SINGLE-CHANNEL SPEECH ENHANCEMENT IN A TRANSIENT NOISE ENVIRONMENT BY EXPLOITING SPEECH HARMONICITY

Kai Wu, V. G. Reju, and Andy W. H. Khong

School of Electrical & Electronic Engineering,
Nanyang Technological University, Singapore,
Email: wu0001ai@e.ntu.edu.sg, {reju, andykhong}@ntu.edu.sg.

ABSTRACT

This paper focuses on the problem of single-channel noise reduction in a transient noise environment for speech enhancement application. A typical speech enhancement algorithm requires an estimate of the noise statistics. However, the problem of noise estimation is challenging when the statistics of the noise vary significantly with time. By exploiting the fact that for speech signal most of the energy is concentrated on the harmonic bands in voiced frames, we propose an algorithm for the estimation of speech presence probability in the time-frequency domain. The estimated speech presence probability is then used for noise estimation for speech enhancement application. Evaluations are conducted to compare the speech enhancement performance between the proposed algorithm and the existing algorithm for various types of transient noise.

Index Terms— Single-channel noise reduction, speech harmonics

1. INTRODUCTION

Single-channel speech enhancement has been investigated intensively for several decades in applications such as hands-free mobile communication, hearing-aids, teleconferencing and speech recognition. However, the problem is challenging due to the fact that no reference signal or spatial information is available for background noise estimation. In addition, the problem becomes more challenging when the statistics of noise change significantly with time.

A typical speech enhancement system consists of two stages: noise spectrum estimation and spectral enhancement of noisy speech signal. For noise estimation, a simple method is the use of a voice activity detector [1] and noise is estimated during speech silence period. However, this method assumes that the noise must be stationary and the speech signal must be sparse. The minimum statistics (MS) based algorithm [2] overcomes the above problem by using minimum of the smoothed noisy signal power spectrum as the noise spectrum. Similar to the MS method, the minima controlled recursive averaging (MCRA) [3] and improved MCRA (IMCRA) [4] are based on the minimum energy tracking but they also include recursive averaging across both time and frequency. Given the noise estimate, various approaches can be applied for enhancement of the noisy signal. The spectral subtraction is one of the simplest methods [5] and many following algorithms have been proposed to mitigate its musical noise problem [6, 7]. By incorporating Gaussian statistical model, the minimum mean-square error (MMSE) log-spectral amplitude (LSA) estimator was proposed to minimize the distortion between clean and estimated speech [8]. In [3, 4, 9], integration of the IMCRA noise estimator and LSA speech spectral enhancement was investigated. In [10], super-Gaussian priors have been applied in MMSE estimator and in [11], time-correlation between speech spectral components are exploited. Another approach proposed for speech enhancement is based on the subspace method in which the algorithm separates the signal and noise into their respective subspaces [12–14]. An overview of the state-of-the-art speech enhancement algorithms can be found in [9, 15, 16].

Although significant progress has been made, speech enhancement in a transient noise environment is still challenging. The conventional MS and IMCRA noise estimation algorithms track the minimum energy in a predefined window in time-frequency (TF) domain. The minimum is then taken as the noise estimate by assuming that the window contains at least a few noise-only frames and that noise is short-term stationary within the window. However, the performance reduces significantly when the noise varies significantly with time. Reduction of window length may help to adapt to fast variation of noise but it may introduce speech distortion if the speech signal is taken as the minimum within a small window. Therefore, efforts were made to mitigate this problem involving the non-stationary nature of noise [17–19].

In this work, we propose a method to suppress the transient noise by exploiting speech characteristics. Exploitation of speech harmonicity has shown to be an effective method in speech enhancement applications, such as regeneration of the degraded harmonics [20, 21] and noise filtering [22, 23]. It has also been applied for speaker tracking in the presence of interference [24]. In this work, the speech harmonic structure is utilized for noise estimation purpose. More specifically, speech harmonics are used to estimate the speech presence probability on each TF point. Recursive averaging is then carried over the past spectral power values that are adjusted by speech presence probability for the estimation of noise. Finally, the estimated noise is used for the spectral gain calculation in LSA estimator for speech enhancement. Simulations are conducted to compare the performance of the proposed method with the existing LSA-IMCRA algorithm [3, 4, 9] in a transient noise environment.

