which is viewed as a data matrix. Then, the data matrix is analysed and its dual geometry is learned in two steps. Specifically, the learned coupling between the time frames and frequency bins naturally leads to a separation between silent and speech frames in the first step, as well as voiced and unvoiced frames in the second step. The three classes are further distinguished based on their frequency and energy content. We evaluate the proposed method using TIMIT database in the presence of different noise levels, and attain encouraging results.

The remainder of the paper is organized as follows. In Section II, we formulate the problem. In Section III, we describe the dual geometry analysis of the time-frequency representation of the measured signal using the method presented in [12], and present the proposed algorithm for V/UV/S classification. In Section IV we show experimental results, which demonstrate the performance of the proposed algorithm on the TIMIT database.

978-1-5090-2152-9/16/$31.00 ©2016 IEEE

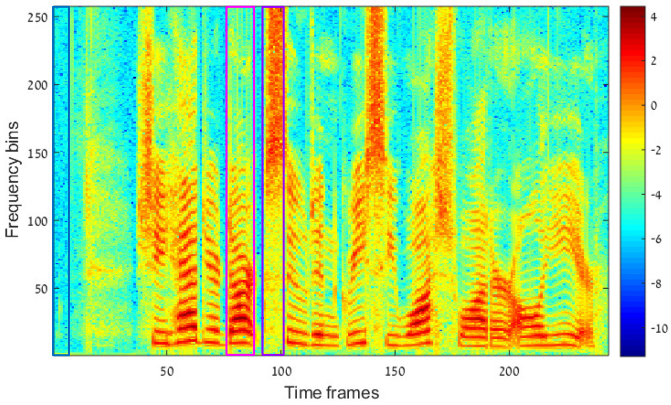

Fig 1. Spectrogram of a clean speech signal. Voiced, unvoiced and silence patterns are marked in pink, purple and blue, respectively.

II. Problem Formulation

Consider a speech signal processed in frames and represented by the log magnitude of the short-time Fourier transform (STFT), given by the data matrix $X \in \mathbb{R}^{n_r \times n_t}$. The parameter n_r is the number of frequency bins and n_t is the number of time frames. Let $\mathcal{X}_t = \{x_{\cdot,i}\}_{i=1}^{n_t}$ be the set of all columns in X. Each time frame, i.e., an element in $\mathcal{X}_t$, is assumed to belong to one of three classes: voiced, unvoiced or silence. Voiced speech consists of approximately constant frequency tones, generated when vowels are spoken. About two thirds of speech phonemes are voiced and they generally exhibit periodicity in time and frequency domains, concentrated at low to middle frequencies due to vocal-tract resonances. In contrast, unvoiced speech is nonperiodic, random-like sound, excited by turbulence sources only and exhibits stronger power at higher frequencies than at lower frequencies [13-14]. Silent frames, in which speech is absent, may consist of an unstructured white Gaussian noise. Fig. 1 depicts the differences between voiced and unvoiced phonemes in the STFT domain. These properties of the measured signal implicitly imply that a different coupling is expected between the time frames and frequency bins in the different classes.

The objective in the paper is to classify each time frame in $\mathcal{X}_t$ by learning the coupled geometry of the measured signal in a hierarchical manner.

III. Hierarchical Dual Geometry Analysis of Time-Frequency Speech Representation

A. Partition trees

A key component in the analysis of hierarchical dual geometry of a data matrix is a partition tree. The tree encodes the similarity between the data points in a fine-to-coarse manner, which conveys an intrinsic hierarchical clustering of the data points. In our setting, data points are either column vectors (time frames) or row vectors (frequency bins). Without loss of generality, we will define the partition trees in this section with respect to the columns of the matrix.

