Computer Speech & Language 70 (2021) 101245

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

Adaptive line enhancer for nonstationary harmonic noise reduction

Aviva Atkins*[,a], Israel Cohen[a], Jacob Benesty[b]

[a] Andrew and Erna Viterbi Faculty of Electrical and Computer Engineering, Technion–Israel Institute of Technology, Haifa 3200003, Israel
[b] INRS-EMT, University of Quebec, 800 de la Gauchetiere Ouest, Suite 6900, Montreal, QC H5A 1K6, Canada

ARTICLE INFO

Article History:
Received 3 January 2020
Revised 6 February 2021
Accepted 2 May 2021
Available online 14 May 2021

Keywords:
Speech enhancement
Noise reduction
Single channel
Adaptive line enhancer
Nonstationary noise

ABSTRACT

In this paper, we propose a frequency-domain adaptive line enhancer (ALE) to reduce nonstationary harmonic noise, such as medical equipment beeps, from a noisy speech signal captured by a single microphone. The reduction of nonstationary noise is very challenging, with the tradeoff between noise reduction and speech distortion, often resulting with much noise residuals. The proposed ALE is a combination of the commonly-used forward adaptive linear filter and a non-causal backward adaptive linear filter used together with an indicator for the presence of transient noise. The proposed combined filter results in less noise residuals while preserving the speech components. We compare the proposed approach to conventional and recent methods, and show that it can outperform these methods, achieving lower distortion, more noise reduction, and overall better speech quality and intelligibility.

1. Introduction

Acoustic background noise is an important factor that degrades both the perceptual speech quality/intelligibility and, unfortunately, it is common in all practical situations, such as telecommunications, teleconferencing, and human-machine interfaces. The process of suppressing this additive noise is known as noise reduction or alternatively as speech enhancement, and it is a fundamental problem that has been researched extensively over the past few decades, e.g., (Benesty et al., 2009; Loizou, 2007; Cohen et al., 2010; Benesty et al., 2018) and references therein. Many methods have been proposed for noise reduction, among them spectral subtraction techniques (Boll, 1979; Miyazaki et al., 2012), optimal filtering including the notorious Wiener filter (Chen et al., 2006; Huang et al., 2014; Lim and Oppenheim, 1979), statistical-model-based algorithms (Ephraim and Malah, 1984; Cohen and Berdugo, 2001; Cohen, 2005; Cohen and Gannot, 2008), subspace methods (Ephraim and Van Trees, 1995; Hu and Loizou, 2003), binary mask methods (Kim et al., 2009; Wang and Brown, 2006), and data-driven supervised learning methods namely deep learning (DL) based on deep neural networks which have recently gained much popularity (Wang and Chen, 2018; Xu et al., 2014; 2015; Zhang and Wang, 2016).

The noise reduction problem is traditionally formulated as a linear filtering problem, where we pass the observed noisy signal through a filter to obtain an estimate of the clean speech signal. In the design of this filter, the aim is to achieve maximal noise suppression without introducing noticeable speech distortion. When a single microphone is used to capture the noisy speech, studies (Hu and Loizou, 2007; Loizou and Kim, 2011) have shown that traditional approaches generally do not provide improvement to the intelligibility while they do improve the speech perceptual quality. For DL algorithms it can also be challenging to improve on both intelligibility and quality; though improving intelligibility, they can suffer from poor sound quality (Lee and Kang, 2019; Zhang and Bhowmik, 2018). Recent studies (Lee et al., 2018; Lee and Kang, 2019; Williamson et al., 2016) include

*Corresponding author.
E-mail addresses: aatkins@campus.technion.ac.il (A. Atkins), icohen@ee.technion.ac.il (I. Cohen), benesty@emt.inrs.ca (J. Benesty).

https://doi.org/10.1016/j.csl.2021.101245
0885-2308/

the processing of the phase spectra in order to achieve better speech quality, alternatively, others (Koizumi et al., 2017; Zhao et al., 2018) incorporate perceptual measures into the training method.

An adaptive line enhancer (ALE) (Ramli et al., 2012; Widrow et al., 1976) modifies both the magnitude and phase, and so can potentially improve both intelligibility and quality. The ALE is a degenerate form of adaptive noise canceling (ANC), both first introduced in (Widrow et al., 1975). They work on the principle of correlation cancellation, suppressing noise from a signal using adaptive filters that self adjust their parameters. They involve producing an estimate of the noise by filtering a reference signal and then subtracting this noise estimate from the primary input containing both the speech signal and noise. They have the advantage of updating the filter coefficients automatically with no need for a-priori knowledge of noise and speech signals. While the ANC requires a reference noise signal that is highly correlated with the noise signal, the ALE consists of a single microphone and a delay element to produce a delayed version of the noisy signal to be used as the reference signal. The delay enables separation of the periodic and stochastic components in a signal; it de-correlates the stochastic components between the input and the delayed input, while leaving the periodic components correlated. The periodic components are extracted by the adaptive filter, usually using a normalized least-mean-square (NLMS) algorithm (Haykin, 2014). Depending on the characteristics of the noise, two different approaches are used. The first is to directly estimate the speech signal from wide-band background noise. In this case, the ALE estimates the speech signal using the pitch period for the delay (Sasaoka et al., 2006). The second case is when the noise is harmonic, e.g., ventilation fan noise and vehicle engine noise. The filter estimates the noise which is then subtracted from the input to obtain the desired speech signal. In this case, the ALE is typically used in conjunction with a traditional noise estimator or even with another ALE to remove any additional wide-band background noise present (Sasaoka et al., 2009). The ALE can also be implemented in the short-time Fourier transform (STFT) domain (Nakanishi et al., 2013), where adaptive filters have less computational cost and can be calculated separately for each frequency bin. The performance of the ALE is highly dependent on the de-correlation delay parameter and the step size of the adaptive filter, which are either fixed or heuristically optimized; hence noise residuals remain, particularly for nonstationary noise.

Recently, Taghia and Martin (Taghia and Martin, 2016) have shown that an ALE with a frequency-dependent step size based on mutual information (MI) can detect the presence of harmonic noise and reduce it. To deal with nonstationary noise, they implemented the algorithm in a block-wise manner and assumed the span of the stationarity of the noise signal is at least as large as the block length. This is not an effective solution for highly nonstationary noise signals such as medical equipment beeps or alarm sounds.

In this paper, we propose using a combination of both a forward linear predictor (FLP) and a non-causal backward linear predictor (BLP), both implemented with the NLMS algorithm, to better address the nonstationarity of the harmonic noise. The FLP on its own (this is the standard NLMS implementation) reduces the noise transient after a delay determined by the adaptive filter step size and de-correlation delay parameters, and hence residuals of the noise remain. In a similar manner, for the BLP on its own, residuals would remain at the end of the transient. By using a combination of the FLP and the BLP, we are able to increase the reduction span of the noise transient, thus reducing the amount of noise residuals. We use an indicator for noise presence to apply the filters only when noise is present, to reduce the amount of distortion to the speech signal. We also apply a set of changing filter lengths, taking the maximal filter length available according to the indicator, to ensure the combined filter spans throughout the noise transient. With this approach, we are able to achieve higher noise reduction and lower signal distortion, improving the speech quality and intelligibility compared to other methods.

The rest of the paper is organized as follows. In Section 2, we describe the signal model and the ALE system. In Section 3, we derive the relevant performance measures. In Section 4, we develop the Wiener filter and present our proposed combined filter. The experimental results are then presented in Section 5, and finally the conclusions are drawn in Section 6.

2. Problem formulation

We consider the following signal model:

$$y(n) = x(n) + v(n), \tag{1}$$

where n is the discrete time index, $y(n)$ is the observed noisy signal, $x(n)$ is the zero-mean desired clean speech signal, and $v(n)$ is the zero-mean noise signal. We assume that $x(n)$ and $v(n)$ are real and uncorrelated signals. Using the STFT, (1) can be expressed as

$$Y(k,m) = X(k,m) + V(k,m), \tag{2}$$

where k (for $k = 0, 1, \ldots, K-1$) is the frequency index, m (for $m = 0, 1, \ldots, M-1$) is the frame index, and $Y(k,m)$, $X(k,m)$, and $V(k,m)$ are the STFTs of $y(n)$, $x(n)$, and $v(n)$, respectively. As $x(n)$ and $v(v)$ are uncorrelated per the assumption, the variance of $Y(k,m)$ is

$$\begin{aligned} \phi_Y(k,m) &= E\left[\left|Y(k,m)\right|^2\right] \\ &= E\left[\left|X(k,m)\right|^2\right] + E\left[\left|V(k,m)\right|^2\right] \\ &= \phi_X(k,m) + \phi_V(k,m), \end{aligned} \tag{3}$$

where $E[\cdot]$ denotes mathematical expectation. We apply a delay τ to the observed signal $Y(k,m)$ and pass this delayed signal through a complex-valued filter $\mathbf{h}(k,m)$ of length L:

$$Z(k,m) = \mathbf{h}^H(k,m)\mathbf{y}(k,m-\tau), \tag{4}$$

where

$$\mathbf{h}(k,m) = [H_0(k,m), H_1(k,m), \ldots, H_{L-1}(k,m)]^T, \tag{5}$$

superscripts ^{H} and ^{T} are the conjugate-transpose and transpose operators, respectively, and

$$\mathbf{y}(k,m-\tau) = [Y(k,m-\tau), Y(k,m-\tau-1), \ldots, Y(k,m-\tau-L+1)]^T. \tag{6}$$

The vectors $\mathbf{x}(k,m-\tau)$ and $\mathbf{v}(k,m-\tau)$ are defined in a similar fashion to $\mathbf{y}(k,m-\tau)$, so we get

$$\mathbf{y}(k,m-\tau) = \mathbf{x}(k,m-\tau) + \mathbf{v}(k,m-\tau). \tag{7}$$

As a result, the error signal is defined as

$$\begin{aligned} E(k,m) &= Y(k,m) - Z(k,m) \\ &= Y(k,m) - \mathbf{h}^H(k,m)\mathbf{y}(k,m-\tau). \end{aligned} \tag{8}$$

We consider decomposing the signal $X(k,m-\tau)$ into two orthogonal components: a part which is correlated to the desired signal, $X(k,m)$, and another part which is uncorrelated to the desired signal, $X(k,m)$, and hence will be considered as an interference component, i.e.,

$$X(k,m-\tau) = \Gamma_X^*(k,m,\tau)X(k,m) + X'(k,m,\tau), \tag{9}$$

where

$$\Gamma_X(k,m,\tau) = \frac{E[X(k,m)X^*(k,m-\tau)]}{E\left[\left|X(k,m)\right|^2\right]} \tag{10}$$

is the inter-frame correlation coefficient of the signal $X(k,m)$ and the interference is

$$X'(k,m,\tau) = X(k,m-\tau) - \Gamma_X^*(k,m,\tau)X(k,m), \tag{11}$$

with $E[X(k,m)X'^*(k,m,\tau)] = 0$ and the superscript * being the complex-conjugate operator. In vector form we can write

$$\mathbf{x}(k,m-\tau) = \boldsymbol{\gamma}_X^*(k,m,\tau)X(k,m) + \mathbf{x}'(k,m,\tau), \tag{12}$$

where the signal correlation vector is

$$\begin{aligned} \boldsymbol{\gamma}_X(k,m,\tau) &= \left[\Gamma_X(k,m,\tau), \Gamma_X(k,m,\tau+1), \ldots, \Gamma_X(k,m,\tau+L-1)\right]^T \\ &= \frac{E[X(k,m)\mathbf{x}^*(k,m-\tau)]}{E\left[\left|X(k,m)\right|^2\right]} \end{aligned} \tag{13}$$

and the signal interference vector is

$$\begin{aligned} \mathbf{x}'(k,m,\tau) &= [X'(k,m,\tau), X'(k,m,\tau+1), \ldots, X'(k,m,\tau+L-1)]^T \\ &= \mathbf{x}(k,m-\tau) - \boldsymbol{\gamma}_X^*(k,m,\tau)X(k,m). \end{aligned} \tag{14}$$

We can implement a similar decomposition for the noise signal $V(k,m)$ to get

$$\mathbf{v}(k,m-\tau) = \boldsymbol{\gamma}_V^*(k,m,\tau)V(k,m) + \mathbf{v}'(k,m,\tau), \tag{15}$$

where $\boldsymbol{\gamma}_V(k,m,\tau)$ is the noise correlation vector and $\mathbf{v}'(k,m,\tau)$ is the noise interference. Plugging these vectors into $Z(k,m)$, we get four components which are uncorrelated, i.e.,

$$\begin{aligned} Z(k,m) &= \mathbf{h}^H(k,m)\mathbf{y}(k,m-\tau) \\ &= \mathbf{h}^H(k,m)\mathbf{x}(k,m-\tau) + \mathbf{h}^H(k,m)\mathbf{v}(k,m-\tau) \\ &= \mathbf{h}^H(k,m)\boldsymbol{\gamma}_X^*(k,m,\tau)X(k,m) + \mathbf{h}^H(k,m)\mathbf{x}'(k,m,\tau) \\ &\quad + \mathbf{h}^H(k,m)\boldsymbol{\gamma}_V^*(k,m,\tau)V(k,m) + \mathbf{h}^H(k,m)\mathbf{v}'(k,m,\tau) \end{aligned} \tag{16}$$

and the error signal can now be written as

$$\begin{aligned} E(k,m) = \;& Y(k,m) - Z(k,m) \\ = \;& X(k,m)\left[1 - \mathbf{h}^H(k,m)\boldsymbol{\gamma}_X^*(k,m,\tau)\right] - \mathbf{h}^H(k,m)\mathbf{x}'(k,m,\tau) \\ & + V(k,m)\left[1 - \mathbf{h}^H(k,m)\boldsymbol{\gamma}_V^*(k,m,\tau)\right] - \mathbf{h}^H(k,m)\mathbf{v}'(k,m,\tau). \end{aligned} \tag{17}$$

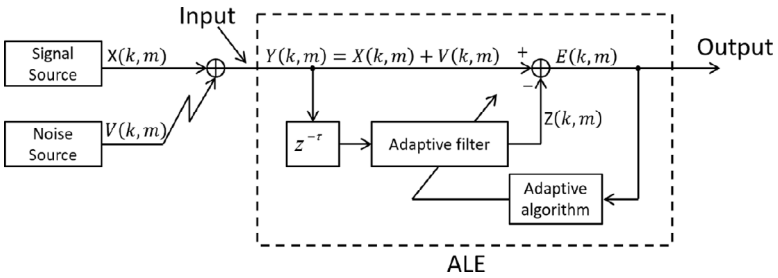

Fig. 1. STFT domain ALE system.