2. PROBLEM FORMULATION

Consider a noisy signal $y(n) = x(n) + d(n)$, where $x(n)$ and $d(n)$ are clean speech signal and additive noise, respectively, and n is the sample index. The short-time Fourier transform (STFT) of the received signal can be represented as $\underline{y}(k,m) = \underline{x}(k,m) + \underline{d}(k,m)$, where $\underline{y}(k,m)$, $\underline{x}(k,m)$ and $\underline{d}(k,m)$ are the STFT coefficients of the received signal, clean speech and noise, respectively, k denotes the frequency bin index and m is the time frame index. Given the uncertainty of noise, two hypotheses $H_0(k,m)$ and $H_1(k,m)$ which, respectively, denotes the absence and presence of a speech signal at the (k,m)th TF point are applied. By assuming a zero-mean com-

plex Gaussian distribution of the STFT coefficients for both speech and noise signals [8,25], the conditional probability density function of the received signal is given as

$$\Pr(\underline{y}(k,m)|H_0(k,m)) = \frac{1}{\pi\lambda_d(k,m)} \exp\left\{ -\frac{|\underline{y}(k,m)|^2}{\lambda_d(k,m)} \right\}, \quad (1)$$

$$\Pr(\underline{y}(k,m)|H_1(k,m)) = \frac{1}{\pi(\lambda_x(k,m)+\lambda_d(k,m))} \exp\left\{ -\frac{|\underline{y}(k,m)|^2}{\lambda_x(k,m)+\lambda_d(k,m)} \right\}, \quad (2)$$

where $\lambda_d(k,m) \triangleq \mathbb{E}\{|\underline{d}(k,m)|^2\}$ denotes the short-term variance of noise signal and $\lambda_x(k,m) \triangleq \mathbb{E}\{|\underline{x}(k,m)|^2|H_1(k,m)\}$ is the short-term variance of speech signal. Furthermore, by applying Bayes rule on (1) and (2), the conditional speech presence probability $p(k,m) \triangleq \Pr(H_1(k,m)|\underline{y}(k,m))$ can be derived as [3,9]

$$p(k,m) = \left\{ 1 + \frac{\{1-f(k,m)\}\{1+\xi(k,m)\}}{f(k,m)\exp\{v(k,m)\}} \right\}^{-1}, \quad (3)$$

where $f(k,m) \triangleq \Pr(H_1(k,m))$ is the a priori speech presence probability, $v(k,m) \triangleq \gamma(k,m)\xi(k,m)/(1+\xi(k,m))$. The variables $\xi(k,m) \triangleq \lambda_x(k,m)/\lambda_d(k,m)$ and $\gamma(k,m) \triangleq |\underline{y}(k,m)|^2/\lambda_d(k,m)$ denote the a priori signal-to-noise ratio (SNR) and the posteriori SNR, respectively.

In general, the speech signal can be estimated by applying a gain function on the noisy signal, i.e.,

$$\underline{\hat{x}}(k,m) = g(k,m)\underline{y}(k,m). \quad (4)$$

Based on the binary hypothesis, an optimal LSA estimator was proposed in [3], where the gain function is derived as

$$g(k,m) = \{g_{H_1}(k,m)\}^{p(k,m)}\{g_{\min}\}^{1-p(k,m)}. \quad (5)$$

In (5),

$$g_{H_1}(k,m) = \frac{\xi(k,m)}{1+\xi(k,m)} \exp\left(\frac{1}{2}\int_{v(k,m)}^{\infty} \frac{e^{-t}}{t} dt \right), \quad (6)$$

is defined as the conditional gain function in the presence of a speech signal [3,8] and $g_{\min}$ is a constant determined by the noise naturalness [3].

