SPEECH ENHANCEMENT IN TRANSIENT NOISE ENVIRONMENT USING DIFFUSION FILTERING

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 transient noise reduction algorithm for speech signals that relies on non-local diffusion filtering. By exploiting the repetitive nature of transient noises we proposed a simple and efficient algorithm, which enabled suppression of various noise types. In this paper, we incorporate a modified diffusion operator in order to obtain a more robust algorithm and further enhancement of the speech. We demonstrate the performance of the modified algorithm and compare it with a competing solution. We show that the proposed algorithm enables improved suppression of various transient interferences without any further computational burden.

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

1. INTRODUCTION

Traditional speech enhancement approaches usually consist of two components: noise power spectrum estimation and estimation of the desired clean speech signal [1] [2]. These methods are based on two fundamental assumptions. The first is that the noise spectrum remains stationary during the observation interval, or slowly varying compared to the spectrum of the speech signal. The second assumption is that the speech signal is not continuously present during the whole observation interval. Accordingly, a simple approach for estimating the noise spectrum is to average the noisy measurement over periods where the speech is absent. Then, by using the noise power spectrum estimate, the speech signal can be retrieved based on a predefined statistical model.

The assumption of pseudo-stationary noise signal poses a major limitation on these traditional algorithms, making them inadequate in many transient noise environments. Among them we mention noise originating from engines, keyboard typing, construction operations, bells, knocking, rings, hammering, etc. In [3] we presented a novel approach for transient noise reduction that relies on non-local (NL) filtering [4]. The proposed algorithm 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. The algorithm consists of three stages. In the first stage, the transient noise is enhanced relying on the strong correlation in time of the speech signal and the discontinuity of the transient noise. In the second stage, a NL filter is employed to extract the transient noise signal. In the third stage, the optimally modified log spectral amplitude (OM-LSA) estimator [2], equipped with a modified noise power spectral density (PSD) estimator, is utilized for the enhancement of the speech. The noise power spectrum estimate is based on the extracted transient signal from the second stage.

In this paper we incorporate a modified NL filter with superior characteristics. In particular, the modified filter is associated with both backward and forward diffusion, which enables simultaneous denoising and sharpening of the desired signal. In addition, it enables to employ a larger number of denoising steps, yielding a more robust algorithm and a more accurate extraction of the transient signal. We demonstrate the improved robustness attained by the modified algorithm on various transient noise types, and show that better results are obtained compared to both the OM-LSA, and the algorithm presented in [3]. This paper is organized as follows. In Section 2, we formulate the problem. In Section 3, the transient noise reduction algorithm is presented. In Section 4, we elaborate on the proposed modification and discuss its characteristics. Finally, in Section 5, we show experimental results that demonstrate the advantages of the proposed method.

2. PROBLEM FORMULATION

Let $y(n)$ be a measured speech signal contaminated with additive noise

$$y(n) = s(n) + d(n) \tag{1}$$

where $s(n)$ is a speech signal and $d(n)$ is a transient noise signal. It is worthwhile noting that the presence of additional stationary noise would not change significantly the derivation of the algorithm, however we omit it for simplicity.

The observation interval is divided into M short-time frames of length N. We assume that the speech signal is modeled as an auto regressive (AR) process in short-time frames. Accordingly, in each time frame $p = 1, \ldots, M$, the source signal is an AR process, given by

$$s(n) = \sum_{l=1}^{L} a_{l;p} s(n-l) + w(n) \tag{2}$$

where $w(n)$ is a white noise excitation signal with zero mean and variance σ_w^2, and $\{a_{l;p}\}_{l=1}^{L}$ are the AR coefficients of frame p.

The transient noise consists of short duration pulses with random amplitudes. It may be written as the output of a filter, excited by an amplitude-modulated random binary sequence [5]:

$$d(n) = h_n * (b(n)v(n)) \tag{3}$$

This research was supported by the Israel Science Foundation (grant no. 1085/05).

where $b(n)$ is a binary valued random sequence of time occurrences of the transient noise, e.g. a Poisson process, $v(n)$ is a continuous valued Gaussian process of transient amplitude, and h_n is an impulse response of a filter that determines the duration and shape of each transient event. We assume that in a single time frame no more than one transient event exists.