Following [15], let T_t be the partition tree of $\mathcal{X}_t$. T_t is composed of a finite sequence of partitions P_l, $0 \leq l \leq L$, with the following properties:

- The partition P_l, which represents the l-th tree-level, consists of $n(l)$ disjoint nonempty subsets of indices in $\{1, \dots, n_t\}$, termed folders and denoted by $I_{l,k}$, $k \in \{1, \dots, n(l)\}$, where $\bigcup_{k=1}^{n(l)} I_{l,k} = \{1, \dots, n_t\}$.
- The coarsest partition $(l = L)$ is composed of a single folder termed root.
- For $l < L$, the partitions are nested such that if $I \in P_l$, then $I \subseteq J$ for some $J \in P_{l+1}$.
- The finest partition $(l = 0)$ is composed of $n(l) = n_t$ singleton folders termed leaves.

The partition tree is the set of all folders at all levels $T_t = \{I_{l,k} \mid 0 \leq l \leq L, 1 \leq k \leq n(l)\}$. Note that we define the folders on the indices of the points and not on the points themselves.

The construction of the tree is performed in a bottom-up fashion, based on pairwise distances between data points. We follow [12] and use the diffusion distance [16], which is more robust to noise compared, e.g., to the Euclidean distance. Furthermore, it yields relatively few levels and the level at which folders are joined is meaningful across the entire dataset. Thus, the tree structure is logically multiscale and follows the structure of the data. The construction is controlled by a constant ε which constrains the size of folders and affects the number of levels in the tree such that increasing ε results in "taller" trees.

The first partition of the data points is generated as follows:

1) Input: A set of n_t columns $\mathcal{X}_t$, a column affinity matrix $A_t \in \mathbb{R}^{n_t \times n_t}$, and a constant ε.
2) Initialization: Set $I_{0,k} = \{k\}$ and $l = 1$.
3) Given an affinity on the data, construct a low-dimensional embedding of the data. Calculate the pairwise diffusion distances $d^{(l)}(i,j)$ $\forall 1 \leq i,j \leq n_t$.
4) Set a threshold $\frac{p}{\varepsilon}$ where $p = median\left(d^{(l)}(i,j)\right)$.
5) For each singleton point i, find its minimal distance $d^{min}(i) = min_j\{d^{(l)}(i,j)\}$.
If $d^{min}(i) > \frac{p}{\varepsilon}$, then i remains as a singleton folder.
Otherwise, i and j form a new folder if j also does not belong to any folder. If j is already part of a folder, then i is added to that folder if $d^{min}(i) < \frac{p}{\varepsilon} 2^{-|I|+1}$.
6) The partition P_l is set to be all the formed folders.

Coarser and coarser partitions are obtained by repeating steps (3)-(6) until all points are merged together at the root. Instead of iterating over points, we iterate over the folders $I_{l-1,k} \in P_{l-1}$. For more detailed description see [12].

B. Earth mover's distance metric

The construction of diffusion distance requires a "good" affinity measure between data points. In our context, an appropriate affinity measure must constitute the notion of dual

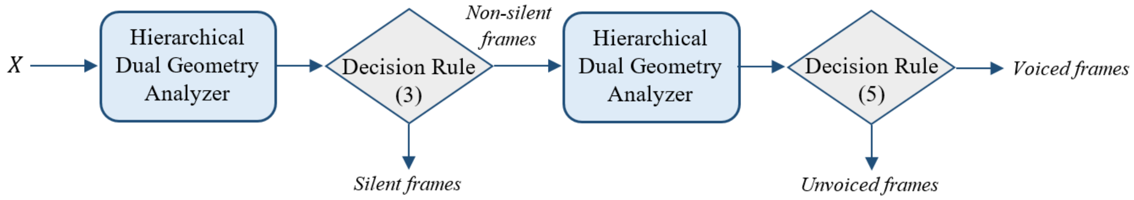

Fig. 2. Block diagram of the two-step proposed classifier.

geometry. Hence, an affinity measure between the time frames must incorporate some organization of the frequency bins and exploit the multiscale neighborhoods defined by the partition tree into the distance.

We use an affinity measure based on the Earth Mover's Distance (EMD), which is usually formulated as a distance between two probability distributions or discrete histograms [17]. The EMD measures the minimal cost of transforming one distribution into the other. Distributions which are a slightly perturbed version of each other will have small EMD, thus the EMD has the attractive property of being insensitive to small distortions.