3. PROPOSED ALGORITHM

From (3)-(6), an estimate of the noise power spectrum $\lambda_d(k,m)$ is required to obtain the gain function $g(k,m)$ in (5). The performance of the speech enhancement system will therefore be affected by the noise estimation algorithm. In this work, we propose to derive the a priori speech presence probability based on the speech harmonic structure. The noise is then estimated by integrating the speech presence probability $p(k,m)$ in (3) with a recursive averaging method over previous TF points.

3.1. Harmonicity based a priori speech presence probability estimation

The speech presence probability $p(k,m)$ in (3) requires an estimate of a priori speech presence probability $f(k,m)$, as well as the estimates of $\xi(k,m)$ and $\gamma(k,m)$. We address the estimation of $f(k,m)$ first, by exploiting the speech harmonic structure. It is well known that speech energy is concentrated on its harmonic bands in the voiced frames and the harmonics are multiple integers of a pitch frequency (see Fig. 1 (a)). We assume, in this work, that the background noise does not have any harmonic structure or it does not share any harmonic band with the speech signal due to the difference in pitch frequency. The aim here is to estimate the harmonic bands of the speech signal and assign a higher a priori speech presence probability on the harmonic bands. The other frequency bins are expected to contain less speech information and might be corrupted by noise. For these frequency bins, a lower a priori speech presence probability will be assigned and these frequency bins will be used for noise estimation.

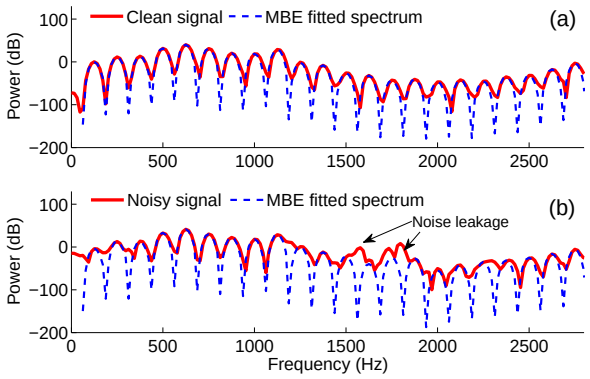

Fig. 1. Multi-band excitation fitting results for (a) a clean speech signal and (b) a noise corrupted signal at SNR = 5 dB.

To estimate the speech harmonics, the multi-band excitation (MBE) fitting method [24, 26] which is commonly employed in speech coding can be applied. Consider, for example, a clean speech voiced frame $\underline{x}(k,m)$. Defining spectrum envelop as $\eta(k,m)$ and excitation spectrum as $\chi(k,m)$, the MBE model is given by

$$\underline{x}(k,m) \triangleq \eta(k,m)\chi(k,m), \quad (7)$$

$$\chi(k,m) = \sum_{q=1}^{Q} \varpi(k - qk_m^{\mathrm{p}}), \quad (8)$$

where $\varpi(k)$ is the DFT coefficient of the STFT analysis window, k_m^{p} is the bin index corresponding to speech pitch frequency at the mth time frame, q is the harmonic band index and Q is the total number of harmonic bands considered.

Given $\underline{y}(k,m)$ at the mth frame, the MBE model fitting can be achieved by estimating $\eta(k,m)$ and k_m^{p} as defined in (7) and (8). We first decouple the spectrum envelope $\eta(k,m)$ into complex amplitude $\eta_q(m)$ for each harmonic band, and specify the frequency interval $[a_q,\ b_q]$ for each of the harmonic bands, where $a_q = \lfloor (q-0.5)k_m^{\mathrm{p}} \rceil$, $b_q = \lfloor (q+0.5)k_m^{\mathrm{p}} \rceil$ and $\lfloor \cdot \rceil$ denotes nearest integer. The MBE fitting process can then be performed on $\underline{y}(k,m)$ by minimizing the fitting error across all the harmonic bands given by

$$\mathcal{E}(m) = \sum_{q=1}^{Q} \sum_{k=a_q}^{b_q} \left| \underline{y}(k,m) - \eta_q(m)\chi(k,m) \right|^2. \quad (9)$$