There are two scenarios; $Z(k,m)$ can either be an estimate of the noise or the desired signal, depending on which signal is still correlated given the delay τ. For the case we are investigating, where the noise is nonstationary harmonic, we will assume that the speech signal is no longer correlated, while the noise is still correlated, meaning that $Z(k,m)$ is the estimate of the noise and $E(k,m)$ is the estimate of the speech signal. The described system is shown in Fig. 1. The ideal conditions for this system are

$$\mathbf{h}^H(k,m)\boldsymbol{\gamma}_X^*(k,m,\tau) = 0, \tag{18}$$

$$\mathbf{h}^H(k,m)\boldsymbol{\gamma}_V^*(k,m,\tau) = 1, \tag{19}$$

and the inherent estimation error is

$$\epsilon_{\text{inherent}}(k,m) = \mathbf{h}^H(k,m)\mathbf{x}'(k,m,\tau) + \mathbf{h}^H(k,m)\mathbf{v}'(k,m,\tau). \tag{20}$$

We can write the speech estimate as follows:

$$X_{\text{est}}(k,m) = E(k,m) = X_{\text{fd}}(k,m) + X'_{\text{ri}}(k,m) + V_{\text{rn}}(k,m), \tag{21}$$

where the filtered desired signal is

$$X_{\text{fd}}(k,m) = X(k,m)\left[1 - \mathbf{h}^H(k,m)\boldsymbol{\gamma}_X^*(k,m,\tau)\right], \tag{22}$$

the residual interference is

$$X'_{\text{ri}}(k,m) = -\mathbf{h}^H(k,m)\mathbf{x}'(k,m,\tau), \tag{23}$$

and the residual noise is

$$V_{\text{rn}}(k,m) = V(k,m)\left[1 - \mathbf{h}^H(k,m)\boldsymbol{\gamma}_V^*(k,m,\tau)\right] - \mathbf{h}^H(k,m)\mathbf{v}'(k,m,\tau). \tag{24}$$

The ideal conditions (18)-(19) are met when the desired signal is no longer correlated with the delayed desired signal, while the noise remains correlated. If $\mathbf{h}^H(k,m)\boldsymbol{\gamma}_X^*(k,m,\tau) \neq 0$, we get distortion of the desired signal, while if $\mathbf{h}^H(k,m)\boldsymbol{\gamma}_V^*(k,m,\tau) \neq 1$, we get less noise reduction. This shows the importance of the delay parameter τ.

3. Performance measures

The narrow-band and full-band input SNRs are, respectively,

$$\text{iSNR}(k,m) = \frac{\phi_X(k,m)}{\phi_V(k,m)} \tag{25}$$

and

$$\text{iSNR}(m) = \frac{\sum_{k=0}^{K-1}\phi_X(k,m)}{\sum_{k=0}^{K-1}\phi_V(k,m)}. \tag{26}$$

The narrow-band output SNR is defined as the ratio of the variance of the filtered desired signal over the variance of the residual interference-plus-noise, i.e.,

$$\text{oSNR}[\mathbf{h}(k,m)] = \frac{\phi_{X_{\text{fd}}}(k,m)}{\phi_{X'_{\text{ri}}}(k,m) + \phi_{V_{\text{rn}}}(k,m)} \tag{27}$$

where

$$\begin{aligned}\phi_{X_{\text{fd}}}(k,m) &= E\left[\left|X_{\text{fd}}(k,m)\right|^2\right] \\ &= \phi_X(k,m)\left|1-\mathbf{h}^H(k,m)\boldsymbol{\gamma}_X^*(k,m,\tau)\right|^2,\end{aligned} \tag{28}$$

$$\begin{aligned}\phi_{X'_{\text{ri}}}(k,m) &= E\left[\left|X'_{\text{ri}}(k,m)\right|^2\right] \\ &= \mathbf{h}^H(k,m)\Phi_{\mathbf{x}'}(k,m,\tau)\mathbf{h}(k,m),\end{aligned} \tag{29}$$

and

$$\begin{aligned}\phi_{V_{\text{rn}}}(k,m) &= E\left[\left|V_{\text{rn}}(k,m)\right|^2\right] \\ &= \phi_V(k,m)\left|1-\mathbf{h}^H(k,m)\boldsymbol{\gamma}_V^*(k,m,\tau)\right|^2+\mathbf{h}^H(k,m)\Phi_{\mathbf{v}'}(k,m,\tau)\mathbf{h}(k,m),\end{aligned} \tag{30}$$

where we have defined the matrices:

$$\Phi_{\mathbf{x}'}(k,m,\tau) = E\left[\mathbf{x}'(k,m,\tau)\mathbf{x}'^H(k,m,\tau)\right], \tag{31}$$

$$\Phi_{\mathbf{v}'}(k,m,\tau) = E\left[\mathbf{v}'(k,m,\tau)\mathbf{v}'^H(k,m,\tau)\right]. \tag{32}$$

Expanding the term:

$$\begin{aligned}\Phi_{\mathbf{x}'}(k,m,\tau) &= E\left[\mathbf{x}'(k,m,\tau)\mathbf{x}'^H(k,m,\tau)\right] \\ &= E\left\{\left[\mathbf{x}(k,m-\tau)-\boldsymbol{\gamma}_X^*(k,m,\tau)X(k,m)\right]\left[\mathbf{x}^H(k,m-\tau)-X^*(k,m)\boldsymbol{\gamma}_X^T(k,m,\tau)\right]\right\} \\ &= \Phi_{\mathbf{x}}(k,m-\tau)-\phi_X(k,m)\boldsymbol{\gamma}_X^*(k,m,\tau)\boldsymbol{\gamma}_X^T(k,m,\tau).\end{aligned} \tag{33}$$

Similarly,

$$\Phi_{\mathbf{v}'}(k,m,\tau) = \Phi_{\mathbf{v}}(k,m-\tau)-\phi_V(k,m)\boldsymbol{\gamma}_V^*(k,m,\tau)\boldsymbol{\gamma}_V^T(k,m,\tau). \tag{34}$$

Plugging these into (27), we get

$$\begin{aligned}\text{oSNR}[\mathbf{h}(k,m)] = \phi_X(k,m)&\left|1-\mathbf{h}^H(k,m)\boldsymbol{\gamma}_X^*(k,m,\tau)\right|^2\times \\ &[\mathbf{h}^H(k,m)\Phi_{\mathbf{x}'}(k,m,\tau)\mathbf{h}(k,m)+ \\ &\phi_V(k,m)\left|1-\mathbf{h}^H(k,m)\boldsymbol{\gamma}_V^*(k,m,\tau)\right|^2+ \\ &\mathbf{h}^H(k,m)\Phi_{\mathbf{v}'}(k,m,\tau)\mathbf{h}(k,m)]^{-1}.\end{aligned} \tag{35}$$

For the ideal conditions (18)-(19), we have

$$\text{oSNR}_{\text{ideal}}[\mathbf{h}(k,m)] = \frac{\phi_X(k,m)}{\mathbf{h}^H(k,m)\Phi_{\mathbf{x}'}(k,m,\tau)\mathbf{h}(k,m)+\mathbf{h}^H(k,m)\Phi_{\mathbf{v}'}(k,m,\tau)\mathbf{h}(k,m)}. \tag{36}$$

We define the full-band output SNR as

$$\text{oSNR}[\mathbf{h}(m)] = \frac{\sum_{k=0}^{K-1}\phi_{X_{\text{fd}}}(k,m)}{\sum_{k=0}^{K-1}\phi_{X'_{\text{ri}}}(k,m)+\sum_{k=0}^{K-1}\phi_{V_{\text{rn}}}(k,m)}. \tag{37}$$

It can be verified (Benesty et al., 2009) that

$$\text{iSNR}[\mathbf{h}(m)] \leq \sum_{k=0}^{K-1}\text{iSNR}[\mathbf{h}(k,m)], \tag{38}$$

$$\text{oSNR}[\mathbf{h}(m)] \leq \sum_{k=0}^{K-1}\text{oSNR}[\mathbf{h}(k,m)]. \tag{39}$$

The noise reduction factor quantifies the amount of noise being rejected by the filter. It is defined as the ratio of the power of the noise at the sensor over the power of the noise remaining at the filter output. The noise reduction factor is usually expected to be lower bounded by 1, however, as we are adding the residual interference this is not necessarily the case. The higher the value, the more the noise is rejected. The narrow-band and full-band noise reduction factors are then

$$\begin{aligned}\xi_{nr}[\mathbf{h}(k,m)] &= \frac{\phi_V(k,m)}{\phi_{X'_{\text{ri}}}(k,m)+\phi_{V_{\text{rn}}}(k,m)}\\ &= \phi_V(k,m)\times[\mathbf{h}^H(k,m)\Phi_{\mathbf{x}'}(k,m,\tau)\mathbf{h}(k,m)+\\ &\quad \phi_V(k,m)\left|1-\mathbf{h}^H(k,m)\boldsymbol{\gamma}^*_V(k,m,\tau)\right|^2+\\ &\quad \mathbf{h}^H(k,m)\Phi_{\mathbf{v}'}(k,m,\tau)\mathbf{h}(k,m)]^{-1},\end{aligned} \tag{40}$$

and

$$\xi_{\text{nr}}[\mathbf{h}(m)] = \frac{\sum_{k=0}^{K-1}\phi_V(k,m)}{\sum_{k=0}^{K-1}\phi_{X'_{\text{ri}}}(k,m)+\sum_{k=0}^{K-1}\phi_{V_{\text{rn}}}(k,m)}. \tag{41}$$

For the ideal conditions (18)-(19):

$$\xi_{\text{nr,ideal}}[\mathbf{h}(k,m)] = \frac{\phi_V(k,m)}{\mathbf{h}^H(k,m)\Phi_{\mathbf{x}'}(k,m,\tau)\mathbf{h}(k,m)+\mathbf{h}^H(k,m)\Phi_{\mathbf{v}'}(k,m,\tau)\mathbf{h}(k,m)}. \tag{42}$$

In practice, the filter might distort the signal. To evaluate the level of this distortion, we define the narrow-band and full-band speech reduction factors, respectively, as

$$\begin{aligned}\xi_{\text{sr}}[\mathbf{h}(k,m)] &= \frac{\phi_X(k,m)}{\phi_{X_{\text{fd}}}(k,m)}\\ &= \frac{1}{\left|1-\mathbf{h}^H(k,m)\boldsymbol{\gamma}^*_X(k,m,\tau)\right|^2}\end{aligned} \tag{43}$$

and

$$\begin{aligned}\xi_{\text{sr}}[\mathbf{h}(m)] &= \frac{\sum_{k=0}^{K-1}\phi_X(k,m)}{\sum_{k=0}^{K-1}\phi_{X_{\text{fd}}}(k,m)}\\ &= \frac{\sum_{k=0}^{K-1}\phi_X(k,m)}{\sum_{k=0}^{K-1}\phi_X(k,m)\left|1-\mathbf{h}^H(k,m)\boldsymbol{\gamma}^*_X(k,m,\tau)\right|^2}.\end{aligned} \tag{44}$$

Thus the speech reduction factor is equal to 1 if there is no distortion and is greater than 1 when distortion occurs. We can clearly see from this ratio that the design of a filter that does not distort the speech signal is dependent on there being no remaining correlation.