We adopt an EMD-like distance, that is, an equivalent version of the EMD that is appropriate for the setting of this paper [12,18]. Without loss of generality, the EMD-like distance between columns $x_{.,i}$ and $x_{.,j}$, given a partition tree T on the rows of X, is defined as:

$$d_{EMD}(i,j) = \sum_{I \in T} \left| m\left(x_{.,i} - x_{.,j}, I\right) \right| w(I) \qquad (1)$$

where $m(x, I)$ is the mean value of column x on folder I, and $w(I) > 0$ is a weight function, depending on the folder I, whose selection is described in [12].

C. V/UV/S classification

The core of the proposed algorithm for V/UV/S clustering lies in learning the coupled geometry of the measured signal in two steps.

In the first step, we learn the hierarchial dual organization of the time-frequency matrix X using a partition tree, as described in the previous section, so that time frames are grouped together based on their similar frequency bins. Our experiments have shown that apart from the root, the coarsest level of the resulting tree for the time frames (P_{L-1}) separates the silent frames from the non-silent frames. It consists of two folders, $I_{L-1,1}$ and $I_{L-1,2}$, one for each class. To label the two folders, we propose to compare their short-time energy values. Let $E(I)$ be the energy value of folder I, given by:

$$E(I) = \frac{1}{|I|} \sum_{t \in I} \sum_{f} X^2[f,t] \qquad (2)$$

The energy of the silence folder is radically low as compared to non-silence folder, thus our decision rule is as follows:

$$E\left(I_{L-1,1}\right) \underset{I_{L-1,1}\ \text{Silence}}{\overset{I_{L-1,1}\ \text{Speech}}{>/<}} E\left(I_{L-1,2}\right) \qquad (3)$$

where $E\left(I_{L-1,1}\right)$ and $E\left(I_{L-1,2}\right)$ are the energy values of folders $I_{L-1,1}$ and $I_{L-1,2}$, respectively. Accordingly, the i-th time frame is a silent frame if it belongs to the folder with the lower energy value.

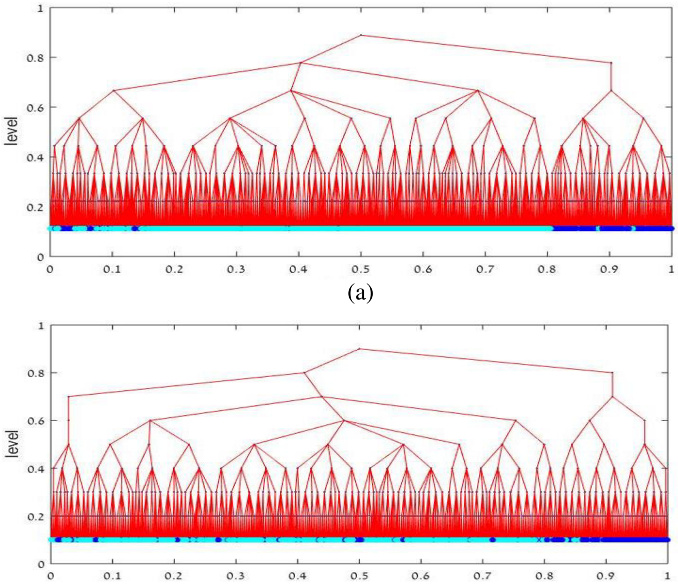

Fig. 3. Partition trees of time frames. (a) Tree resulting from the first step, and (b) from the second step. The leaves are colored in blue and magenta according to the true labelling: (a) silence and non-silence, and (b) unvoiced and voiced, respectively.