The minimization of (9) is with respect to two variables, $\eta_q(m)$ and k_m^{p}. Equating the derivative of (9) with respect to $\eta_q(m)$ to zero, we have

$$\eta_q(m) = \frac{\sum_{k=a_q}^{b_q} \underline{y}(k,m)\chi^*(k,m)}{\sum_{k=a_q}^{b_q} |\chi(k,m)|^2}, \quad (10)$$

where $(\cdot)^*$ denotes complex conjugate operator. Substituting (10) into (9), the error function of (9) is then computed with respect to all

pitch frequency bin indices of interest $k_m^{\mathrm{p}} \in [k_{\min}^{\mathrm{p}},\ k_{\max}^{\mathrm{p}}]$. Finally, the minimum of $\mathcal{E}(m)$ is determined and the corresponding k_m^{p} is selected as the estimated $\widehat{k}_m^{\mathrm{p}}$ of the mth frame. The MBE estimated signal can therefore be given by $\widehat{\underline{x}}_{\mathrm{MBE}}(k,m) = \widehat{\eta}(k,m)\widehat{\chi}(k,m)$. It worth noting that $\widehat{\underline{x}}_{\mathrm{MBE}}(k,m)$ cannot be directly taken as the enhanced output signal since it may contain artificial effects. In this work, $\widehat{\underline{x}}_{\mathrm{MBE}}(k,m)$ is used to estimate $f(k,m)$ for noise estimation purpose.

Figure 1 (a) shows the MBE fitting result for a 32 ms voiced frame of a clean speech using (9) to (10). It can be seen that the MBE approximation, shown by the dashed line, is capable of fitting the clean speech spectrum with only a very small error. Figure 1 (b) shows the result for the same voiced frame corrupted by a power drill noise, where the noise energy leakage mainly occurs at approximately 1600 and 1800 Hz. Comparing Figs. 1 (a) and (b), we note that the speech harmonics which are less corrupted by the noise can be identified by MBE fitting, while the other frequencies outside the harmonic bands can be used for noise estimation.

In order to estimate $f(k,m)$, we propose to apply a linear mapping on the log spectrum of $\widehat{\underline{x}}_{\mathrm{MBE}}(k,m)$, i.e.,

$$\widehat{f}(k,m) = \frac{\mathcal{L}|\widehat{\underline{x}}_{\mathrm{MBE}}(k,m)| - \min\limits_{k}(\mathcal{L}|\widehat{\underline{x}}_{\mathrm{MBE}}(k,m)|)}{\max\limits_{k}(\mathcal{L}|\widehat{\underline{x}}_{\mathrm{MBE}}(k,m)|) - \min\limits_{k}(\mathcal{L}|\widehat{\underline{x}}_{\mathrm{MBE}}(k,m)|)}, \quad (11)$$

where $\mathcal{L}|\widehat{\underline{x}}_{\mathrm{MBE}}(k,m)| = 20\log_{10}|\widehat{\underline{x}}_{\mathrm{MBE}}(k,m)|$. The rational behind (11) is that the fitted harmonic bands with a higher energy should be assigned a higher a priori speech presence probability, while the other frequencies should be assigned lower probability values. Further, a normalized fitting error for each of the harmonic bands can be defined as [24],

$$\varepsilon_q(m) = \frac{\sum_{k=a_q}^{b_q} |\underline{y}(k,m) - \widehat{\underline{x}}_{\mathrm{MBE}}(k,m)|^2}{\sum_{k=a_q}^{b_q} |\underline{y}(k,m)|^2}, \quad (12)$$