3. PROPOSED ALGORITHM

3.1. Transient Noise Enhancement

We utilize the differences between the transient noise and the AR source signal. The transient noise, modeled as a short duration pulse, introduces discontinuity in the signal. Thus, decorrelating the noisy measurement $y(n)$ in each time frame using the AR parameters of the source signal has the following effect. First, the scale of the source signal amplitude is reduced to almost that of the original excitation signal, whereas the scale of the transient noise remains unchanged or increased. Second, the source signal is decorrelated, whereas the transient noise is smeared. Let $\tilde{y}_p(n)$ be the decorrelated measurement in time frame p

$$\tilde{y}_p(n) = w(n) + \tilde{d}_p(n) \tag{4}$$

written as the sum of the source excitation signal $w(n)$, and a smeared version of the transient noise $\tilde{d}_p(n)$, given by

$$\tilde{d}_p(n) = d(n) - \sum_{l=1}^{L} a_{l;p} d(n-l). \tag{5}$$

We apply the short-time Fourier transform (STFT) in order to further enhance the difference between the transient noise and the source. From (4) and (5), we have that the STFT of the decorrelated signal in time frame p and frequency bin k is

$$\tilde{y}_{p,k} = w_{p,k} + (1 - A(p,k))\, d_{p,k} \tag{6}$$

where $w_{p,k}$ is the STFT of the excitation signal, $A(p,k)$ is the transfer function of the source AR filter (under the multiplicative transfer function (MTF) approximation), and $d_{p,k}$ is the STFT of the transient noise. Let H_0 denote the set of time frames free of transient noise occurrence and let H_1 denote the set of time frames that contain a transient occurrence. Then, from (3) we obtain that $d_{p,k}$ can be written as

$$d_{p,k} = \begin{cases} H(k) v_{p,k} & p \in H_1 \\ 0 & p \in H_0 \end{cases} \tag{7}$$

where $H(k)$ is the MTF approximation of the transient noise system h_n, $v_{p,k}$ is the transform of $(b(n)v(n))$, and $b(n)$ in frame $p \in H_1$ is an impulse.

3.2. Transient Noise Extraction using Nonlocal Filter

Equations (6) and (7) imply that all time frames contain the AR source excitation signal, whereas only time frames that contain transient occurrences have distinct shape, which may vary according to the specific noise type. We note that silent segments are disregarded in our analysis. The divergence between time frame shapes is exploited in order to extract the transient noise from the decorrelated measurement. Let $\tilde{\mathbf{y}}$ be an $M \times N$ matrix, defined as

$$\tilde{\mathbf{y}} = [\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_M]^T \tag{8}$$

where $\tilde{y}_p$ is a vector of the STFT samples from all frequency bins of the pth time-frame of the decorrelated signal (6), given by

$$\tilde{y}_p = [\tilde{y}_{p,0}, \ldots, \tilde{y}_{p,N-1}]^T. \tag{9}$$

Let $k : \mathbb{C}^N \times \mathbb{C}^N \rightarrow \mathbb{R}^+$ be a kernel representing a notion of similarity between two vectors $\tilde{y}_p$ (time frames). Based on the relation defined by the kernel, we form a weighted graph or a Euclidean manifold, where the time frame vectors are the vertices and the kernel sets the weights of the edges connecting the vectors. Generally, the choice of the specific kernel function is application-oriented to yield meaningful connections that convey the local geometry of the data. Then, a Markov random walk on the manifold is defined by [6]

$$p(\tilde{y}_p, \tilde{y}_l) = \frac{k(\tilde{y}_p, \tilde{y}_l)}{d(\tilde{y}_p)} \tag{10}$$

where $d(\tilde{y}_p) = \sum_{l=1}^{M} k(\tilde{y}_p, \tilde{y}_l)$. Consequently, $p(\tilde{y}_p, \tilde{y}_l)$ represents the probability of transition in a single step from node $\tilde{y}_p$ to node $\tilde{y}_l$. Let K denote the matrix corresponding the kernel function k, and let $P = D^{-1}K$ be the matrix corresponding to the transition probability function p, both defined on the vectors $\tilde{y}_p$, where D is a diagonal matrix with $D_{pp} = d(\tilde{y}_p)$. Thus, advancing the random walk forward by a single step can be written simply as matrix multiplication $P\tilde{\mathbf{y}}$. Consequently, running the random walk t steps forward is equivalent to $P^t\tilde{\mathbf{y}}$. Advancing the Markovian process a single step forward is equivalent to averaging over similar time frames (in the kernel sense), and hence constitutes a "denoising" iteration [4], which can be written as[1]

