# CLUSTERING AND SUPPRESSION OF TRANSIENT NOISE IN SPEECH SIGNALS USING DIFFUSION MAPS

*Ronen Talmon*[1]*, Israel Cohen*[1] *and Sharon Gannot*[2]

[1] Department of Electrical Engineering
Technion - Israel Institute of Technology
Technion City, Haifa 32000, Israel
{ronenta2@tx, icohen@ee}.technion.ac.il

[2] School of Engineering
Bar-Ilan University
Ramat-Gan, 52900, Israel
gannot@eng.biu.ac.il

## ABSTRACT

Recently we have presented a novel approach for transient noise reduction that relies on non-local (NL) filtering. In this paper, we modify and extend our approach to support clustering and suppression of a few transient noise types simultaneously, by introducing two novel concepts. We observe that voiced speech spectral components are slowly varying compared to transient noise. Thus, by applying an algorithm for noise power spectral density (PSD) estimation, configured to track faster variations than pseudo-stationary noise, the PSD of speech components may be estimated. In addition, we utilize *diffusion maps* to embed the measurements into a new domain. We obtain a new representation which enables clustering of different transient noise types. The new representation is incorporated into a NL filter as a better affinity metric for averaging over transient instances. Experimental results show that the proposed algorithm enables clustering and suppression of multiple transient interferences.

***Index Terms***— Speech enhancement, speech processing, acoustic noise, impulse noise, transient noise.

## 1. INTRODUCTION

Recently [1] [2] we have presented a novel approach for transient noise reduction that relies on non-local (NL) filtering [3]. Transient noise includes noise originating from engines, keyboard typing, construction operations, bells, knocking, rings, hammering, etc. The algorithm presented in [1] handles speech signals contaminated with repeating transient noise events, utilizing the fact that a distinct pattern appears multiple times. Specifically, the locations of the repeating pattern are identified, and the transient noise is extracted by averaging over all of these instances.

A fundamental part of the algorithm in [1] is the distinction between speech and transients. In order to enhance this difference, the measurements are whitened using linear prediction coding. In each short time frame, the linear prediction coefficients (LPC) of the speech are estimated and used to whiten the signal in the frame. As a result, speech components are whitened, whereas transients maintain their impulsive nature. Unfortunately, this preprocessing suffers from two limitations. First, estimating the LPC from the noisy measurement might be a difficult task yielding inaccurate estimation. Second, each transient is altered differently as different coefficients are estimated in each frame. Consequently, the whitening process may exclude the assumption that the same pattern appears multiple times.

In this paper, we propose a different, more robust, approach to distinguish between transients and speech. Common speech enhancement algorithms include a noise power spectral density (PSD) estimation component, which exploits the fact that the spectrum of environmental noise is usually slowly varying with time compared to that of speech. Hence, the noise PSD is estimated by smoothing the signal power and tracking only slow variations while neglecting fast energy bursts. We utilize the same concept for "transient enhancement". We observe that voiced speech spectral components are slowly varying compared to transient noises. Thus, by applying a common algorithm for noise PSD estimation, configured to track faster variations than pseudo-stationary noise, the speech PSD may be estimated from measurements of the speech signal contaminated by transients. We note that exploiting the rate of change of the signal was previously introduced in RASTA [9], where bandpass filtering of the short-term power spectrum is employed to suppress slowly and rapidly varying interferences.

In addition, the proposed algorithm enables clustering and suppression of a few transient noise types. We utilize *diffusion maps* [4] to embed the measurements into a new space and obtain a new representation. The Euclidean distance in the diffusion maps space, which is termed *diffusion distance*, is a robust perceptual metric for comparison. In particular, it enables clustering of different transient noise types. Moreover, when incorporated into the NL filter, it provides a better affinity metric for averaging over transient instances.

This paper is organized as follows. In Section 2, we formulate the problem. In Section 3, the algorithm for clustering and suppression of transient noise is presented. Finally, in Section 4, we show experimental results that demonstrate the advantages of the proposed method.

## 2. PROBLEM FORMULATION

Let $y(n)$ be a measured speech signal corrupted by $L$ additive transient noise types

$$y(n) = s(n) + \sum_{l=1}^{L} d^l(n) \tag{1}$$

where $s(n)$ is a clean speech signal and $d^l(n)$ is the $l$th contaminating transient noise type.