which gives a deviation measure of the noisy signal spectrum from the MBE model. The total harmonic energy $\mathcal{P}(m)$ for the mth frame can therefore be defined by a summation of the energy on each harmonic band tuned by the fitting error $\varepsilon_q(m)$, i.e.,

$$\mathcal{P}(m) = \sum_{q=1}^{Q}\left\{(1-\varepsilon_q(m))\sum_{k=a_q}^{b_q}|\widehat{\underline{x}}_{\mathrm{MBE}}(k,m)|^2\right\}. \quad (13)$$

From (13), it can be seen that for speech-dominant frames whose energies are concentrated on the harmonic bands with smaller fitting errors $\varepsilon_q(m)$, $\mathcal{P}(m)$ would approach to a higher value. Therefore, a weighting can be applied across time frames as

$$w(m) = \begin{cases} 1, & \text{if } \mathcal{P}(m) \geq \mathcal{P}^{\mathrm{u}}, \\ \dfrac{\mathcal{P}(m) - \min_m(\mathcal{P}(m))}{\mathcal{P}^{\mathrm{u}} - \min_m(\mathcal{P}(m))}, & \text{otherwise}, \end{cases} \quad (14)$$

such that $w(m) \to 1$ indicates a speech frame and vise versa. The variable $\mathcal{P}^{\mathrm{u}}$ is an upper-bound threshold. In this work, we have used $\mathcal{P}^{\mathrm{u}} = 1.5 \cdot \mathrm{median}_m(\mathcal{P}(m))$ in order to be more conservative on keeping the frames which might contain speech information. With $w(m)$, $\widehat{f}(k,m)$ can be rewritten as

$$\widehat{f}(k,m) \Leftarrow \widehat{f}(k,m) \cdot w(m). \quad (15)$$

It worth noting that for the unvoiced speech frames the value of $\mathcal{P}(m)$ and hence $\widehat{f}(k,m)$ will be small. However, due to the fact that the voiced speech segments may present before the unvoiced speech segments and the recursive nature of the noise estimation described in Sec. 3.2, the unvoiced segments of the speech will not be completely removed.

Table 1. Summary of the proposed algorithm.

Initialize $\bar{\lambda}_d(k,1) = |\underline{y}(k,1)|^2$, $\widehat{\lambda}_d(k,1) = |\underline{y}(k,1)|^2$, $g_{H_1}(k,1) = 1$. Estimate the harmonics for all the frames of $\underline{y}(k,m)$ using (9)-(10) and $\widehat{f}(k,m)$ using (11)-(15).

for frame index $m \geq 2$:

1. Estimate $\widehat{\gamma}(k,m)$ using $\widehat{\gamma}(k,m) = \frac{|\underline{y}(k,m)|^2}{\widehat{\lambda}_d(k,m-1)}$.
2. Estimate $\widehat{\xi}(k,m)$ using (16) and $g_{H_1}(k,m)$ using (6).
3. Estimate $\widehat{p}(k,m)$ using (3).
4. Update the noise estimate $\widehat{\lambda}_d(k,m)$ using (17)-(19).
5. Compute $g(k,m)$ using (5) with another speech presence probability $\widehat{p}'(k,m)$ derived in [3].
6. Enhance the noisy signal using (4).

end

3.2. Noise estimation and speech enhancement

Given $\widehat{f}(k,m)$, noise estimation and hence the speech enhancement is performed in an iterative manner. From (3), this implies that $\xi(k,m)$ and $\gamma(k,m)$ need to be estimated recursively using each time frame signal. In this work, we employ the a priori SNR estimator [3,4], which has been reported to achieve better performance than the commonly used "decision-directed" approach [25]. In particular, the a priori SNR can be recursively estimated by