Another way to measure the distortion of the desired signal due to the filtering is via the desired signal distortion index, which is defined as the mean-squared error (MSE) between the desired signal and the filtered desired signal normalized by the variance of the desired signal. The closer the distortion index is to 0, the less the distortion. The narrow-band and full-band desired signal distortion indexes are then

$$\begin{aligned}\nu_{\text{sd}}[\mathbf{h}(k,m)] &= \frac{E\left[\left|X(k,m)-X_{\text{fd}}(k,m)\right|^2\right]}{\phi_X(k,m)}\\ &= \frac{\phi_X(k,m)\left|\mathbf{h}^H(k,m)\boldsymbol{\gamma}^*_X(k,m,\tau)\right|^2}{\phi_X(k,m)} = \left|\mathbf{h}^H(k,m)\boldsymbol{\gamma}^*_X(k,m,\tau)\right|^2\end{aligned} \tag{45}$$

and

$$\nu_{\text{sd}}[\mathbf{h}(m)] = \frac{\sum_{k=0}^{K-1}E\left[\left|X(k,m)-X_{\text{fd}}(k,m)\right|^2\right]}{\sum_{k=0}^{K-1}\phi_X(k,m)}, \tag{46}$$

where for the ideal conditions there is no distortion. By making the appropriate substitutions, one can derive the following relationships:

$$\frac{\text{oSNR}[\mathbf{h}(k,m)]}{\text{iSNR}(k,m)} = \frac{\xi_{\text{nr}}[\mathbf{h}(k,m)]}{\xi_{\text{sr}}[\mathbf{h}(k,m)]}, \tag{47}$$

$$\frac{\text{oSNR}[\mathbf{h}(m)]}{\text{iSNR}(m)} = \frac{\xi_{\text{nr}}[\mathbf{h}(m)]}{\xi_{\text{sr}}[\mathbf{h}(m)]}. \tag{48}$$

4. Optimal filters

We define the narrow-band MSE as

$$\begin{aligned} J[\mathbf{h}(k,m)] &= E\left[\left|E(k,m)\right|^2\right] \\ &= E\left[\left|X_{\mathrm{fd}}(k,m)\right|^2\right] + E\left[\left|V_{\mathrm{rn}}(k,m)\right|^2\right] + E\left[\left|X'_{\mathrm{ri}}(k,m)\right|^2\right] \\ &= \phi_X(k,m)\left|1-\mathbf{h}^H(k,m)\boldsymbol{\gamma}_X^*(k,m,\tau)\right|^2 + \mathbf{h}^H(k,m)\Phi_{\mathbf{x}'}(k,m,\tau)\mathbf{h}(k,m) \\ &\quad + \phi_V\left|1-\mathbf{h}^H(k,m)\boldsymbol{\gamma}_V^*(k,m,\tau)\right|^2 + \mathbf{h}^H(k,m)\Phi_{\mathbf{v}'}(k,m,\tau)\mathbf{h}(k,m) \\ &= J_{\mathrm{d}}[\mathbf{h}(k,m)] + J_{\mathrm{r}}[\mathbf{h}(k,m)], \end{aligned} \tag{49}$$

where

$$\begin{aligned} J_{\mathrm{d}}[\mathbf{h}(k,m)] = \;& E\left[\left|X_{\mathrm{fd}}(k,m)\right|^2\right] \\ &= \phi_X(k,m)\left|1-\mathbf{h}^H(k,m)\boldsymbol{\gamma}_X^*(k,m,\tau)\right|^2 \end{aligned} \tag{50}$$

and

$$\begin{aligned} J_{\mathrm{r}}[\mathbf{h}(k,m)] = \;& E\left[\left|V_{\mathrm{rn}}(k,m)\right|^2\right] + E\left[\left|X'_{\mathrm{ri}}(k,m)\right|^2\right] \\ &= \phi_V\left|1-\mathbf{h}^H(k,m)\boldsymbol{\gamma}_V^*(k,m,\tau)\right|^2 + \mathbf{h}^H(k,m)\Phi_{\mathbf{v}'}(k,m,\tau)\mathbf{h}(k,m) \\ &\quad + \mathbf{h}^H(k,m)\Phi_{\mathbf{x}'}(k,m,\tau)\mathbf{h}(k,m). \end{aligned} \tag{51}$$

We can easily see the relation between the MSEs and some of the performance measures:

$$\mathrm{oSNR}[\mathbf{h}(k,m)] = \frac{J_{\mathrm{d}}[\mathbf{h}(k,m)]}{J_{\mathrm{r}}[\mathbf{h}(k,m)]} \tag{52}$$

and

$$\xi_{\mathrm{nr}}[\mathbf{h}(k,m)] = \frac{\phi_V(k,m)}{J_{\mathrm{r}}[\mathbf{h}(k,m)]}. \tag{53}$$

We can define the full-band MSE as

$$\begin{aligned} J[\mathbf{h}(m)] &= \frac{1}{K}\sum_{k=0}^{K-1} J[\mathbf{h}(k,m)] \\ &= \frac{1}{K}\sum_{k=0}^{K-1} J_{\mathrm{d}}[\mathbf{h}(k,m)] + \frac{1}{K}\sum_{k=0}^{K-1} J_{\mathrm{r}}[\mathbf{h}(k,m)] \\ &= J_{\mathrm{d}}[\mathbf{h}(m)] + J_{\mathrm{r}}[\mathbf{h}(m)]. \end{aligned} \tag{54}$$

It is clear that minimization of the narrow-band MSE for each index k is equivalent to minimization of the full-band MSE.

4.1. Wiener

We derive the Wiener filter by minimizing the narrow-band MSE, $J[\mathbf{h}(k,m)]$:

$$\mathbf{h}_{\mathrm{W}}(k,m) = \Phi_{\mathbf{y}}^{-1}(k,m-\tau)\left[\phi_X(k,m)\boldsymbol{\gamma}_X^*(k,m,\tau) + \phi_V(k,m)\boldsymbol{\gamma}_V^*(k,m,\tau)\right], \tag{55}$$

where

$$\Phi_{\mathbf{y}}(k,m-\tau) = \Phi_{\mathbf{x}}(k,m-\tau) + \Phi_{\mathbf{v}}(k,m-\tau). \tag{56}$$

When the clean speech signal is no longer correlated after delay τ, i.e. $\boldsymbol{\gamma}_X(k,m,\tau) = 0$, the Wiener filter reduces to

$$\mathbf{h}_{\mathrm{W}}(k,m)\Big|_{\boldsymbol{\gamma}_X(k,m,\tau)=0} = \phi_V(k,m)\Phi_{\mathbf{y}}^{-1}(k,m-\tau)\boldsymbol{\gamma}_V^*(k,m,\tau). \tag{57}$$

If we define another error signal:

$$E_V(k,m) = V(k,m) - \mathbf{h}_V^H(k,m)\mathbf{y}(k,m-\tau), \tag{58}$$

where $\mathbf{h}_V(k,m)$ is another filter of length L, we can clearly see that this filter estimates the noise. The appropriate MSE is

$$\begin{aligned} J[\mathbf{h}_V(k,m)] &= E\left[\left|E_V(k,m)\right|^2\right] \\ &= E\left[\left|V(k,m)\right|^2\right] + \mathbf{h}_V^H(k,m)E\left[\mathbf{y}(k,m-\tau)\mathbf{y}^H(k,m-\tau)\right]\mathbf{h}_V^H(k,m) \\ &\quad - \mathbf{h}_V^H(k,m)E[\mathbf{y}(k,m-\tau)V^*(k,m)] - E\left[V(k,m)\mathbf{y}^H(k,m-\tau)\right]\mathbf{h}_V(k,m) \\ &= \phi_V(k,m) + \mathbf{h}_V^H(k,m)\Phi_{\mathbf{y}}(k,m-\tau)\mathbf{h}_V(k,m) \\ &\quad - \phi_V(k,m)\mathbf{h}_V^H(k,m)\boldsymbol{\gamma}_V^*(k,m,\tau) - \phi_V(k,m)\boldsymbol{\gamma}_V^T(k,m,\tau)\mathbf{h}_V(k,m) \end{aligned} \tag{59}$$

Minimizing $J[\mathbf{h}_V(k,m)]$, we get the Wiener filter *for the noise estimate*:

$$\mathbf{h}_{V,\mathrm{W}}(k,m) = \phi_V(k,m)\Phi_{\mathbf{y}}^{-1}(k,m-\tau)\boldsymbol{\gamma}_V^*(k,m,\tau), \tag{60}$$

which is the same filter we obtained previously in (57) when the clean speech is not correlated. This indicates that for the Wiener filter in (55), $\mathbf{h}_W^H(k,m)\mathbf{y}(k,m-\tau)$ can be a good estimate for the noise, and as a result $E_W(k,m) = Y(k,m) - \mathbf{h}_W^H(k,m)\mathbf{y}(k,m-\tau)$ can be a good estimate for the desired speech signal, though clearly if $\boldsymbol{\gamma}_X(k,m,\tau) \neq 0$ we can expect to get some signal distortion.

The interesting thing about this method is that we can solve the Wiener filter in (55) adaptively for each frequency bin. Applying the method of gradient descent (Haykin, 2014), which is an iterative adjustment applied to the filter in the direction opposite to the gradient of the cost function, the narrow-band MSE, we can write

$$\mathbf{h}(k,m+1) = \mathbf{h}(k,m) - \frac{\tilde{\mu}(k)}{2}\frac{\partial J[\mathbf{h}(k,m)]}{\partial \mathbf{h}}\Big|_{\mathbf{h}=\mathbf{h}(k,m)}, \tag{61}$$

where $\tilde{\mu}(k)$ is a fixed positive step size parameter per frequency bin, and the scaling factor of $\frac{1}{2}$ has been introduced merely for mathematical convenience. For the Least Mean Square (LMS) (Haykin, 2014) we ignore the expectation operator of the cost function, and so we can formulate the updating rule as

$$\mathbf{h}(k,m+1) = \mathbf{h}(k,m) + \tilde{\mu}(k)\mathbf{y}(k,m-\tau)E^*(k,m), \tag{62}$$

where $0 < \tilde{\mu}(k) < \frac{2}{\lambda_{\max}}$, and $\lambda_{\max}$ is the greatest eigenvalue of the correlation matrix $\Phi_{\mathbf{y}}(k,m-\tau)$. The LMS is suboptimal compared to the Wiener filter, as the learning curve of the algorithm will asymptotically exceed the minimal MSE by an amount termed the excess MSE. However, unlike the Wiener filter, the algorithm does not require knowledge of statistical characteristics, and it does not require inverting the correlation matrix, making it simple. This made the LMS extremely popular and many variants to the algorithm have been developed, namely the simple NLMS (Haykin, 2014), which normalizes the step parameter with the power of the input, making it dimensionless and easier to choose:

$$\mathbf{h}(k,m+1) = \mathbf{h}(k,m) + \frac{\mu(k)}{\mathbf{y}^H(k,m-\tau)\mathbf{y}(k,m-\tau) + \delta}\mathbf{y}(k,m-\tau)E^*(k,m), \tag{63}$$

where $0 < \mu(k) < 2$ is a step-size parameter which should be smaller than 1 here for the algorithm to converge, and $\delta > 0$ is the regularization parameter, which can be quite large depending on the amount of noise. In this standard form, the filter is updated across all frames and frequencies. For stationary harmonic noise we can expect to get some estimation inaccuracy in segments where speech is present as there could remain some speech correlation, even when the noise is absent. For non-stationary harmonic noise we can get additional estimation inaccuracy also in segments containing the edges of the transients, given the convergence speed of the algorithm. These inaccuracies lead to speech distortion.

4.2. Proposed combined approach

The conventional ALE is a forward linear predictor typically implemented by an adaptive NLMS algorithm, where a filter of length L is used and is updated across all frames, i.e.,

$$E(k,m) = Y(k,m) - \sum_{l=0}^{L-1} H_l(k,m)Y(k,m-l-\tau) = Y(k,m) - \mathbf{h}^H(k,m)\mathbf{y}(k,m-\tau) \tag{64}$$

and the filter $\mathbf{h}(k,m)$ is found using (63). We propose using a combination of both a forward linear predictor (FLP) and a backward linear predictor (BLP), to get the combined linear predictor (CLP). The BLP is a non-causal predictor; however, we deem the use of it acceptable as a few frames delay is reasonable in most applications, while we get the added advantage of using future information for better noise estimation. The combination of the FLP and the BLP is done by applying on each frame the filter (either FLP or BLP) that provides the smallest spectral error.