$$\hat{\tilde{d}}_p = [P\tilde{\mathbf{y}}]_p^T = \sum_l p(\tilde{y}_p, \tilde{y}_l)\tilde{y}_l \tag{11}$$

where $\hat{\tilde{d}}_p$ is the estimation of the smeared transient noise signal in the STFT domain after a single iteration. It is worthwhile noting that performing consecutive iterations may enhance the signal further. Let $\hat{\tilde{d}}_p^t = \left[P^t\tilde{\mathbf{y}}\right]_p^T$ denote the extracted smeared transient noise signal after performing t denoising iterations.

The choice of the kernel is of key importance in this method. We use the following Gaussian kernel, as defined in [3]

$$k\left(\tilde{y}_p, \tilde{y}_l\right) = \exp\left\{-\frac{\|\phi_{\tilde{y}}(p) - \phi_{\tilde{y}}(l)\|^2}{2\sigma^2}\right\} \tag{12}$$

where $\phi_{\tilde{y}}(p)$ is the short-time PSD of the pth frame of the decorrelated measurement (4). This specific choice of the kernel is motivated by the desire to exploit the reoccurring distinct shape of time frames containing a transient event, which may be conveyed by the frame PSD $\phi_{\tilde{y}}(p)$. This particular kernel leads to the following result. Time frames that contain transient noise occurrences are similar (in the kernel sense) to other frames that contain transient noise of the same shape. On the other hand, time frames free of transient noise are similar to other frames free of transient noise. In either case, when applying (11), the speech excitation signal $w_{p,k}$, which has random characteristics, is summed destructively, and the resulting frame is "denoised" from the excitation signal. As a consequence, the transient noise signal is *extracted*.

By applying inverse filtering to (5) with the spectral envelope of the AR source $1/(1 - A(p,k))$ on the extracted transient noise, we obtain an estimate of the transient signal in the STFT domain $\hat{d}_{p,k}$. Since the kernel is based only on the spectral shape, it provides an estimate of the transient signal short-time PSD rather than

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

an estimate of the signal itself. Accordingly, $\hat{\phi}_d(p,k)$ denotes the short-time PSD estimate of the transient noise, calculated based on the extracted transient noise signal from the output of the diffusion filter.

3.3. OM-LSA with a Modified Noise Spectrum Estimator

For enhancing the speech, we use an OM-LSA version equipped with a modified noise power spectrum estimate. From the output of the diffusion filter we obtain an estimate of the PSD of the transient noise signal $\hat{\phi}_d(p,k)$. Thus, we adjust the calculation of the optimal spectral gain function to rely on a sum of the transient noise PSD estimate $\hat{\phi}_d(p,k)$ and the stationary noise PSD estimate, obtained using the MCRA approach [7]. The calculation of the optimal spectral gain function is controlled by both the stationary and transient noise parts, and thus, attenuation of transient occurrences 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 [2] and the references therein.

4. A MODIFIED DIFFUSION FILTERING

By using spectral decomposition of P, t consecutive denoising iterations can be presented as

$$\left[P^t\tilde{\mathbf{y}}\right]_{p,k} = \sum_{j=0}^{M-1} \lambda_j^t b_{j,k}\psi_j(p) \tag{13}$$

where $\{\lambda_j, \psi_j\}$ are the eigenvalues and right eigenvectors of P, and $b_{j,k}$ are the inner products between the left eigenvectors φ_j and the decorrelated signal at frequency bin k. From (13), we observe that the decay rate of the eigenvalues has an impact on both the number of significant components and on the number of denoising iterations.

Now, consider the operator $P^{(1)} = 2P - P^2$ proposed in [4], which has the same eigenvectors as P, but its eigenvalues satisfy $2\lambda_j - \lambda_j^2 = 1 - (1-\lambda_j)^2$. Therefore the new operator has a much smaller suppression of the large eigenvalues[2]. Naturally, this modification can be performed recursively, yielding the following family of operators

$$P^{(i+1)} = 2P^{(i)} - P^{(i)^2} \tag{14}$$

where the suppression of the large eigenvalues decreases with each iteration i. Thus, replacing P with $P^{(i)}$ in (13) yields a larger number of significant components and enables to apply more denoising iterations.