The transient noise is modeled as the output of a filter excited by an amplitude-modulated random binary sequence [5], given by

$$d^l(n) = h^l(n) * \left(b^l(n)v^l(n)\right) \tag{2}$$

where the impulse response $h^l(n)$ models the duration and shape of the events of the $l$th transient noise type, $b^l(n) \in \{0,1\}$ is a binary valued random sequence of time locations of transient noise events of type $l$, and $v^l(n)$ is a continuous valued random process of transient amplitudes. In this work, we use a fixed impulse response, which implies that the transient events of each type have the same spectral features up to random amplitudes. This assumption leads to the observation that the same pattern appears in the measurement several times, and is of a key importance in the proposed algorithm.

We apply the short-time Fourier transform (STFT) to convey the spectral difference between the transients and the speech. We use time-frames of length $N$, and denote the number of frames in the finite observable interval by $M$. Let $y_{p,k}$ be the STFT of $y(n)$ in time frame $p$ and frequency bin $k$. Using (1), it can be written as

$$y_{p,k} = s_{p,k} + \sum_{l=1}^{L} d^l_{p,k} \tag{3}$$

where $s_{p,k}$ and $d^l_{p,k}$ are the STFT of $s(n)$ and $d^l(n)$ respectively.

We assume that in a single short time frame no more than one single transient exists (of any type). Accordingly, we define $\mathcal{H}_0$ to be the set of time frame indices without a transient, and $\mathcal{H}_l$ to be the set of time frame indices consisting of the $l$th transient noise type.

## 3. PROPOSED ALGORITHM

### 3.1. Transient Noise Enhancement

We adopt the concept commonly used for PSD estimation of (pseudo) stationary noise. A noise estimate is obtained by averaging over past spectral power values, and temporal smoothing is carried out to reduce fluctuations of speech segments. We observe that as stationary noise is slowly varying compared to speech, speech is slowly varying compared to transient noise. Thus, a speech estimate might be obtained by *short-term* averaging over past spectral power values, and the temporal smoothing reduces the abrupt transients.

Here we propose to use the optimally modified log spectral amplitude (OM-LSA) method [6] with fast recursion to enhance the transient noise and suppress the speech. We configure the minima controlled recursive averaging (MCRA) algorithm [7], which is employed in the OM-LSA for estimating the noise PSD, to track fast variations. We use very short time frames of length 16ms in order to reduce the variations of the speech between sequential frames. In addition, the following temporal smoothing is carried out

$$\hat{\phi}_s(p,k) = \alpha\hat{\phi}_s(p-1,k) + (1-\alpha)|y_{p,k}|^2 \tag{4}$$

where $\hat{\phi}_s(p,k)$ is the PSD estimate of the speech, and $\alpha$ is a recursion parameter. We choose a relatively small recursion parameter $\alpha = 0.5$ to enable quick tracking of speech components. However, the recursion parameter should not be too small to discard abrupt changes of transients.

The described modification enables to capture most of the speech parts, but sudden changes characterizing speech phonemes onsets are overlooked. Phoneme onset identification is obtained using two sliding windows for the speech PSD estimation. The first window is causal, and used to detect the minimum power in previous frames, as described in the original algorithm [7]. The second window is anti-causal, and used to detect the minimum power in future frames. We note that the window should be shorter than a typical speech phoneme, but longer than a typical transient. The PSD estimate is taken as the maximum of the two minima detected in the two windows. Now, at the beginning of a speech phoneme, the minimum in the causal window is low, conveying the power level of the background noise before the phoneme. On the other hand, the minimum in the anti-causal window is high, representing the power level of the phoneme (assuming the window is shorter than the phoneme). Consequently, taking the maximum of the two minima yields the desired estimate of the power level of the phoneme. It is worthwhile noting that a transient instance is not captured in this process. Since both windows are longer than a transient, the minima in such windows must be the power level of the background signal (either speech or background noise) before or after the transient.