In addition, we propose estimating the noise (Cohen, 2003) and updating the filter only when the nonstationary harmonic noise is present by using a noise presence detector (Cohen and Berdugo, 2001). The detector can be developed, for example, by using a DL algorithm on a database containing medical equipment beeping sounds, similar to the implementation in (Ariav et al., 2018; Zhang and Wu, 2013; Ivry et al., 2019). These algorithms perform in a frame-by-frame manner without introducing temporal delay. As we anyway introduce a few frames of delay for the calculation of the backward filter, it would be possible to take advantage of this oracle information for the detector implementation as well. It is possible to develop the detector with

traditional methods as well, such as with the periodicity measure (Tucker, 1992), with the advantage of not having to collect a large database or the need to train on it and ensure the resulting model can be generalized (this is actually one of the advantages to the proposed ALE compared to a DL enhancer). The development of such a detector is outside the scope of this paper, and we will assume an ideal detector, which assumes knowledge of the noise signal. Let $\mathscr{H}_0$ and $\mathscr{H}_1$ be two hypotheses denoting noise presence and absence, respectively, and let $\mathscr{I}(k,m)$ be a noise indicator, given by

$$\mathscr{I}(k,m)=\begin{cases}1, & V(k,m)\in\mathscr{H}_0\\ 0, & V(k,m)\in\mathscr{H}_1\end{cases}. \tag{65}$$

Another option we propose for the noise indicator would be to use the MI approach calculation for step size $\mu(k)$ to identify the frequencies that contain harmonic noise (Taghia and Martin, 2016). This frequency only indicator requires some temporal delay, as the MI is calculated for a block of frames.

For filter length $L>1$, when the noise transients are only starting, not enough samples of noise are present for the entire filter length. Instead of taking the samples where noise is not present to zero (pre-windowing), we use a set of filters with changing length, based on the available amount of noise samples, until the maximal filter length of L. We denote the forward predictor per filter length as

$$\begin{aligned}E_f^{\ell}(k,m) &= Y(k,m)-\mathscr{I}(k,m)\cdot\sum_{i=0}^{\ell}H_{\ell,i}^{*}(k,m)Y(k,m-i-\tau)\\ &= Y(k,m)-\mathscr{I}(k,m)\cdot\mathbf{h}_{\ell}^{H}(k,m)\mathbf{y}_{\ell}(k,m-\tau),\end{aligned} \tag{66}$$

where

$$\begin{aligned}&\mathbf{h}_{\ell}(k,m)=\left[H_{\ell,0}(k,m),\ldots,H_{\ell,\ell}(k,m)\right]^{T},\\ &\mathbf{y}_{\ell}(k,m-\tau)=\left[Y(k,m-\tau),Y(k,m-\tau-1),\ldots,Y(k,m-\tau-\ell)\right]^{T},\end{aligned} \tag{67}$$

and

$$\ell=\left\{\ell^{'}\,|\sum_{i=0}^{\ell^{'}}\mathscr{I}(k,m-i-\tau)=\ell^{'}+1,\quad \ell^{'}=0,1,\ldots,L-1\right\}. \tag{68}$$

The forward predictor, the FMLNLMS (forward-mapped-L-NLMS), is then

$$E_f(k,m)=E_f^{\max\ell}(k,m). \tag{69}$$

Similarly, we define the backward predictor per filter length as

$$\begin{aligned}E_b^{p}(k,m) &= Y(k,m-L)-\mathscr{I}(k,m-L)\cdot\sum_{i=0}^{p}G_{p,i}^{*}(k,m)Y(k,m-i+\tau)\\ &= Y(k,m-L)-\mathscr{I}(k,m-L)\cdot\mathbf{g}_{p}^{H}(k,m)\mathbf{y}_{p}(k,m+\tau),\end{aligned} \tag{70}$$

where

$$\begin{aligned}&\mathbf{g}_{p}(k,m)=\left[G_{p,0}(k,m),\ldots,G_{p,p}(k,m)\right]^{T},\\ &\mathbf{y}_{p}(k,m+\tau)=\left[Y(k,m+\tau),Y(k,m+\tau-1),\ldots,Y(k,m+\tau-p)\right]^{T},\end{aligned} \tag{71}$$

$$p=\left\{\ell^{'}\,|\sum_{i=0}^{\ell^{'}}\mathscr{I}(k,m-i+\tau)=\ell^{'}+1,\quad \ell^{'}=0,1,\ldots,L-1\right\}, \tag{72}$$

and the backward predictor, the BMLNLMS (backward-mapped-L-NLMS), is then

$$E_b(k,m)=E_b^{\max p}(k,m). \tag{73}$$

Finally, the signal estimate for the combined linear predictor, the CMLNLMS (combined-mapped-L-NLMS), is the filter which provides the smallest spectral error per frame:

$$E_c(k,m)=\begin{cases}E_b(k,m+L), & \left|E_b(k,m+L)\right|^2\le\left|E_f(k,m)\right|^2\text{and}\left|E_b(k,m+L)\right|^2\le\left|Y(k,m)\right|^2\\ E_f(k,m), & \left|E_b(k,m+L)\right|^2>\left|E_f(k,m)\right|^2\text{and}\left|E_f(k,m)\right|^2\le\left|Y(k,m)\right|^2\\ Y(k,m), & \text{else}\end{cases}. \tag{74}$$

The FMLNLMS reduces the noise transient after a delay determined by the step size and the de-correlation delay parameters, and hence residuals of the noise would remain at the beginning of the noise transient. For the BMLNLMS, accordingly, residuals would remain at the end of the transient. By using the combined CMLNLMS filter, we are able to increase the reduction span of the noise transient, thus reducing the amount of noise residuals, as will be demonstrated in the experimental results section.

As we also apply the filters based on a noise presence indicator (Mousazadeh and Cohen, 2013), we do not process the speech signal where it is not necessary and so we do not incur the inherent estimation error which is dependent on the speech correlation. Thus we are able to preserve the speech components where noise is not present.

Note that if the spectral error of either the FMLNLMS or BMLNLMS is larger than the noisy signal spectra, we do not use either of the filters' results, rather we use the noisy signal itself. This is done to avoid over-estimation of the noise.

5. Experimental results

The evaluation is done on speech signals taken from the TIMIT database (Garofolo, 1993), including 20 different utterances, each utterance is from a different speaker, half of the utterances are from male speakers and half from female speakers. The signals are sampled at 16 kHz and degraded by nonstationary harmonic noise with overall SNR in the range $[0, 20]$ dB. The noisy signals are transformed to the time-frequency domain using STFT, which is implemented with overlapping Hamming analysis. The overlap-add method is used for the signal reconstruction in the time domain. For the noise, different nonstationary harmonic noise signals such as heart monitor beeping, train door beeping, house alarm, smoke alarm, and rail road crossing bells were collected to form a database of 26 different signals (BBC Sound Effects, http://bbcsfx.acropolis.org.uk ; Freesound, https://freesound.org ; SoundBible, https://soundbible.com ; YouTube, https://www.youtube.com). The database contains both real-life recorded signals and synthesized signals generated for sound effects. The signals were converted from stereo to mono and down-sampled to 16 kHz for our use.

To implement the ideal noise detector, we define a threshold value relative to the maximum spectrum of the nonstationary harmonic noise. If the noise spectrum is above this threshold, we consider the noise to be present; if it is equal or below it, we consider the noise to be absent:

$$\mathcal{I}(k,m) = \begin{cases} 1, & \left[20\log_{10}\left|V(k,m)\right| - \max\left\{20\log_{10}\left|V(k,m)\right|\right\}\right] > I_{\text{threshold}} \\ 0, & \text{else} \end{cases} \tag{75}$$

This ideal detector assumes knowledge of the noise signal. By using different $I_{\text{threshold}}$ values and the frequency only detector based on the MI approach, we can emulate the impact of a non-ideal noise detector, which would not assume such knowledge.

We evaluate the performance of our proposed approach using the distortion index and noise reduction factor, where we investigate the measures as defined in Section 3, which we designate the orthogonal decomposition (OD) measures where the residual interference is treated as part of the noise. We compare the OD measures to the traditional performance measures where the residual interference is part of the filtered desired signal (Benesty et al., 2009). In addition, we use the perceptual evaluation of speech quality (PESQ) measure (ITU-T, 2007) and the short-time objective intelligibility (STOI) measure (Taal et al., 2011), where larger values of PESQ and STOI indicate better quality and intelligibility, respectively.

5.1. Correlation vector

To evaluate the correlation vector $\boldsymbol{\gamma}_X(k,m,\tau)$, we define the vector:

$$\mathbf{x}_1(k,m,\tau) = [X(k,m), X(k,m-1), \ldots, X(k,m-\tau), \ldots, X(k,m-\tau-L+1)]^T \in \mathbb{C}^{(L+\tau)\times 1} \tag{76}$$

and its correlation matrix:

$$\boldsymbol{\Phi}_{\mathbf{x}_1}(k,m,\tau)_{(L+\tau)\times(L+\tau)} = E\left[\mathbf{x}_1(k,m,\tau)\mathbf{x}_1^H(k,m,\tau)\right]. \tag{77}$$

The correlation vector of speech is then the transpose of the first row vector of $\boldsymbol{\Phi}_{\mathbf{x}_1}(k,m,\tau)$ from location τ, normalized by its first element:

$$\boldsymbol{\gamma}_X(k,m,\tau) = \frac{E[X(k,m)\mathbf{x}^*(k,m-\tau)]}{E\left[\left|X(k,m)\right|^2\right]} = \tag{78}$$

$$= \frac{\boldsymbol{\Phi}_{\mathbf{x}_1}(k,m,\tau)(1,\tau:(\tau+L))}{\boldsymbol{\Phi}_{\mathbf{x}_1}(k,m,\tau)(1,1)} \tag{79}$$

and the delayed correlation matrix is the matrix block from row and column $\tau+1$:

$$\boldsymbol{\Phi}_{\mathbf{x}}(k,m-\tau) = \boldsymbol{\Phi}_{\mathbf{x}_1}(k,m,\tau)(\tau+1:\tau+L,\tau+1:\tau+L). \tag{80}$$

We evaluate in a similar fashion the noise correlation vector $\boldsymbol{\gamma}_V(k,m,\tau)$ and the delayed correlation matrix $\boldsymbol{\Phi}_{\mathbf{v}}(k,m-\tau)$.

We examine the correlation vectors behavior in Fig. 2. We see that the absolute value of the correlation vector of the speech signal is reduced compared to that of the noise and it decays faster per delay τ. Hence, $Z(k,m)$ would be a better estimate of the noise than of the speech signal, though we can expect distortion as the delayed speech signal correlation is clearly not zero. This demonstrates the existing tradeoff between the noise reduction and distortion. The noise reduction will be better for smaller delay τ as the noise correlation is higher, but the speech signal distortion will be higher as also the speech correlation is higher.

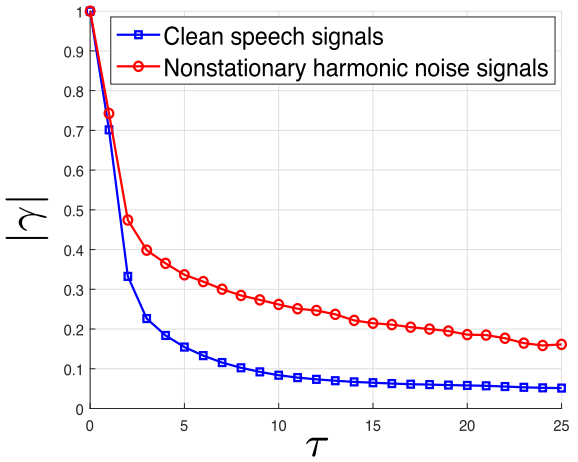

Fig. 2. Absolute value of the correlation vectors vs. delay parameter τ for all database clean speech signals (blue square) and for all database nonstationary harmonic noise signals (red circle), for filter length $L = 1$, window length $K = 512$ and 75% overlap. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

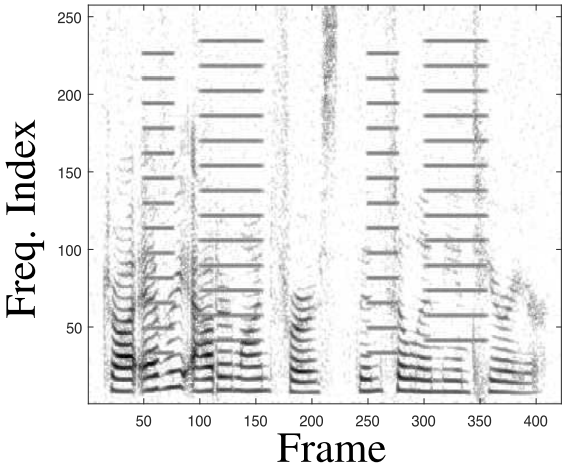

Fig. 3. Spectrogram of a 3.4 s female speech signal corrupted by a synthetic nonstationary harmonic noise at 10 dB SNR.

For smaller window lengths or for larger percentage overlaps, the correlations can be expected to decay more slowly. We will continue the analysis with overlap of 75% to retain high correlation, and use window length $K = 512$.