Another characteristic of these modified operators emerges from [8], where it was shown that P converges to $I + \mathcal{L}$, where $\mathcal{L}$ is the backward Fokker-Planck diffusion operator. Therefore the modified operator converges to

$$P^{(1)} = 2P - P^2 = I - (I-P)^2 \approx (I-\mathcal{L})(I+\mathcal{L}). \tag{15}$$

Accordingly, by using $P^{(1)}$, each denoising iteration implies running the diffusion forward (destructive summation of the excitation signal), followed by running the diffusion backward (sharpening the transient signal).

[2]It can be shown that the eigenvalues satisfy $1 = \lambda_0 > \lambda_1 \geq \ldots \geq 0$.

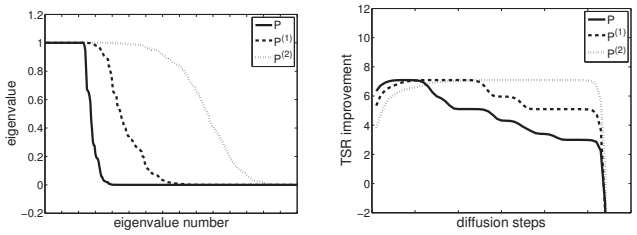

Fig. 1. Left: eigenvalues of the diffusion operators. Right: the TSR improvement (in dB) obtained as a function of the number of denoising steps.

Table 1. Evaluation of the Transient Signal Estimation (in dB).

Noise Type	Input TSR	Output TSR	
		P	$P^{(1)}$
Metronome	-8.7	5.5	7.4
Door Knocks	-6.7	4.0	4.7
Kitchen Pocks	-0.3	6.0	8.2
Household Clicks	-7.5	6.6	8.9

5. EXPERIMENTAL RESULTS

In this section we evaluate the performance of the proposed method with the modified operators and compare the results with the results of the OM-LSA estimator.

In the first experiment we generate signals according to the time domain model. The source signal is simulated as a *stationary* AR source (2), and the transient occurrences are determined according to the Gaussian-Poisson distributions (3).

For measuring the performance of the transient signal extraction we use the transient to signal ratio (TSR) defined as

$$\text{TSR} = 10\log_{10}\frac{\mathbb{E}\left\{d^2(n)\right\}}{\mathbb{E}\left\{(\hat{d}(n)-d(n))^2\right\}}. \tag{16}$$

Figure 1 presents the results of the first experiment obtained by using P, $P^{(1)}$ and $P^{(2)}$ as diffusion operators. We present the eigenvalues of the modified operators. According to the analysis in Section 4, slower decay of eigenvalues in an inverted parabolic shape is observed when applying the modified operators. In Fig. 1 we also present the TSR improvement as a function of diffusion steps. First, we observe that all operators enable approximately the same maximum performance. Second, the modified operators enable broader range of diffusion steps, making the algorithm more robust to further denoising and to an arbitrary choice of the number of steps. It is worthwhile noting that it does not imply an additional computation burden since the diffusion process may be implemented via the spectral decomposition (13).

In the second experiment we use recorded speech signals and transient noises, sampled at 16 KHz. The various recorded transient noises are taken from [9]. The measurements are constructed according to (1), with additional low variance computer generated Gaussian white noise. We use short-time frames of length 256 both for the LPC estimation and for the STFT. In each time frame, we estimate an AR envelope consisting of $L = 20$ coefficients. 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.

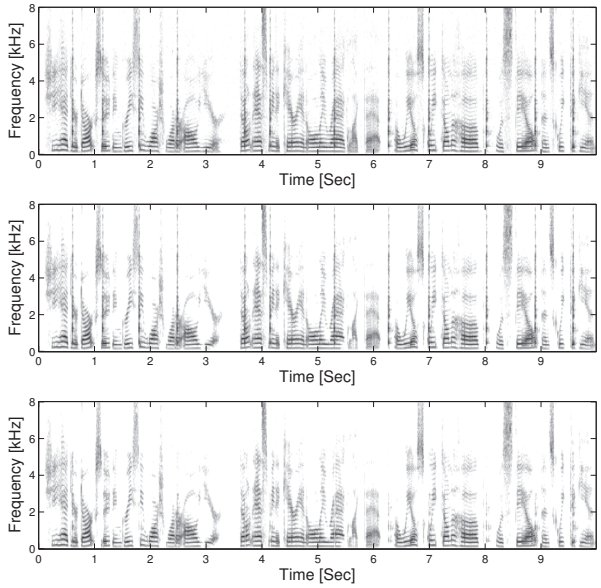

Fig. 2. Signal spectrograms. Top: the noisy measurement. Middle: the enhanced speech obtain by the OM-LSA. Bottom: the enhanced speech obtain by the proposed algorithm.