### 3.2. Transient Noise Clustering using Diffusion Maps

Let $\tilde{y}(n)$ and $\tilde{y}_{p,k}$ be the measured signal in the time and the STFT domains after the application of the OM-LSA algorithm for the transient noise enhancement, as described in Section 3.1. The STFT coefficients of $\tilde{y}(n)$ from all frequency bins of each time frame $p$ are collected into a vector $\tilde{\mathbf{y}}_p$

$$\tilde{\mathbf{y}}_p = [\tilde{y}_{p,0}, \ldots, \tilde{y}_{p,N-1}]^T,\ p = 1, \ldots, M. \tag{5}$$

We define an affinity metric $k(\tilde{\mathbf{y}}_p, \tilde{\mathbf{y}}_l)$ between pairs of such vectors using the following Gaussian kernel

$$k\left(\tilde{\mathbf{y}}_p, \tilde{\mathbf{y}}_l\right) = \exp\left\{-\|\boldsymbol{\phi}_{\tilde{y}}(p) - \boldsymbol{\phi}_{\tilde{y}}(l)\|^2/2\sigma^2\right\} \tag{6}$$

where $\sigma^2$ is the variance of the Gaussian kernel which determines the scale of the affinity metric, and $\boldsymbol{\phi}_{\tilde{y}}(p)$ is a vector of length $N$, given by

$$\boldsymbol{\phi}_{\tilde{y}}(p) = [\phi_{\tilde{y}}(p,0), \ldots, \phi_{\tilde{y}}(p,N-1)]^T \tag{7}$$

where $\phi_{\tilde{y}}(p,k)$ is the short-time PSD of $\tilde{y}(n)$ in time-frame $p$ and frequency bin $k$, defined by $\phi_{\tilde{y}}(p,k) = \mathbb{E}\left\{\frac{1}{N}\tilde{y}_{p,k}\tilde{y}^*_{p,k}\right\}$, where $\mathbb{E}\{\cdot\}$ is an expectation. For more details regarding this specific choice of a kernel see [1].

We view the vectors $\{\tilde{\mathbf{y}}_p\}_{p=1}^M$ as nodes of an undirected symmetric graph. Two nodes $\tilde{\mathbf{y}}_p$ and $\tilde{\mathbf{y}}_l$ are connected by an edge with weight $k(\tilde{\mathbf{y}}_p, \tilde{\mathbf{y}}_l)$, that corresponds to the affinity between $\tilde{\mathbf{y}}_p$ and $\tilde{\mathbf{y}}_l$. We continue with the construction of a random-walk on the graph nodes by normalizing the kernel $k$ as

$$p(\tilde{\mathbf{y}}_p, \tilde{\mathbf{y}}_l) = k(\tilde{\mathbf{y}}_p, \tilde{\mathbf{y}}_l)/d(\tilde{\mathbf{y}}_p) \tag{8}$$

where $d(\tilde{\mathbf{y}}_p) = \sum_{l=1}^M k(\tilde{\mathbf{y}}_p, \tilde{\mathbf{y}}_l)$. Consequently, $p(\tilde{\mathbf{y}}_p, \tilde{\mathbf{y}}_l)$ represents the probability of transition in a single step from node $\tilde{\mathbf{y}}_p$ to node $\tilde{\mathbf{y}}_l$. Similarly, let $p_t(\tilde{\mathbf{y}}_p, \tilde{\mathbf{y}}_l)$ be the probability of transition in $t$ steps from node $\tilde{\mathbf{y}}_p$ to node $\tilde{\mathbf{y}}_l$. Let $K$ denote the matrix corresponding to the kernel function $k$, and let $P = D^{-1}K$ be the matrix corresponding to the function $p$, where $D$ is a diagonal matrix with $D_{pp} = d(\tilde{\mathbf{y}}_p)$. Accordingly, $P^t$ is the matrix corresponding to the function $p_t$.

Results from spectral theory [8] can be employed to describe $P$, enabling to study the geometric structure of $\{\tilde{\mathbf{y}}_p\}$ 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 positive eigenvalues, written in a descending order $1 = \lambda_0 > \lambda_1 \geq \lambda_2 \geq \ldots$, satisfying $P\psi_j = \lambda_j\varphi_j$.

The construction of the random walk leads to a definition of a new affinity metric between any two vectors [4]

$$\begin{aligned} D_t^2(\tilde{\mathbf{y}}_p, \tilde{\mathbf{y}}_l) &= \|p_t(\tilde{\mathbf{y}}_p, \cdot) - p_t(\tilde{\mathbf{y}}_l, \cdot)\|^2_{\varphi_0} \\ &= \sum_{q=1}^{M} \left(p_t(\tilde{\mathbf{y}}_p, \tilde{\mathbf{y}}_q) - p_t(\tilde{\mathbf{y}}_l, \tilde{\mathbf{y}}_q)\right)^2/\varphi_0(q) \end{aligned} \tag{9}$$

for any integer $t$. This metric is termed *diffusion distance* as it relates to the evolution of the transition probability distribution $p_t(\tilde{\mathbf{y}}_p, \tilde{\mathbf{y}}_l)$. It enables to describe the relationship between pairs of vectors in terms of their graph connectivity. Consequently, the main advantage of the diffusion distance is that local structures and rules of transitions are integrated together into a global metric. In recent years, this metric was shown to be very useful in various applications from different fields [10].