5.2. Least squares

For illustrative purposes, we start the analysis with the least-squares (LS) filter (Haykin, 2014; Manolakis et al., 2005) and a synthetic nonstationary harmonic noise corrupting a 3.4 s long female speech signal from the database. The noisy speech signal is shown in Fig. 3. The noise is generated by adding harmonics in different frequencies modulated by trapezoidal trains in the time domain to white Gaussian noise (WGN). We use a threshold value of -25 dB for the noise detector, and compare between the OD performance measures in Section 3, and the traditional performance measures.

Observing Fig. 4, which shows the resulting performance measures for the different filters, we see that, as expected, the speech signal is still correlated to its delayed version. As we increase delay τ, the correlation and hence the signal distortion decrease, and as the noise signal also becomes less correlated as τ increases, the noise reduction also decreases. These trends are true for all filter types: the FLP, BLP, and the proposed CLP. The residual interference is considered part of the distortion for the traditional definition compared to the OD definition. Accordingly, we see that the traditional distortion index decreases as τ increases to a higher level than the OD distortion index. While on the other hand, the noise reduction for the OD definition is lower than the noise reduction for the traditional definition, as we compare the original noise level to the residual noise plus residual interference. In fact, in Fig. 5, which demonstrates the distortion index and noise reduction factor for the CLP filter at

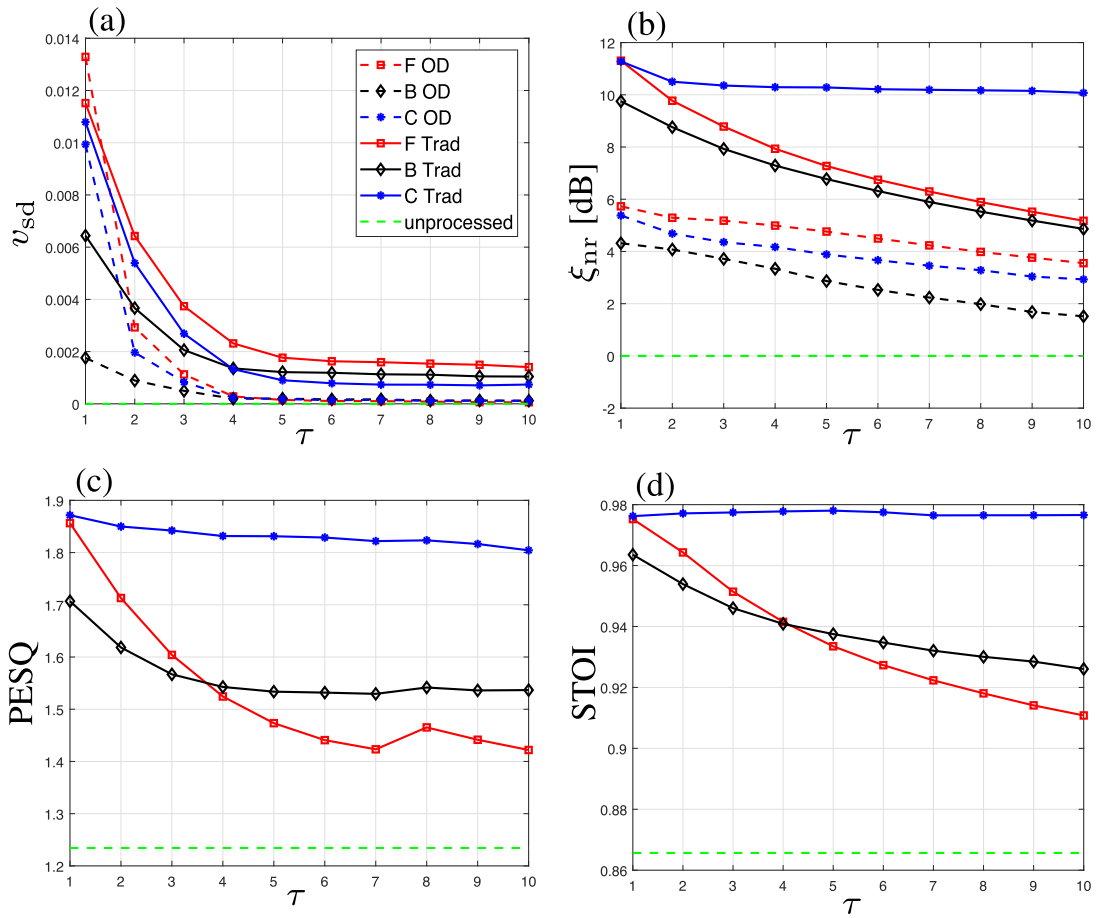

Fig. 4. (a) Distortion index, (b) noise reduction factor, (c) PESQ, and (d) STOI for LS ALE filter applied on the 3.4 s female speech signal corrupted by synthetic non-stationary harmonic noise at 10 dB SNR for varying delay τ and filter length $L = 3$, for the different LS filters: FLP (red square), BLP (black diamond), and CLP (blue asterisk). Solid lines are for traditional performance measures definition, the dashed lines are for the OD definition in Section 3. Dashed green is for the unprocessed noisy signal. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

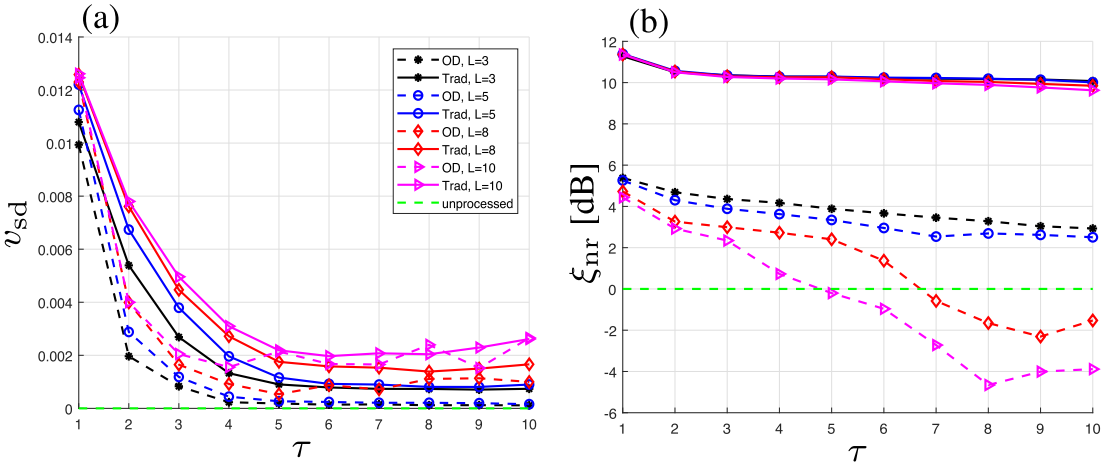

Fig. 5. (a) Distortion index and (b) noise reduction factor for CLP LS ALE filter applied on the 3.4 s female speech signal corrupted by synthetic nonstationary harmonic noise at 10 dB SNR for varying delay τ and different filter lengths $L \in \{3, 5, 8, 10\}$. Solid lines are for traditional performance measures definition, the dashed lines are for the OD definition in Section 3. Dashed green is for the unprocessed noisy signal. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

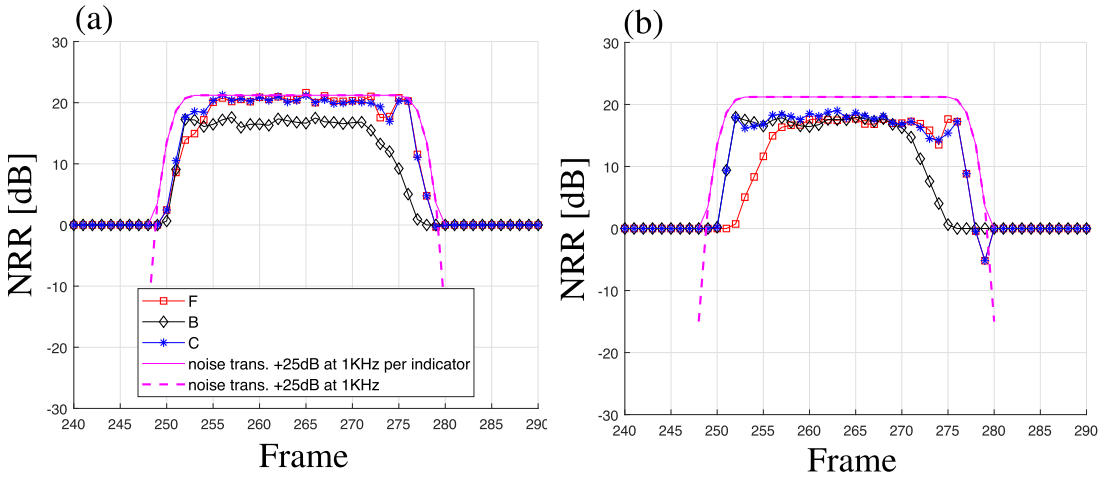

Fig. 6. Average noise reduction ratio (NRR) per frame for the LS filters applied on a 3.4 s female speech signal corrupted by synthetic nonstationary harmonic noise at 10 dB SNR, filter length $L = 3$ and with delay (a) $\tau = 1$ and (b) $\tau = 3$, where red square is for the forward filter, black diamond is for the backward filter, blue asterisk is for the combined filter, solid magenta is the spectral noise level at 1 KHz given the specified indicator shifted by 25 dB, and dashed magenta is the spectral noise level at 1 KHz shifted by 25 dB. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

different filter lengths, we see that the OD noise reduction factor can go below 0 dB. This means that from the OD perspective we are enhancing the noise and not suppressing it. This does not reflect in the perception of quality of the enhanced signal seen in the PESQ level, or in the intelligibility seen in the STOI level, where we have improved levels compared to the unprocessed noisy signal. The results for PESQ and STOI at the different filter lengths are very similar, so only the result for a single filter length is shown in Fig. 4. The general trend for PESQ and STOI is also a decrease in value as we increase the delay; hence the noise reduction decrease is more dominant than the distortion decrease.

We observe the dependence on filter length in Fig. 5. For the OD distortion index, longer filter length means we are deducting more correlated components from the current speech signal, so the distortion increases as the filter length increases. For the traditional distortion index, both the amount of correlated speech being deducted from the speech and the amount of residual interference increase as we increase the filter length; therefore the distortion index also increases. For the OD noise reduction factor, for longer filter lengths we get more residual interference which is more dominant than the improvement in the noise modeling, so we get lower noise reduction. For the CLP filter, the traditional noise reduction dependence on the delay τ becomes smaller. This will be explained presently. In Fig. 4 we see that for the traditional definition the CLP has better noise reduction than the FLP and the BLP. For the OD definitions though, the CLP has lower noise reduction. Again, this does not reflect the PESQ and STOI levels which are better for the CLP.

Fig. 6 illustrates the reduction of the noise transients (noise only) for each filter, given the delay parameter and filter length. As expected, the forward and backward filters do not reduce the entire transient. The FLP reduces the transient starting after delay τ until the end of the transient, while the BLP starts from the beginning until $\tau + 1$ from the end of the transient. Using the combination of both FLP and BLP, the CLP reduction of the transient spans the entire length of the transient. At the edges of the transient, the noise level is low; however the LS filter uses the same average filter coefficients to estimate it, so it is possible to overestimate the noise. One way to mitigate this, as was implemented here, is to use the filtered results only if the spectral error is smaller than the noisy signal spectra. Another case where the noise can be overestimated is when the transient edge overlaps speech components, and the noise is estimated erroneously based on the higher-leveled speech. This becomes more pronounced with longer delay, as can be seen in Fig. 6. In this we see the clear impact of the noise detector. For a higher threshold used, we will get less overestimation of the level of the noise; however, as we can expect, we will get also less noise reduction since we are taking into account less noise to be reduced. For the CLP filter, since we span the entire transient regardless of τ, the traditional noise reduction dependence on the delay τ becomes smaller, as we saw earlier. Some dependence remains, as the correlation of the noise does reduce as τ increases, and therefore we see that for larger delay τ we still have less noise reduction.

Given the results for the LS filters using the OD definitions, we continue the analysis with the traditional performance measures only.