$$\begin{aligned}\widehat{\xi}(k,m) =& \alpha_\xi g_{H_1}^2(k,m-1)\widehat{\gamma}(k,m-1) \\ &+ (1-\alpha_\xi)\max\{\widehat{\gamma}(k,m)-1, 0\}, \end{aligned} \quad (16)$$

where $g_{H_1}(k,m)$ has been defined in (6) and $\widehat{\gamma}(k,m)$ is the posteriori SNR which will be estimated iteratively using $\widehat{\lambda}_d(k,m-1)$. Therefore, $\widehat{p}(k,m)$ can be obtained by applying $\widehat{f}(k,m)$, $\widehat{\xi}(k,m)$ and $\widehat{\gamma}(k,m)$ on (3).

For noise estimation, a recursive averaging is applied over the past spectral power values that is adjusted by $\widehat{p}(k,m)$ [4], i.e.,

$$\begin{aligned}\bar{\lambda}_d(k,m) =& \bar{\lambda}_d(k,m-1)\widehat{p}(k,m) \\ &+ \bar{\lambda}_{d,H_0}(k,m)(1-\widehat{p}(k,m)), \end{aligned} \quad (17)$$

where

$$\bar{\lambda}_{d,H_0}(k,m) = \alpha_d\bar{\lambda}_d(k,m-1) + (1-\alpha_d)|\underline{y}(k,m)|^2, \quad (18)$$

denotes the averaged noise estimate under the assumption of speech absence, and α_d is a predefined smoothing parameter. From (17), if $\widehat{p}(k,m) \to 1$ which indicates a speech presence at corresponding TF point, noise estimation will be held and $\bar{\lambda}_d(k,m-1)$ will be taken as $\bar{\lambda}_d(k,m)$. If $\widehat{p}(k,m) \to 0$, the recursive averaging will be carried over the past spectral power values as indicated in (18). It worth noting that the estimate of noise will be biased when the recursive averaging in (17) and (18) is used [3, 4], Thus, a constant compensation factor β can be applied as [4]

$$\widehat{\lambda}_d(k,m) = \beta \cdot \bar{\lambda}_d(k,m). \quad (19)$$

Finally, speech enhancement can be achieved using (4)-(6). It worth mentioning that two kinds of speech presence probability $p(k,m)$ are used in LSA-IMCRA algorithm [3, 4, 9]. One is based on the TF distribution of $\widehat{\xi}(k,m)$ and it is used for gain calculation

Table 2. Segmental SNR improvement for different types of transient noise (dB).

	WGN		Power drill noise		Destroyer engine noise		Babble noise	
input SNR	LSA-IMCRA	Proposed	LSA-IMCRA	Proposed	LSA-IMCRA	Proposed	LSA-IMCRA	Proposed
0 dB	3.51	**7.62**	3.45	**6.38**	4.86	**7.71**	3.50	**3.52**
5 dB	3.05	**6.29**	3.24	**5.76**	4.20	**6.65**	3.19	**3.47**
10 dB	2.47	**4.38**	2.99	**4.20**	3.79	**4.87**	2.77	**2.91**

Table 3. Segmental LSD reduction for different types of transient noise (dB).

	WGN		Power drill noise		Destroyer engine noise		Babble noise	
input SNR	LSA-IMCRA	Proposed	LSA-IMCRA	Proposed	LSA-IMCRA	Proposed	LSA-IMCRA	Proposed
0 dB	5.73	**10.12**	2.89	**4.16**	3.28	**4.48**	2.19	**2.22**
5 dB	4.95	**7.95**	2.28	**2.96**	2.48	**2.93**	1.73	**1.74**
10 dB	4.22	**5.74**	1.75	**1.94**	1.74	1.66	1.28	1.26