We use the right eigenvectors of the transition matrix $P$ to obtain a new data-driven description of the $M$ vectors $\{\tilde{\mathbf{y}}_p\}$ via a family of mappings that are termed *diffusion maps* [4]. Let $\Psi_t(\tilde{\mathbf{y}}_p)$ be the diffusion mappings of the $M$ vectors $\{\tilde{\mathbf{y}}_p\}$ into a Euclidean space $\mathbb{R}^\ell$, defined as

$$\Psi_t(\tilde{\mathbf{y}}_p) = \left[\lambda_1^t \psi_1(p), \cdots, \lambda_\ell^t \psi_\ell(p)\right]^T \tag{10}$$

where $\ell$ is the new space dimensionality ranging between 1 and $M-1$. We note that a fast decay of $\{\lambda_j\}$ may enable dimensionality reduction, as coordinates in (10) become negligible for large $\ell$.

It can be shown [4] that the diffusion distance (9) is equal to the Euclidean distance in the diffusion maps space when using all $\ell = M-1$ eigenvectors. This result provides a justification for using the Euclidean distance in the new space for spectral clustering purposes. 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 $\ell$ eigenvectors, yielding efficient comparisons.

In Section 4, we show that embedding the vectors into the diffusion maps space naturally organizes the data into separate clusters of speech and transient noises.

### 3.3. Transient Noise PSD Estimation using Diffusion Filters

Similarly to (6), we now define a *new* Gaussian kernel $\bar{k}$ based on diffusion distance

$$\bar{k}\left(\tilde{\mathbf{y}}_p, \tilde{\mathbf{y}}_l\right) = \exp\left\{-\|\Psi_t(\tilde{\mathbf{y}}_p) - \Psi_t(\tilde{\mathbf{y}}_l)\|^2 / 2\bar{\sigma}^2\right\} \tag{11}$$

and, similarly to (8), construct a corresponding random-walk to obtain a new transition probability function $\bar{p}(\tilde{\mathbf{y}}_p, \tilde{\mathbf{y}}_l)$. We denote by $\bar{K}$ and $\bar{P}$ the matrices corresponding to the new kernel and transition probability functions. We emphasize that unlike the kernel (6) used in [1] which relies on the Euclidean distance, we use here a kernel that relies on diffusion distance. The use of a diffusion distance conveys the capability to distinguish between different types of transients, and hence, the proposed algorithm enables handling few transient types simultaneously.

Let $\tilde{Y}$ be a matrix consisting of the set of vectors $\{\tilde{\mathbf{y}}_p\}$, $\tilde{Y} = [\tilde{\mathbf{y}}_1, \tilde{\mathbf{y}}_2, ...\tilde{\mathbf{y}}_M]^T$. Advancing the random-walk on this set a single step forward can be written as $\bar{P}\tilde{Y}$. Similarly, propagating the random-walk $t$ steps forward corresponds to raising $\bar{P}$ to the power of $t$ and applying it on $\tilde{Y}$ as $\bar{P}^t\tilde{Y}$. A single step of the random walk on the graph is given by[1]

$$\left[\bar{P}\tilde{Y}\right]_p = \sum_{l=1}^{M} \bar{p}(\tilde{\mathbf{y}}_p, \tilde{\mathbf{y}}_l)\tilde{\mathbf{y}}_l \tag{12}$$

and can be interpreted as averaging over similar time frames according to $\bar{p}$. Since the diffusion distance leads to clustering of frames into classes of speech and transient noises, the random-walk iteration approximately averages over all the frames from the same class, i.e.

$$\left[\bar{P}\tilde{Y}\right]_p \approx \sum_{l \in \mathcal{H}_i} \bar{p}(\tilde{\mathbf{y}}_p, \tilde{\mathbf{y}}_l)\tilde{\mathbf{y}}_l, \quad \tilde{\mathbf{y}}_p \in \mathcal{H}_i. \tag{13}$$

[1] $[X]_i$ extracts the $i$th row of the matrix $X$.