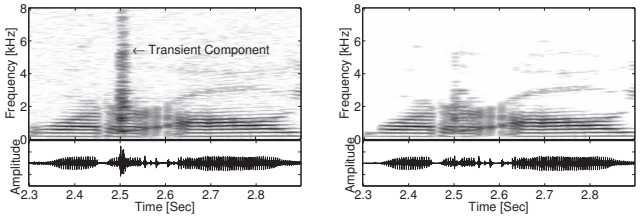

Fig. 3. Signal spectrograms and waveforms near the transient component at 2.5s. Left: the noisy measurement. Right: the enhanced speech obtain by the proposed algorithm.

Table 1 compares the TSR obtained by the proposed method using P and $P^{(1)}$ as diffusion operators, at the input and output of the algorithm second stage. As shown, usage of the modified operator enables improved TSR. These improved results were enabled by the characteristics of the modified operator. When using $P^{(1)}$, optimal results were obtained by employing a larger number of diffusion steps ($131,072$ compared to 128). Figure 2 shows spectrograms of the noisy measurement of a speech signal contaminated by metronome interference, and the enhanced speech obtained by the OM-LSA and the proposed algorithm (using $P^{(1)}$ as the diffusion operator). It is worthwhile noting that we randomly change the gaps between the metronome transient occurrences in order to fully demonstrate the proposed algorithm robustness. We notice that in the OM-LSA output the metronome signal is not suppressed, and the noisy measurement remains unchanged. However, we observe at the output of the proposed algorithm, that the metronome percussive events were completely removed, while maintaining the speech components undistorted. Figure 3 zooms in to the area near the transient event at 2.5s, and further illustrates the removal of the transient component and the preservation of the speech, obtained by the proposed method.

Table 2. Enhancement Evaluation in Transient Occurrence Periods (in dB).

Noisy Type	Input SNR	Output SNR		
		OM-LSA	P	$P^{(1)}$
Metronome	-9.50	-9.47	0.08	1.46
Door Knocks	-7.66	-7.43	-0.81	2.05
Kitchen Pocks	-13.88	-13.49	-6.97	-4.09
Household Clicks	-13.13	-12.96	-4.18	-1.55

In order to evaluate the enhancement of the speech we measure the commonly used signal to noise ratio (SNR). Table 2 summarizes the SNR, calculated only in transient occurrence time frames. We observe that the proposed algorithm obtains superior SNR compared to the OM-LSA. In addition, the use of $P^{(1)}$ increases the speech enhancement. As described before, the improved performance is obtained by employing more iterations, which is enabled by $P^{(1)}$ characteristics.

6. CONCLUSIONS

We have presented a non-local diffusion filter for handling speech corrupted with transient interferences. The proposed approach exploits the intrinsic geometric structure of the measurements. In particular, it relies on the divergence between speech components and sharp impulses of repeating transient noise occurrences. By relying on the diffusion interpretation of the non-local filters, we utilize a modified diffusion operator with more attractive characteristics. Experimental results show that for repetitive and short transient occurrences, the proposed algorithm achieves superior performance, obtaining better transient noise extraction and further enhanced speech signal. In addition, we demonstrate the robustness of the proposed method using various types of transient interferences.

7. REFERENCES

[1] Y. Ephraim and D. Malah, "Speech enhancememt using a minimum mean square error short time spectral amplitude estimator," *IEEE Trans. Acoust., Speech Signal Process.*, pp. 1109–1121, Dec. 1984.

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

[3] R. Talmon, I. Cohen, and S. Gannot, "Transient noise reduction for speech enhancement using diffusion filters," *submitted to IEEE Trans. Audio, Speech, Lang. Process.*, 2009.

[4] 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.

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

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

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

[9] [Online]. Available: http://www.freesound.org.