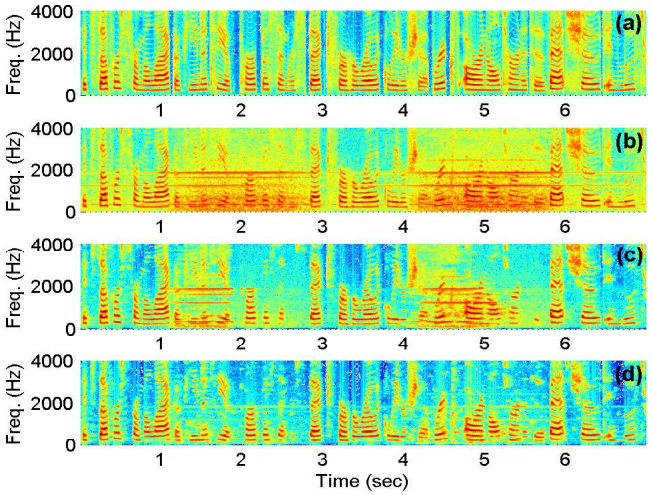

Fig. 2. The spectrograms of (a) a clean speech signal, (b) a signal corrupted by a power drill interference noise at segmental SNR $= 5$ dB during 1 s to 3 s and 4 s to 6 s, (c) enhanced signal by the LSA-IMCRA algorithm and (d) enhanced signal by the proposed algorithm.

in (5). The other is derived from IMCRA and it is used for noise estimation. In this work, we follow the same framework where our $\widehat{p}(k,m)$ is used only for noise estimation purpose. The same TF-distribution of $\widehat{\xi}(k,m)$ based speech presence probability estimator proposed in [3] is used for (5). The proposed algorithm is summarized in Table 1.

4. SIMULATION RESULTS

Evaluations are conducted to compare the performance of the proposed algorithm with the existing LSA-IMCRA algorithm [3,4,9] in a transient noise environment. Both male and female speech signals with length of 7 s sampled at 16 kHz from the TIMIT database [27] were used as clean signals. For noise, white Gaussian noise (WGN), power drill noise, destroyer engine noise and babble noise obtained from NOISEX-92 database [28] were used. The noise was added in 20 dB SNR during the whole signal length, in which two transient periods with length of 2 s were simulated by reducing the corresponding segmental SNR to a range of 0 to 10 dB. For STFT analysis, a Hamming window of 512 samples length (32 ms) with 75% overlap was used. For higher frequency resolution, the 512 samples data frame was padded with zeros and transformed to DFT domain using 1024 point FFT. For the proposed method, the speech pitch frequency was estimated from a range of $[90,\ 300]$ Hz corresponding to $[k^{\mathrm{p}}_{\min},\ k^{\mathrm{p}}_{\max}] = [7,\ 20]$, and $Q = \lfloor k^{\mathrm{m}}_{\mathrm{spch}}/\widehat{k}^{\mathrm{p}}_{m} \rceil$ where $k^{\mathrm{m}}_{\mathrm{spch}} = 321$ corresponding to 5 kHz is the maximum frequency-bin index of speech signal considered, $\widehat{k}^{\mathrm{p}}_{m}$ is the estimated pitch frequency-bin index and $\lfloor \cdot \rceil$ denotes nearest integer. The other parameters were $\alpha_d = 0.85$, $\alpha_\xi = 0.95$, $g_{\min} = -15$ dB, $\beta = 1.2$.

Two objective measures are used for performance evaluation. To measure the noise suppression performance, the segmental SNR is defined as [3, 9, 19]

$$\mathrm{SNR} = 10\log_{10}\frac{\mathbb{E}_n\{\underline{x}^2(n)\}}{\mathbb{E}_n\{|\underline{\widehat{x}}(n)-\underline{x}(n)|^2\}}, \tag{20}$$

where $\mathbb{E}_n(\cdot)$ is the expectation over the samples in a time segment in which the transient noise is present. Also, in order to measure the distortion, the segmental log spectral distance (LSD) is defined as [9, 19]

$$\mathrm{LSD} = \mathbb{E}_m\left\{\frac{2}{K}\sum_{k=1}^{K/2}\left|\mathcal{L}'\widehat{\lambda}_x(k,m)-\mathcal{L}'\lambda_x(k,m)\right|^2\right\}^{\frac{1}{2}}, \tag{21}$$

where $\mathcal{L}'\lambda(k,m) = \max\{10\log_{10}\lambda(k,m),\ \delta\}$ and δ is a small value defined as $\delta = \max_{(k,m)}\{10\log_{10}\lambda(k,m) - 50\}$ in order to confine the log spectrum within 50 dB dynamic range [9, 19]. The operator $\mathbb{E}_m(\cdot)$ denotes expectation over time frames.