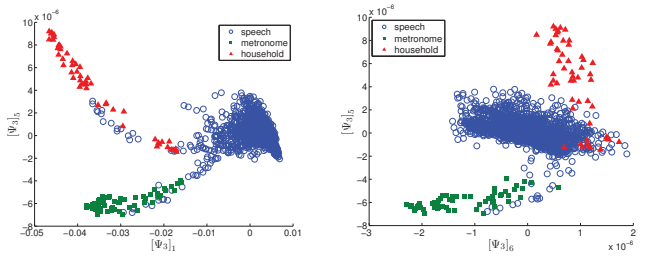


**Fig. 1**. Scatter plot of the diffusion maps embedding $\Psi_3$. Left: scatter plot of the 1st and 5th coordinates. Right: scatter plot of the 6th and 5th coordinates.

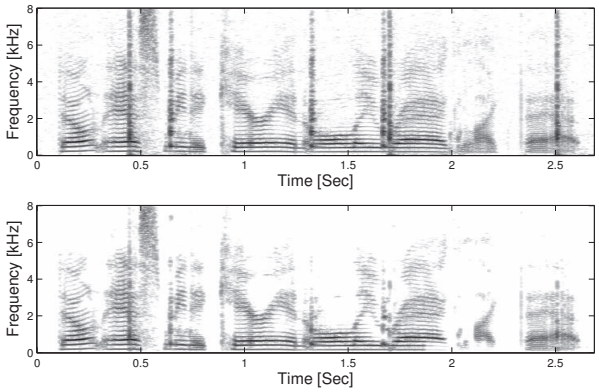


**Fig. 2**. Signal spectrograms. Top: the noisy measurement. Bottom: the enhanced speech obtain by the proposed algorithm.

As a result, since each transient type has the same spectral pattern (2), the transient instances are averaged together and enhanced. On the other hand, the "random" speech is averaged non-coherently, and therefore suppressed. After a few random-walk iterations, the instances of the different transient types may be extracted from the measurements, and their PSD can be estimated.

### 3.4. Transient Noise Suppression

The last part of the algorithm is similar to [1]. For enhancing the speech, we use an OM-LSA version with a modified noise PSD estimate. After a few iterations of the diffusion filter (13) we obtain an estimate of the PSD of the transient noises. Thus, we adjust the calculation of the optimal spectral gain to rely on a sum of the transient noises and the stationary noise PSD estimates. Since the calculation of the optimal spectral gain function is now controlled by both the stationary and transient noises, additional suppression of the transients is attainable. For more details regarding the optimal gain function derivation and estimation of the speech presence probability and the noise spectrum, we refer the readers to [6] and the references therein.

## 4. EXPERIMENTAL RESULTS

In this section we evaluate the performance of the proposed method. In our experiment we use recorded speech signals and transient noises, sampled at 16 KHz. The various recorded transient noise types are taken from [11]. We use STFT frames of 256 samples. The

**Table 1**. Enhancement Evaluation in Transient Occurrence Periods.

| Transient Noise | SNR Improvement [dB] | LSD Improvement [dB] |
|---|---|---|
| Metronome & Household Noise | 13.4 | 6.8 |
| Metronome & Kitchen Pocks | 12.9 | 6.6 |
| Door Knocks & Household Noise | 11.4 | 6.5 |

corresponding time frame length is 16ms, which is longer than the duration of the tested transients (approximately 10ms). In order to compare different noise signals of various durations and shapes, we maintain a constant value of the noise maximum amplitude, which equals to the maximum amplitude of the speech. For the diffusion maps embedding we use $\ell = 10$ dimensions and the number of diffusion steps is $t = 3$. These values were chosen since in empirical testing they yielded the best performance.

Measurements of 10s length are constructed according to (1), with additional computer-generated Gaussian white noise with SNR level of 20dB. In this experiment, the two corrupting transient noises are of a metronome and household sounds. The measurements consist of 20 transients of each type.