In the second step, the same approach is applied to a matrix, similar to X, which is constructed only using the non-silent frames. Similarly, the partition tree on the time dimension groups the voiced frames and unvoiced frames in two different folders, $I_{L-1,1}$ and $I_{L-1,2}$. We suggest to detect each folder based on the ratio of energies at high frequencies to that at lower frequencies, given by:

$$R(I) = \frac{\sum_{t \in I} \sum_{f > 150} X^2[f,t]}{\sum_{t \in I} \sum_{f < 150} X^2[f,t]} \qquad (4)$$

where the folder with lower values of the ratio (4) is assumed consisting of voiced frames:

$$R\left(I_{L-1,1}\right) \underset{I_{L-1,1}\ \text{Voiced}}{\overset{I_{L-1,1}\ \text{Unvoiced}}{>/<}} R\left(I_{L-1,2}\right) \qquad (5)$$

We note that we found it difficult to classify speech signals into three categories (V/UV/S) in a single step. In contrast, the proposed algorithm allows a good classification of the time frames using the two-step procedure, as we will show in Section IV. Moreover, the two-step algorithm may be seen as a construction of a new tree in a top-down manner. Namely, the root is separated into speech and silent folders, and then, the speech folder is further separated into voiced and unvoiced phonemes. The third step of such a construction, which we leave to a future study, would include separating the voiced and unvoiced folders into sub-folders of phonemes.

We summarize the proposed approach in Algorithm 1, and a flow chart of the proposed classifier is illustrated in Fig. 2.

Algorithm 1

```
Step 1:
Input: A logarithmic magnitude of the STFT – matrix $X$.
Initialization:
    1. Calculate an initial affinity measure on either the rows or columns of $X$. We assume columns here and use an affinity measure based on the Euclidean distance.
    2. Calculate an initial partition tree on the columns $T_t$.
Iterations: For $n \geq 1$,
    3. Calculate the tree-based distance (1) between each pair of rows.
    4. Calculate row affinity measure
           $A_r^{(n)}(i,j) = e^{-d_{EMD,r}(i,j)/\sigma_r}$.
    5. Calculate partition tree on the rows $T_r^{(n)}$.
    6. Calculate the tree-based distance (1) between each pair of columns.
    7. Calculate column affinity measure
           $A_t^{(n)}(i,j) = e^{-d_{EMD,t}(i,j)/\sigma_t}$.
    8. Calculate partition tree on the columns $T_t^{(n)}$.
End for
Decision rule: Label folders $I_{L-1,1}$ and $I_{L-1,2}$ in $T_t^{(n)}$ as silence and non-silence using (3).

Step 2:
Apply steps 1-8 on the part of $X$ corresponding to non-silent frames.
Decision rule: Label folders $I_{L-1,1}$ and $I_{L-1,2}$ in $T_t^{(n)}$ as voiced and unvoiced using (5).
```

IV. Experimental Results

The performance of the proposed algorithm for V/UV/S classification is evaluated on the TIMIT database [19]. A subset of the TIMIT database, including 80 short sentences from 10 different (male and female) speakers (8 from each speaker), is used for the evaluation. In our experiment we refer to a rather complicated scenario, in which multiple speakers are concatenated and various levels of white Gaussian noise are added. The speech signals are sampled at 16kHz and processed in time frames of 512 samples with 50% overlap.

Fig. 3 demonstrates the hierarchical partitioning of the time frames via the partition trees. Fig. 3(a) results from the first step of the proposed algorithm, and Fig. 3(b) from the second. Both trees clearly organize the time frames into two meaningful clusters, colored in blue and magenta. In Fig. 3(a), the time frames are grouped into two dominant folders of silent and non-silent frames, and in Fig. 3(b) the non-silent frames are separated into voiced and unvoiced.