5.3. Adaptive filtering

All other experiments are done with noise signals from the collected database. We use the adaptive NLMS filter, where we compare between the forward filter (FMLNLMS) in (69), the backward filter (BMLNLMS) in (73), and the proposed combined filter (CMLNLMS) in (74). In addition, we compare the performance of our proposed approach to the conventional forward NLMS with fixed step size, as well as to the MI approach in (Taghia and Martin, 2016). We also propose joining the MI approach and the

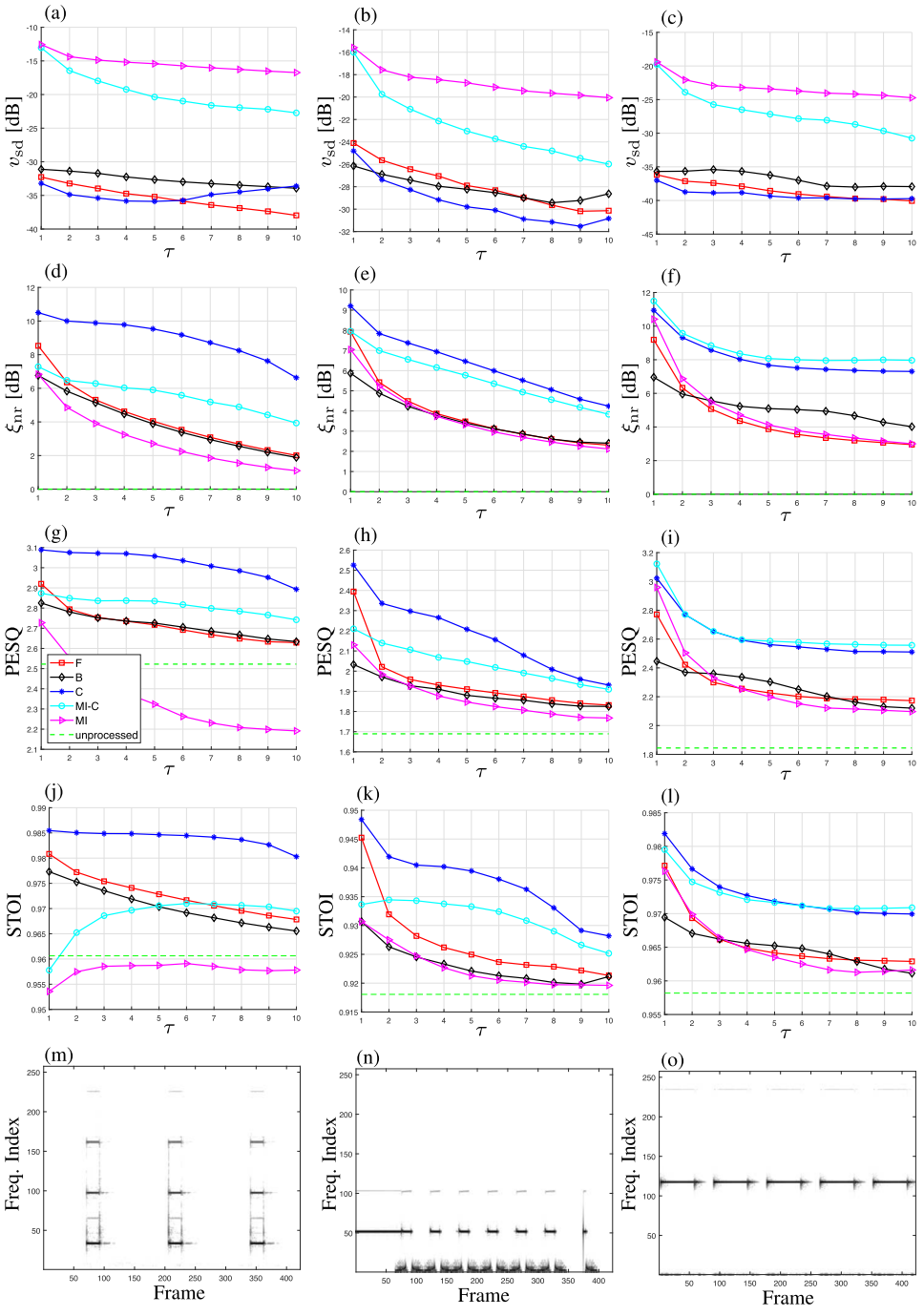

Fig. 7. Average (a-c) distortion index, (d-f) noise reduction factor, (g-h) PESQ, and (j-l) STOI, for the adaptive filters applied on ten different speech signals each from a different speaker, half female half male, corrupted by 3 different types of nonstationary harmonic noise whose spectrograms are: (m) hospital beeping 1, (n) hospital beeping 2, and (o) house fire alarm, at 10 dB SNR for varying delay τ and filter length $L = 3$, where red square is for the forward filter, black diamond is for the backward filter, blue asterisk is for the proposed combined filter, cyan circle is for the combined filter using MI, and magenta triangle is for the standard MI. Dashed green is the result for the unprocessed noisy signal.(a,d,g,j) are the average results for noise (m), (b,e,h,k) are the average results for noise (n) and (c,f, i,l) are respectively the average results for noise (o). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

combined filter by using the MI approach calculation for step size $\mu(k)$ and the identification of the frames that contain noise, i.e. use the MI as the noise indicator with the calculated step size. We compare this to the previously mentioned methods as well.

We should note that for the MI calculation in (Taghia and Martin, 2016) a decision block was included to determine whether the clean signal is corrupted by harmonic noise, represented by coefficient Q_μ which would get 1 if it is corrupted by harmonic noise and 0 otherwise. This coefficient multiplies the step size, so if $Q_\mu = 0$, the step size is zero and the noisy signal does not get processed. It is set according to a threshold level, which was decided upon based on a comparison of random noise types such as babble and WGN to harmonic noises such as vehicle engine and traffic. For the highly nonstationary harmonic noises that we evaluate, the resulting coefficient Q_μ is typically zero. Hence, we disregard this coefficient in our analysis and always process the noisy signal.

In Fig. 7 we show the average performance measures for the different adaptive filters applied on different speech signals from different speakers corrupted by three different types of real nonstationary harmonic signals, specifically two different hospital beeping noises and a house fire alarm. Similar to the LS, we see that using the combined filter we are improving on all the measures compared to the forward or backward filters standalone, except for additional distortion in some cases. We clearly see that the proposed combined filter outperforms the MI approach. As expected, the proposed combined approach achieves less distortion than the MI approach, since it employs an indicator on frames as well as frequencies. It is worth noting that also for the MI-combined approach which implements the MI indicator while using the combination of forward and backward NLMS filters we get an improvement over the standard MI approach. In fact, with the MI-combined we can get better results for some of the measures compared to the combined, when the nonstationary harmonic noise is almost constant per frame. This can also occur for large values of the delay parameter and a long filter, as demonstrated in Fig. 8, where we compare the combined and the MI-combined for varying delay parameter and different filter lengths for hospital beeping 1 noise. For the proposed combined approach, an appropriate selection of step size μ is still required for optimal results. This remains unchanged from the

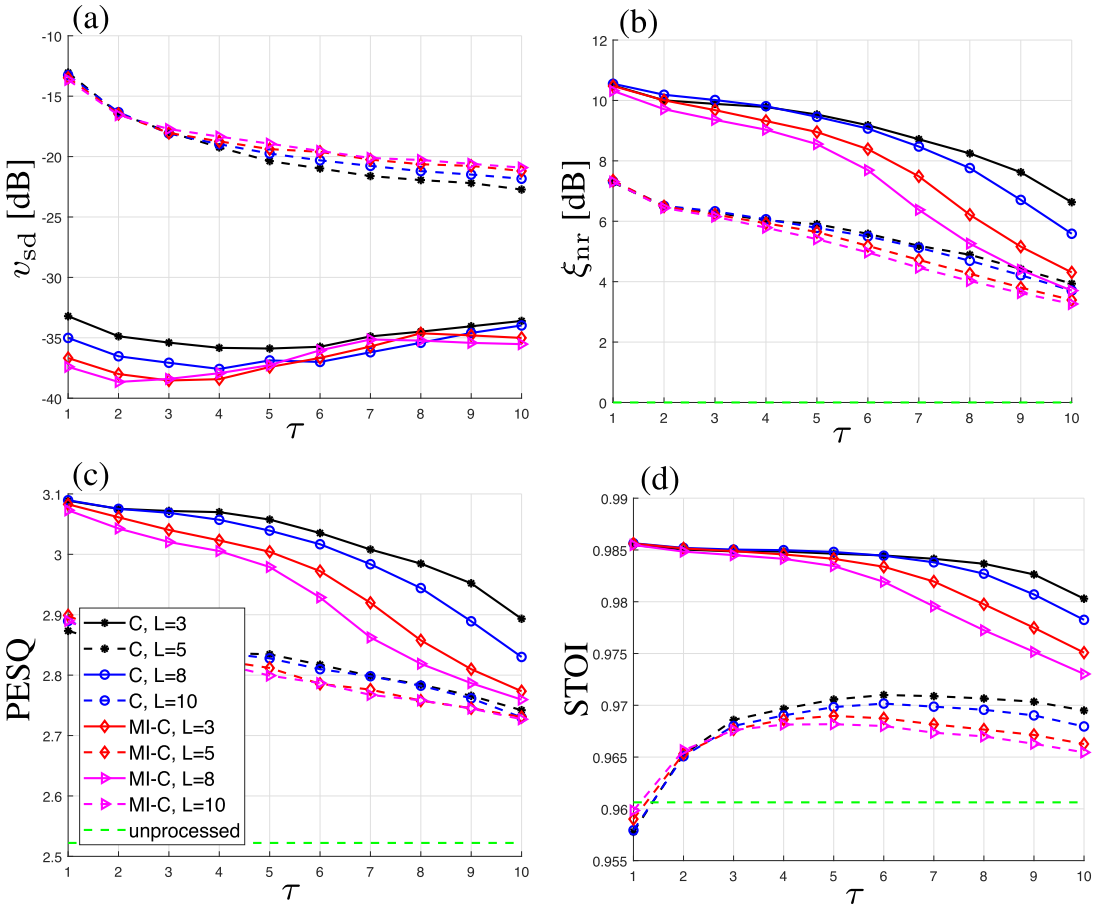

Fig. 8. Average (a) distortion index, (b) noise reduction factor, (c) PESQ, and (d) STOI, for the adaptive filters applied on ten different speech signals each from a different speaker, half female half male, corrupted by hospital beeping 1 noise at 10 dB SNR for varying delay τ and different filter lengths L, where solid lines are for the combined filter and dashed lines are for the MI-combined filter. Dashed green is the result for the unprocessed noisy signal. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

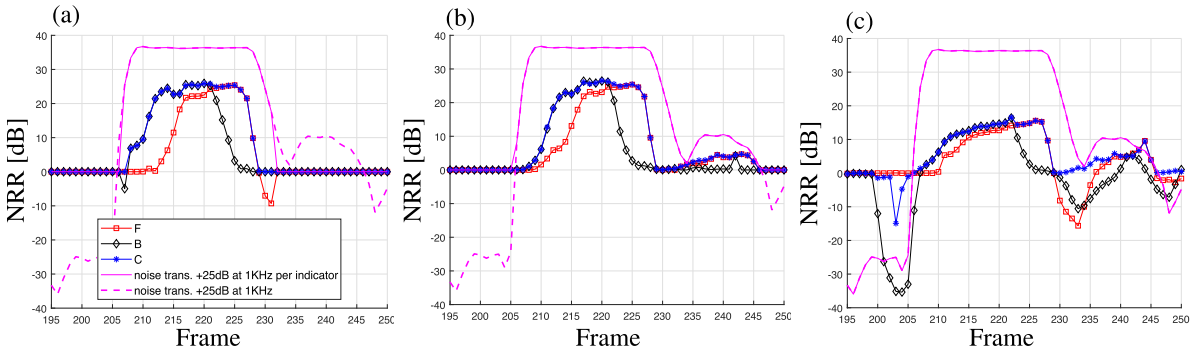

Fig. 9. Average NRR per frame at 1 KHz for adaptive filters applied on a 3.4 s female speech signal corrupted by a hospital beeping 1 noise at 10 dB SNR, filter length $L = 3$ and with delay $\tau = 3$ for (a) MLNLMS with $\mu = 0.5$ and indicator threshold=−25 dB, (b) MLNLMS with $\mu = 0.5$ and indicator threshold=−40 dB, and (c) MI-MLNLMS. Where red square is for the forward filter, black diamond is for the backward filter, blue asterisk is for the combined filter, solid magenta is the spectral noise level given the specified indicator shifted by 25 dB, and dashed magenta is the spectral noise level shifted by 25 dB. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Average results for ten different speech signals each from a different speaker, half female half male, degraded by three different types of nonstationary harmonic noise at 10 dB SNR: hospital beeping 1, hospital beeping 2, and house fire alarm, with delay $\tau = 1$ and filter length $L = 3$. The performance of the CMLNLMS with different noise indicators is presented, in addition to the unprocessed noisy condition, in terms of quality measure PESQ, intelligibility measure STOI, distortion index v_{sd}, noise reduction factor ξ_{nr}, and overall oSNR.