Figure 1 illustrates the diffusion maps embedding according to (10). Each point in Fig. 1 represents the embedding of a single vector $\tilde{\mathbf{y}}_p$ in $\mathbb{R}^2$. The frame content (speech, metronome, and household sounds) appears in different shapes (round, rectangular, triangular). In Fig. 1 (Left) we show a scatter plot of the 1st and 5th coordinates of the diffusion map (10) of the vectors, and in Fig. 1 (Right) we show a scatter plot of the 6th and 5th coordinates of the diffusion map of the vectors. We observe a clear clustering of the signals. However, when using merely two coordinates, we see some overlaps and outliers, i.e. points from one type in the area of another. Our empirical testing show that by using $\ell = 10$ dimensions, the diffusion maps embedding provides adequate separation of the points and minimal overlaps.

In Fig. 2 we show the spectrograms of part of the noisy measurement, and the denoised speech using the proposed algorithm. We observe in Fig. 2 (Bottom) that the proposed estimator achieves reduction of both transient types, while imposing very low distortion. This figure illustrates the suppression of the transients and the preservation of the speech. The results are evaluated using two common objective measures - signal to noise ration (SNR) and log spectral distortion (LSD), which are calculated only in periods with transients. The obtained SNR improvement is 13.4dB and the obtained LSD improvement is 6.8dB.

The described experiment was repeated with different pairs of transient noises taken from [11]. The obtained results are summarized in Table 1. We clearly observe that the proposed algorithm achieve significant speech enhancement in periods of transient interference. These results emphasize the advantage of the proposed algorithm in obtaining good transient noise reduction, while preserving speech components, even under the adverse conditions created by the presence of transient noise events. Moreover, they demonstrate the robustness of the algorithm to different transient types.

In addition, we tested the algorithm with a single type of transients similarly to [1] as well. The obtained results using the proposed algorithm are similar to the results obtain by the algorithm in [1]. Therefore, we did not present them in this paper, as they appear in [1] and [2]. However, we emphasize that the algorithm proposed in this paper enables a similar improvement in more challenging scenarios which include more than a single type of transient interference.

## 5. CONCLUSIONS

We introduced an algorithm for clustering and suppression of transient noises in speech signals. The proposed algorithm consists of three filters in cascade: a preprocessing OM-LSA for enhancing the transients, non-local filter for estimating the transients PSD, and OM-LSA for suppressing the transients and enhancing the speech. Here, we improve and extend an existing work by introducing two new concepts. The first relies on the observation that speech is slowly varying compared to transients, as (pseudo) stationary noise is slowly varying compared to speech. Hence, we propose to employ a stationary noise PSD estimation algorithm, equipped with two sliding windows and configured to track rapid variations, for estimating the speech PSD. The second concept is based on embedding the original measurements in a new space. We employ *diffusion maps*, which is a state-of-the-art method for manifold learning and dimensionality reduction, to obtain clustering of the transients and a robust affinity metric. The clustering capability of this approach was demonstrated, and a more comprehensive study will be conducted in future work. Experimental results show that the proposed algorithm can successfully handle simultaneous suppression of several transients, while maintaining the speech undistorted.

## 6. REFERENCES


[1] R. Talmon, I. Cohen, and S. Gannot, "Transient noise reduction for speech enhancement using diffusion filters," *to appear in IEEE Trans. Audio, Speech Lang. Process.*, 2010.

[2] R. Talmon, I. Cohen, and S. Gannot, "Speech enhancement in transient noise environment using diffusion filtering," *Proc. 35th IEEE Internat. Conf. Acoust. Speech Signal Process., ICASSP-2010, Dallas, Texas*, pp. 4782–4785, Mar. 2010.

[3] A. Singer, Y. Shkolnisky, and B. Nadler, "Diffusion interpretation of non local neighborhood filters for signal denoising," *SIAM J. Imaging Sciences*, vol. 2, no. 1, pp. 118–139, 2009.

[4] R. Coifman and S. Lafon, "Diffusion maps," *Applied and Computational Harmonic Analysis*, vol. 21, pp. 5–30, Jul. 2006.

[5] S. Vaseghi, *Advanced Digital Signal Processing and Noise Reduction*, John Wiley & Sons Ltd., 3rd edition, 2006.

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

[7] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Process. Let.*, vol. 9, no. 1, pp. 12–15, Jan. 2002.

[8] F. R. K. Chung, *Spectral Graph Theory*, CBMS-AMS, 1997.

[9] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[10] S. Lafon, Y. Keller, and R. R. Coifman, "Data fusion and multicue data matching by diffusion maps," *IEEE Tran. on Pattern Analisys and Machine Intelligence*, vol. 28, no. 11, pp. 1784–1797, Nov. 2006.

[11] [online]. Available: http://www.freesound.org.