We further evaluate the performance of the proposed algorithm for voice-unvoiced classification in the form of success rate, that is, the percentage of correct classifications relatively to the number of time frames. We compare our performance to the algorithm presented in [9], which we term PEFAC. PEFAC provides an estimate of voiced speech per time frame, as a part of a pitch tracking scheme, via a supervised learning procedure based on GMM. The results are shown in Table 1 for clean and noisy signals, contaminated with white Gaussian noise at 30dB, 20dB and 10dB SNR. The proposed algorithm outperforms PEFAC for low noise levels due to some fundamental differences between the algorithms. First, the proposed algorithm is independent of any specific speaker features such as pitch frequency. Thus, it can provide solution to the V/UV/S classification problem in the presence of multiple speakers, unlike PEFAC which is designed solely for a single speaker framework. Second, the proposed algorithm is purely data-driven in the sense that it does not require training data and is free from the problem of statistical mismatch between test and training datasets. The limitation of the algorithm is being rather sensitive to noise, since noise distorts the time-frequency representation, as illustrated in Fig. 4.

Another source of errors stems from the fact that the TIMIT dataset is manually marked. Manual labeling affects the accuracy of the boundaries and a few milliseconds of error is inevitable (Fig. 5). These built-in mistakes decrease the performance of our algorithm and should be taken into account when comparing to other methods.

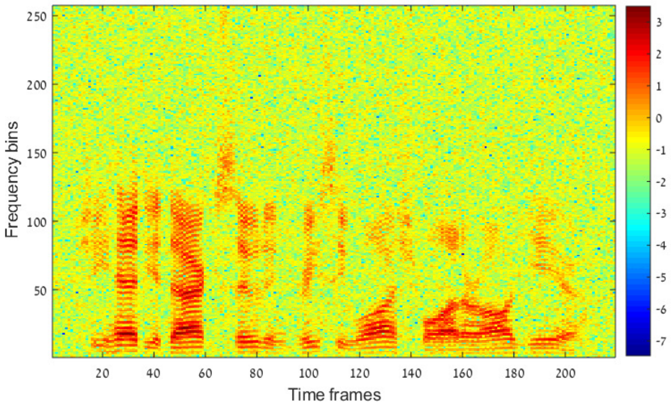

Fig. 4. Spectrogram of a noisy speech signal with 10dB SNR.

	Clean	30dB SNR	20dB SNR	10dB SNR
Hierarchical Dual Geometry Classifier	93.9%	91.9%	88.2%	63.8%
PEFAC	79.7%	79.7%	79.4%	78.8%

Table 1. Comparison of overall success rates between Hierarchical Dual Geometry Classifier and PEFAC. Noisy signal obtained using white Gaussian noise.

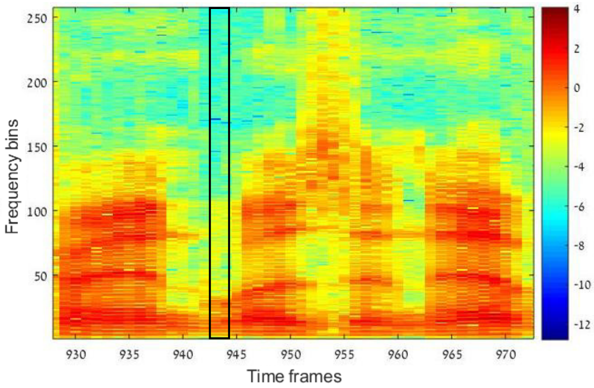

Fig. 5. Illustration of labelling errors in the TIMIT dataset. This is a part of the spectrogram, focusing on time frames 943 and 944 (marked in black). The ground truth annotation of both frames is silence, although it is clear that their energy content is greater than zero.
In our framework, the frames are correctly detected voiced, which unjustly yields a classification error.

V. Conclusion

We have introduced an algorithm for V/UV/S classification. The algorithm learns the coupled geometry between time frames and frequency bins of a speech signal via a two-step unsupervised procedure. Such a classification presents some significant advantages, e.g., being independent of the availability of training databases.

The algorithm has been evaluated using TIMIT database and attained good results in the presence of multiple speakers and noises of different levels.

Further improvement can be made by adding a prior noise suppression stage, in order to reduce noise in the cases where the SNR is below a certain level [20].

References

[1] D. Dov and I. Cohen, “Voice activity detection in presence of transients using the scattering transform,” in *Proc. IEEE 28th Convention of Electrical & Electronics Engineers in Israel (IEEEI)*, pp. 1–5, 2014.