Method	PESQ	STOI	v_{sd}	ξ_{nr} [dB]	oSNR [dB]
Hospital beeping 1					
Unprocessed	2.5222	0.9606	0	0	10
MI-CMLNLMS	2.8733	0.9578	0.0569	7.2855	16.7384
CMLNLMS -25dB	3.0887	**0.9855**	**0.0014**	10.4996	20.5596
CMLNLMS -40dB	**3.118**	0.985	0.0071	**10.5385**	**20.5733**
Hospital beeping 2					
Unprocessed	1.6893	0.9181	0	0	10
MI-CMLNLMS	2.2099	0.9336	0.0312	7.9484	17.5639
CMLNLMS -25dB	**2.5256**	**0.9484**	**0.0042**	9.2039	19.2213
CMLNLMS -40dB	2.4382	0.9174	0.0205	**9.3499**	**19.2804**
House fire alarm					
Unprocessed	1.8442	0.9582	0	0	10
MI-CMLNLMS	**3.1218**	0.9795	0.0126	**11.4893**	**21.2375**
CMLNLMS -25dB	3.0227	**0.9819**	**0.0008**	10.939	20.9498
CMLNLMS -40dB	3.0199	0.9819	0.0013	11.0794	21.0885

conventional NLMS. Larger values of the step size lead to faster convergence which results in better noise reduction; however, we also get more distortion. From $\mu = 0.5$ and larger there is no significant improvement in the noise reduction, but as mentioned the distortion increases, hence, we continue the analysis with $\mu = 0.5$. Another approach not investigated here, would be to use the maximum between the MI calculated step to a fixed value such as the selected 0.5. This is expected to improve the results of the combined compared to the MI-combined for cases such as the house fire alarm.

Fig. 9 illustrates the adaptive filters behavior for the noise transients reduction. We get similar behavior to the LS filter, where the combined filter has increased reduction span of the transient compared to the forward or backward filters standalone. We see a similar improvement using the combined filter with the MI approach compared to the standard MI approach which uses a forward NLMS (shown by the red square curve). Fig. 9 also demonstrates the reduction of the transient given different indicators: an indicator with threshold value of −25 dB, an indicator with threshold value of −40 dB, and an indicator based on the MI, as used in the MI-CMLNLMS described above. Table 1 shows that the range of results for the CMLNLMS between an almost ideal detector (−40 dB threshold) and a degenerate detector that only detects harmonic frequencies (MI) is not large, when the optimal results are not necessarily achieved for the better detector. Therefore, a reasonable amount of misdetection or false-detection in an implemented indicator is not expected to have a large impact.

In (69) and (73) we calculate the filters for a set of filters, from length 1 to length L per the noise presence and use the maximal filter length available. The advantage of the filter length set compared to a single fixed length for each standalone forward or backward filter is clear. We can start filtering when the samples where noise is present are few. For the combined filter, the main

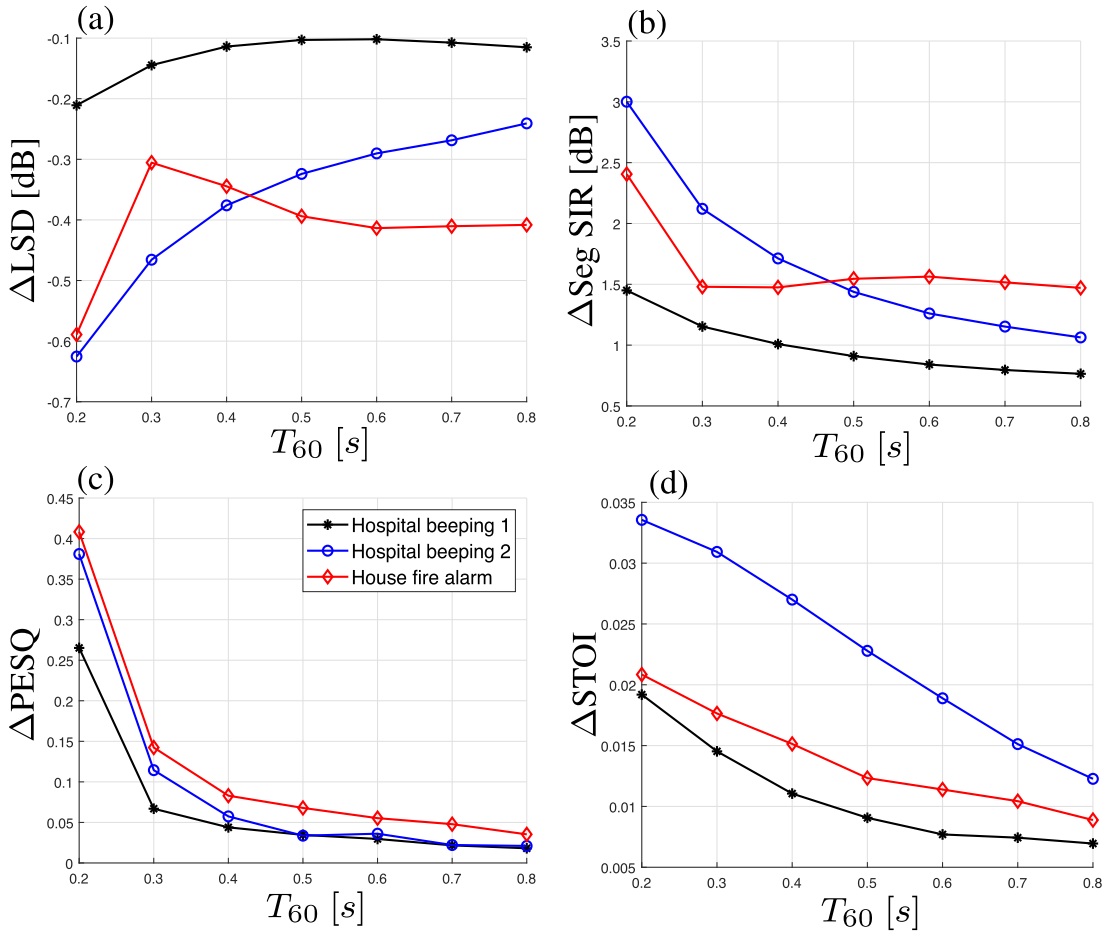

Fig. 10. Average (a) ΔLSD, (b) ΔSegSIR, (c) ΔPESQ, and (d) ΔSTOI, for the proposed CMLNLMS adaptive filter with delay $\tau = 1$, filter length $L = 3$, step size $\mu = 0.5$, and indicator threshold=−25 dB applied on ten different speech signal each from a different speaker, half female half male, corrupted by hospital beeping 1 (black asterisk), hospital beeping 2 (blue circle) and house fire alarm (red diamond) noises at 10 dB SNR, for reverberant room size $5 \times 6 \times 4$ m, with fixed 1 m distance between the speech source and the microphone and fixed 2 m distance between the noise source and the microphone, for varying T_{60}. Δ is calculated between the processed signal results to the unprocessed reverberant noisy signal results. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

advantage is for the scenario when there is no overlap between the forward and the backward filters, which can happen when the transient is short compared to the delay. From Fig. 8 we see that at larger delays shorter filters have better performance. This can be expected, as at large delay the correlation of both the noise and the signal to their delayed version is reduced. Therefore, we recommend keeping the filter length short, to reduce the required computing resources, to reduce the distortion which generally increases as the filter length is longer, and to achieve overall better performance. As we do not want to introduce a large delay into the system, and given the shown results, we recommend to choose τ to be 1 or at least no more than a few frames. For smaller window size, this should scale up, i.e. a larger delay could be used.

Additional background noise that is stationary and random, such as WGN, gets decorrelated by the delay parameter and so is not expected to have an impact on the results. A more interesting case is when the environment is reverberant. The reverberations, which are reflections of the source signal from walls and other obstacles arriving at different delays to the microphone are correlated to the source signal, and hence can impact the ALE performance. We generate the reverberations by convolving the anechoic speech and anechoic nonstationary harmonic noise with synthetic Room Impulse Responses (RIRs). The RIRs are generated by Matlab implementation Habets (2014) of the image method Allen and Berkley (1979). We perform the following test: we vary T_{60} while keeping the distances between the microphone-to-sources fixed. The performance is now evaluated using the segmental signal-to-interference ratio (SegSIR) and the log spectral distance (LSD), see Habtes et al. (2008) for the equations, as well as PESQ and STOI. As a reference for these performance measures we use the (properly delayed) anechoic speech signal. In Fig. 10 we show the performance of the proposed combined CMLNLMS filter, when we simulated a room size of $5 \times 6 \times 4$ m (length×width×height) with 1 m distance between microphone to speech source and 2 m distance between microphone to noise source. The results were averaged across different utterances from different speakers, and are presented as Δ which is the difference between the resulting performance for the processed signal to the results for the unprocessed signal. The results for three different nonstationary harmonic noise types are shown. We clearly observe that even under different reverberant conditions

Table 2
Results for speech database degraded by nonstationary harmonic noise database at 0,10, and 20 dB SNR, with delay $\tau = 1$ and filter length $L = 3$. The performance of the conventional NLMS with fixed step size $\mu = 0.05$, the MI approach, the proposed the joint MI-CMLNLMS, and the proposed CMLNLMS with step size $\mu = 0.5$ are presented, in addition to the unprocessed noisy condition, in terms of quality measure PESQ, intelligibility measure STOI, distortion index v_{sd}, noise reduction factor ξ_{nr}, and overall oSNR.

Method	PESQ	STOI	v_{sd}	ξ_{nr} [dB]	*oSNR* [dB]
iSNR 20 [dB]					
Unprocessed	3.1405	0.986	0	0	20
NLMS-0.05	2.8187	0.961	0.1263	2.6729	21.1417
MI	3.208	0.9778	0.0471	3.4674	22.9966
MI-CMLNLMS	3.231	0.9782	0.0465	3.6651	23.1704
CMLNLMS-0.5	**3.4202**	**0.9885**	**0.024**	**5.6781**	**25.5897**
iSNR 10 [dB]					
Unprocessed	2.3409	0.9458	0	0	10
NLMS-0.05	2.17	0.9307	0.1294	4.606	13.0758
MI	2.5134	0.9551	0.0494	5.6614	15.1939
MI-CMLNLMS	2.5519	0.9568	0.0487	6.0447	15.553
CMLNLMS-0.5	**2.6913**	**0.9712**	**0.0231**	**8.2737**	**18.194**
iSNR 0 [dB]					
Unprocessed	1.662	0.8663	0	0	0
NLMS-0.05	1.6107	0.8669	0.136	6.0182	4.4684
MI	1.8383	0.8988	0.0561	7.1133	6.6275
MI-CMLNLMS	1.8618	0.9019	0.055	7.6637	7.1537
CMLNLMS-0.5	**1.8957**	**0.9185**	**0.0216**	**9.7319**	**9.678**

(varying T_{60}) the proposed filter improves on the performance as ΔPESQ, ΔSTOI, and ΔSegSIR are positive and ΔLSD is negative. It is interesting to note that as the reverberation time is larger we generally get reduction in the performance of the ALE filter, meaning that the relationship between the correlation vectors as shown in Fig. 2 has shifted. Specifically for the house fire alarm from a relatively low T_{60} ΔSegSIR and ΔLSD become almost flat. This noise has very short stops between the repetitions of the harmonic so that even with a small T_{60} the noise and the reverberations of the noise are overlapping, and effectively that harmonic becomes present across all the frames, almost like a stationary harmonic noise, which is easier to reduce. We should note that the proposed method does not improve on the speech reverberations as it focuses on the harmonic noise reduction. It could be followed by another stage to clean up the speech reverberations.

Table 2 shows the mean results for the database of noise signals degrading the database of speech signals at different SNR conditions (20 dB, 10 dB, and 0 dB respectively), for the different methods: the conventional NLMS with fixed step size, the MI approach, the proposed combined filter, and the joint MI-combined filter as well as the unprocessed noisy condition. We can clearly see that with the proposed approach of the CMLNLMS it is possible to achieve the best results for the different performance measures. Second to the CMLNLMS would be the proposed MI-CMLNLMS approach. We should note that with a smaller step size for the conventional NLMS with fixed step size it is possible to decrease the distortion so that it is lower than the combined filter; however, we would get even lower noise reduction and oSNR which are already lower than the combined filter.

To conclude this section, we present in Fig. 11 the resulting spectrograms of the enhanced signals for the MI approach and for the proposed CMLNLMS with $\mu = 0.5$ and -25 dB threshold. By applying the proposed approach, we are able to remove most of the highly nonstationary harmonics with little residuals remaining, while with the MI approach clearly more residuals remain. We should note that the higher frequency harmonic was not reduced in this example as it was below the level of the threshold used for the noise indicator.

6. Conclusions

ALE is commonly implemented with a forward NLMS adaptive filter. Here we proposed using an additional non-causal backward NLMS adaptive filter and taking a combination of the forward and backward filters according to the minimal spectral error to reduce nonstationary harmonic noise. The combination of these two filters enables better reduction of the noise transients, compared to using only the forward filter, which would only start reducing each noise transient after the de-correlation delay. We showed that our proposed approach is effective compared to conventional forward NLMS with fixed step size and the MI approach; it improves the noise reduction and improves the speech quality and the speech intelligibility, while introducing little distortion. In addition, we showed that the improvement in results holds for different SNR conditions.