[2] D. Dov, R. Talmon and I. Cohen, “Audio-visual voice activity detection using diffusion maps,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 732–745, 2015.

[3] D. Dov, R. Talmon, and I. Cohen, “Kernel method for voice activity detection in the presence of transients,” To appear in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

[4] D. Dov, R. Talmon, and I. Cohen, "Kernel-based Sensor Fusion with Application to Audio-Visual Voice Activity Detection." arXiv preprint arXiv:1604.02946, 2016.

[5] R. G. Bachu, S. Kopparthi, B. Adapa and B. D. Barkana, “Voiced/ unvoiced decision for speech signals based on zero-crossing rate and energy,” in K. Elleithy (Ed.), *Advanced Techniques in Computing Sciences and Software Engineering*, pp. 279-282, 2010.

[6] A. M. Noll, “Cepstrum pitch determination,” *J. Acoust. Soc. Amer.*, vol 41, pp.293-309, Feb. 1967.

[7] J. K. Shah, A. N. Iyer, B. Y. Smolenski, and R. E. Yantorno, “Robust voiced/unvoiced classification using novel features and Gaussian mixture model,” in *Proc. ICASSP 2004*, Montreal, Quebec, Canada, 2004.

[8] F. Qi, C. Bao and Y. Liu, “A novel two-step SVM classifier for voiced/unvoiced/silence classification of speech,” in *International Symposium on Chinese Spoken Language Processing*, pp. 77–80, 2004.

[9] S. Gonzalez and M. Brookes, “PEFAC – a pitch estimation algorithm robust to high levels of noise,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 22, no. 2, pp. 518–530, Feb. 2014.

[10] M. Gavish, B. Nadler, and R. R. Coifman, “Multiscale wavelets on trees, graphs and high dimensional data: Theory and applications to semi supervised learning,” in *Proc. ICML 2010*, 2010, pp. 367–374.

[11] R. R. Coifman and M. Gavish, “Harmonic analysis of digital data bases,” *Wavelets and Multiscale analysis*, Birkhäuser Boston, 2011, pp. 161-197.

[12] J. I. Ankenman, “Geometry and analysis of dual networks on questionnaires,” Ph.D. dissertation, Yale University, 2014.

[13] J. K. Lee, C. D. Yoo, “Wavelet speech enhancement based on voiced/ unvoiced decision,” Korea Advanced Institute of Science and Technology The 32nd International Congress and Exposition on Noise Control Engineering, Jeju International Convention Center, Seogwipo, Korea, August 25-28, 2003.

[14] H. Deng, and D. O'Shaughnessy, “Voiced-unvoiced-silence speech sound classification based on unsupervised learning,” 2007 IEEE International Conference on Multimedia and Expo, pp.176-179, July 2007.

[15] G. Mishne, R. Talmon, et al., “Hierarchical coupled geometry analysis for neuronal structure and activity pattern discovery,” submitted arXiv:1511.02086v1

[16] R. R. Coifman and S. Lafon, “Diffusion maps,” *Appl. Comput. Harmon. Anal.*, vol. 21, no. 1, pp. 5–30, July 2006.

[17] Y. Rubner, C. Tomasi and L. J. Guibas, “The earth mover's distance as a metric for image retrieval,” *Int. J. Comput. Vision 40*, pp. 99–121, 2000.

[18] R. R. Coifman and W. E. Leeb, “Earth mover’s distance and equivalent metrics for spaces with hierarchical partition trees,” Yale University, Tech. Rep., 2013.

[19] J. S. Garofolo, “Getting started with the DARPA TIMIT CD-ROM: An acoustic-phonetic continuous speech database,” Nat. Inst. of Standards and Technology (NIST), Gaithersburg, MD, 1993.

[20] I. Cohen and B. Berdugo, “Speech enhancement for non-stationary noise environments,” *Signal Process.*, vol. 81, no. 11, pp. 2403–2418, 2001.