We used the combined filter with a noise indicator allowing for lower distortion and better noise reduction. Though the development of the indicator was outside the scope of this paper, we showed that by using different indicators, including a degenerate indicator that only detects the harmonic frequencies, we were able to improve on the results. Finally, we explained the behavior of the filter for the different parameters, the step size, the de-correlation delay, and the filter length, and provided recommendations on the selection of these parameters. The proposed ALE system was shown to reduce nonstationary harmonic

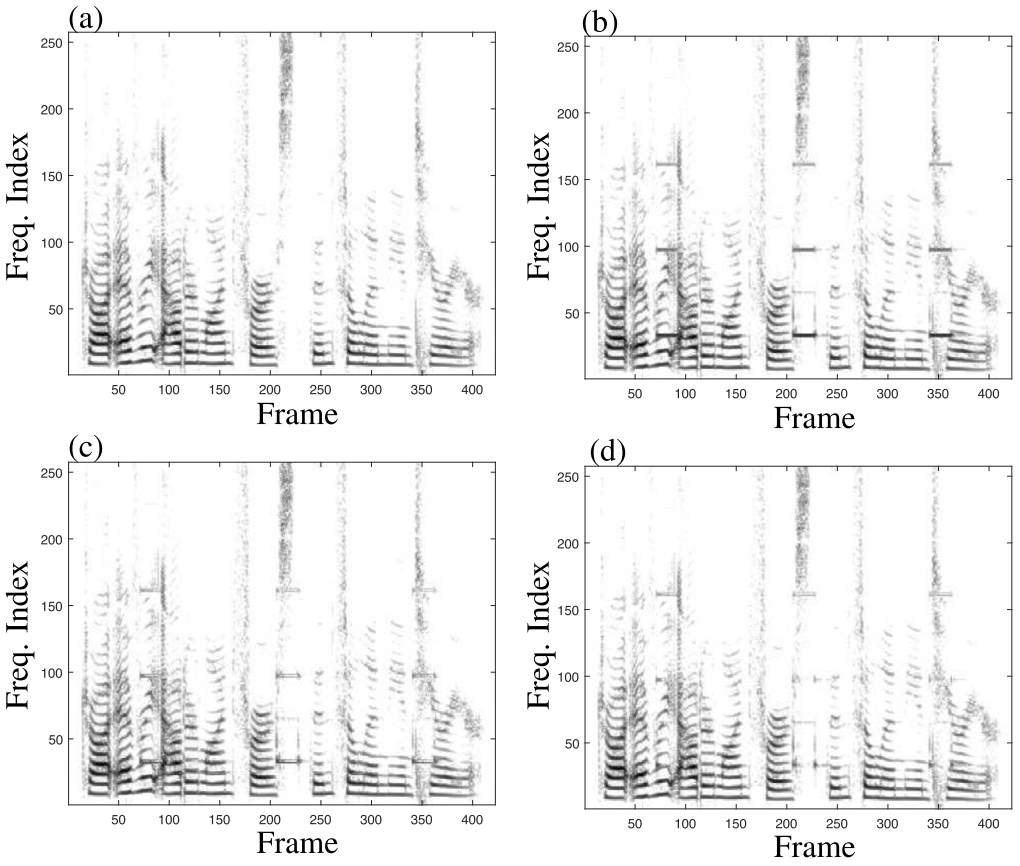

Fig. 11. Spectrograms of (a) 3.4 s female clean speech signal from the database, (b) speech signal corrupted by nonstationary harmonic hospital beeping 1 noise at 10 dB SNR, (c) enhanced signal obtained from the MI approach, and (d) enhanced signal obtained from the proposed CMLNLMS with $\mu = 0.5$. The results are obtained with filter length $L = 3$, de-correlation delay $\tau = 1$, and indicator threshold of -25 dB for the combined filter.

noise. For reduction of the random wide-band components of the noise, the ALE can be used as the first stage of a two stage system, where it is followed by a traditional speech enhancement method.

Future work will check the impact of using a frequency dependent threshold value for the noise detector implementation and include subjective listening tests to further validate the results. Additionally, the performance of the proposed method under reverberant conditions, particularity with low signal-to-noise ratio, should be further explored.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by the Israel Science Foundation (grant No. 576/16), and the ISF-NSFC joint research program (grant No. 2514/17). The authors thank the anonymous reviewers for their constructive comments which helped to improve the presentation of this paper.

References

Allen, J.D., Berkley, D.A., 1979. Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Amer 65 (4), 943–950.
Ariav, I., Dov, D., Cohen, I., 2018. A deep architecture for audio-visual voice activity detection in the presence of transients. Signal Process. 142, 69–74.
Benesty, J., Chen, J., Huang, Y., Cohen, I., 2009. Noise Reduction in Speech Processing. Springer-Verlag, Berlin, Germany.
Benesty, J., Cohen, I., Chen, J., 2018. Fundamentals of Signal Enhancement and Array Signal Processing. Wiley-IEEE Press, Singapore.
Boll, S.F., 1979. Supression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Process. ASSP-27 (2).
Chen, J., Benesty, J., Huang, Y., Doclo, S., 2006. New insights into the noise reduction wiener filter. IEEE/ACM Trans. Audio Speech Lang. Process. 14 (4), 1218–1234.
Cohen, I., 2003. Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. IEEE Trans. Speech Audio Process. 11 (5), 466–475.

Cohen, I., 2005. Speech enhancement using super-gaussian speech models and noncausal a priori SNR estimation. Speech Commun. 47 (3), 336–350.
Cohen, I., Benesty, J., (Eds.), S.G., 2010. Speech Processing in Modern Communication: Challenges and Perspectives. Springer-Verlag, Berlin.
Cohen, I., Berdugo, B., 2001. Spectral enhancement by tracking speech presence probability in subbands. In: Proc. IEEE Workshop on Hands Free Speech Communication, HSC'01, pp. 95–98.
Cohen, I., Gannot, S., 2008. Spectral enhancement methods. In: Benesty, J., Sondhi, M.M., Huang, Y.A. (Eds.), Springer Handbook of Speech Processing. Springer, pp. 873–901.chapter 44
Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum mean square error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. ASSP-32 (6), 1109–1121.
Ephraim, Y., Van Trees, H.L., 1995. A signal subspace approach for speech enhancement. IEEE Trans. Speech Audio Process. 3 (4), 251–266.
Garofolo, J.S., 1993. Getting Started with the DARPA TIMIT CD-ROM: An Acoustic Phonetic Continuous Speech Database. National Institute of Standards and Technology (NIST).
Habets, E. A. P., 2014. Rir generator. https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator.
Habtes, E.A.P., Gannot, S., Cohen, I., 2008. Joint dereverberation and residual echo suppression of speech signals in noisy environments. IEEE Trans. Audio Speech Lang. Process. 16 (8), 1433–1451.
Haykin, S., 2014. Adaptive Filter Theory, fifth ed. International ed. Upper Saddle River: Pearson.
Hu, Y., Loizou, P.C., 2003. A generalized subspace approach for enhancing speech corrupted by colored noise. IEEE Trans. Speech Audio Process. 11 (4), 334–341.
Hu, Y., Loizou, P.C., 2007. A comparative intelligibility study of single-microphone noise reduction algorithms. J. Acoust. Soc. Amer 122 (3), 1777–1786.
Huang, G., Benesty, J., Long, T., Chen, J., 2014. A family of maximum SNR filters for noise reduction. IEEE/ACM Trans. Audio Speech Lang. Process. 22 (12), 2034–2047.
ITU-T, 2007. Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs. ITU-T Recommendation P.862.2. ITU-T.
Ivry, A., Berdugo, B., Cohen, I., 2019. Voice activity detection for transient noisy environment based on diffusion nets. IEEE J. Sel.Top. Signal Process. 13, 254–264.
Kim, G., Lu, Y., Hu, Y., Loizou, P., 2009. An algorithm that improves speech intelligibility in noise for normal-hearing listeners. J. Acoust. Soc. Am. 126 (3), 1486–1494.
Koizumi, Y., Niwa, K., Hioka, Y., Kobayashi, K., Haneda, Y., 2017. DNN-based source enhancement self-optimized by reinforcement learning using sound quality measurements. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), pp. 81–85.
Lee, J., Kang, H.G., 2019. A joint learning algorithm for complex-valued T-F masks in deep learning-based single-channel speech enhancement systems. IEEE/ACM Trans. Audio Speech Lang. Process. 27 (6), 1098–1108.
Lee, J., Skoglund, J., Shabestary, T., Kang, H.G., 2018. Phase-sensitive joint learning algorithms for deep learning-based speech enhancement. IEEE Signal Process. Lett. 25 (8), 1276–1280.
Lim, J.S., Oppenheim, A.V., 1979. Enhancement and bandwidth compression of noisy speech. Proc. IEEE 67 (12), 1586–1604.
Loizou, P.C., 2007. Speech Enhancement Theory and Practice. CRC Press, Boca Raton.
Loizou, P.C., Kim, G., 2011. Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions. IEEE Trans. Audio Speech Lang. Process. 19 (12), 47–56.
Manolakis, D.G., Ingle, V.K., Kogon, S.M., 2005. Statistical and Adaptive Signal Processing. Artech House, Boston.
Miyazaki, R., Saruwatari, H., Inoue, T., Takahashi, Y., Shikano, K., Kondo, K., 2012. Musical-noise-free speech enhancement based on optimized iterative spectral subtraction. IEEE Trans. Acoust. Speech Signal Process. 20 (7), 2080–2094.
Mousazadeh, S., Cohen, I., 2013. Voice activity detection in presence of transient noise using spectral clustering. IEEE Trans. Audio Speech Lang. Process. 21 (6), 1261–1271.
Nakanishi, I., Namba, H., Li, S., 2013. Speech enhancement based on frequency domain ale with adaptive de-correlation parameters. Int. J. Comput. Theory Eng. 5 (2).
Ramli, R.M., Noor, A.O.A., Samad, S.A., 2012. A review of adaptive line enhancers for noise cancellation. Aust. J. Basic Appl. Sci. 6 (6), 337–352.
Sasaoka, N., Shimada, K., Sonobe, S., Itoh, Y., Fujii, K., 2009. Speech enhancement based on adaptive filter with variable step size for wideband and periodic noise. In: Proc. IEEE Int. Midwest Symp. Circuits Syst., pp. 648–652.
Sasaoka, N., Watanabe, M., Itoh, Y., Fujiit, K., 2006. A study on step size control for noise reconstruction system with ale. In: Proc. Int. Symp. Intell. Signal Process. Commun. (ISPACS), pp. 307–310.
Taal, C.H., Hendriks, R.C., Heusdens, R., Jensen, J., 2011. An algorithm for intelligibility prediction of time-frequency weighted noisy speech. IEEE Trans. Audio Speech Lang. Process. 19 (7), 2125–2136.
Taghia, J., Martin, R., 2016. A frequency-domain adaptive line enhancer with step-size control based on mutual information for harmonic noise reduction. IEEE Trans. Audio Speech Lang. Process. 24.
Tucker, R., 1992. Voice activity detection using a periodicity measure. In: Proc. IEEE I, vol. 139, pp. 377–380.
Wang, D., Brown, G.J., 2006. Computational Auditory Scene Analysis: Principles, Algorithms, and Applications. Wiley interscience, Hoboken, New Jersey.
Wang, D.L., Chen, J., 2018. Supervised speech separation based on deep learning: an overview. IEEE/ACM Trans. Audio Speech Lang. Process. 26 (10), 1702–1726.
Widrow, B., Glover, J.R., McCool, J.M., Kaunitz, J., Williams, C.S., Hearn, R.H., R.Zeidler, J., Dong, E., Goodlin, R.C., 1975. Adaptive noise cancelling: principles and applications. Proc. IEEE 63 (12).
Widrow, B., McCool, J.M., Larimore, M.G., 1976. Stationary and nonstationary learning characteristics of the LMS adaptive filter. Proc. IEEE 64 (8).
Williamson, D.S., Wang, Y., Wang, D., 2016. Complex ratio masking for monaural speech separation. IEEE/ACM Trans. Audio Speech Lang. Process. 24 (3), 483–492.
Xu, Y., Du, J., Dai, L.R., Lee, C.H., 2014. An experimental study on speech enhancement based on deep neural networks. IEEE Signal Process. Lett. 21 (1), 65–68.
Xu, Y., Du, J., Dai, L.R., Lee, C.H., 2015. A regression approach to speech enhancement based on deep neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. 23 (1), 7–19.
Zhang, T., Bhowmik, A.K., 2018. Enhancing speech in noisy and reverberant environments using deep learning techniques. In: Proc. SID Symp. Dig. Tech. Pap., vol. 49, pp. 467–470.
Zhang, X., Wang, D.L., 2016. A deep ensemble learning method for monaural speech separation. IEEE/ACM Trans. Audio Speech Lang. Process. 24 (5), 967–977.
Zhang, X.L., Wu, J., 2013. Deep belief networks based voice activity detection. IEEE Trans. Audio Speech Lang. Process. 21 (4), 697–710.
Zhao, Y., Xu, B., Giri, R., Zhang, T., 2018. Perceptually guided speech enhancement using deep neural networks. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), pp. 5074–5078.