Figure 2 compares the proposed algorithm with the LSA-IMCRA algorithm. Figure 2 (a) shows the spectrogram of a clean female speech signal while Fig. 2 (b) shows the noisy signal in which the noise level of a power drill interference noise was increased during 1 s to 3 s and 4 s to 6 s, both at a segmental SNR of 5 dB. The LSA-IMCRA algorithm, shown in Fig. 2 (c), achieves noise suppression to some extent. However, the algorithm fails to suppress the noise during 1 to 2 s and 4 to 5 s due to the fact that the algorithm requires time to estimate the actual noise spectrum using the IMCRA noise estimator. The segmental SNR improvement and LSD reduction are 3.6 dB and 2.7 dB, respectively, for the LSA-IMCRA algorithm. From Fig. 2 (d), it is clear that the proposed algorithm can achieve better suppression performance than LSA-IMCRA. In this case, 6.9 dB SNR improvement and 3.7 dB LSD reduction is achieved by the proposed algorithm.

Table 2 and 3 summarize the segmental SNR and LSD reduction for various noise types and noise levels averaged over 30 trials of different speech signals. In general, the proposed algorithm achieves better performance than the LSA-IMCRA algorithm. For example, the proposed algorithm achieves additional 3.2 dB segmental SNR improvement and 3.0 dB LSD reduction than the LSA-IMCRA algorithm for the case of WGN at 5 dB input SNR.

5. CONCLUSION

A speech harmonicity based single-channel speech enhancement algorithm is proposed. An a priori speech presence probability is estimated by utilizing speech harmonic structure in voiced frames. Noise estimation is then achieved by integrating the recursive averaging framework with the estimated speech presence probability and used for speech enhancement. Objective evaluations show that the proposed algorithm achieves better performance than the existing LSA-IMCRA algorithm for various types of transient noise.

6. REFERENCES

[1] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Letters*, vol. 6, no. 1, pp. 1–3, 1999.

[2] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech and Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

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

[4] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Trans. Speech and Audio Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.

[5] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust. Speech and Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[6] Z. Goh, K.-C. Tan, and B. T. G. Tan, "Postprocessing method for suppressing musical noise generated by spectral subtraction," *IEEE Trans. Speech and Audio Process.*, vol. 6, no. 3, pp. 287–292, May 1998.

[7] H. Gustafsson, S. E. Nordholm, and I. Claesson, "Spectral subtraction using reduced delay convolution and adaptive averaging," *IEEE Trans. Speech and Audio Process.*, vol. 9, no. 8, pp. 799–807, Nov 2001.

[8] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust. Speech and Signal Process.*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[9] I. Cohen and S. Gannot, "Spectral enhancement methods," in *Springer Handbook of Speech Processing*, pp. 873–902. Springer, 2008.

[10] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Speech and Audio Process.*, vol. 13, no. 5, pp. 845–856, Sept 2005.

[11] I. Cohen, "Relaxed statistical model for speech enhancement and a priori snr estimation," *IEEE Trans. Speech and Audio Process.*, vol. 13, no. 5, pp. 870–881, Sept 2005.

[12] Y. Ephraim and H. L. V. Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech and Audio Process.*, vol. 3, no. 4, pp. 251–266, Jul 1995.

[13] Yi Hu and P.C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Trans. Speech and Audio Process.*, vol. 11, no. 4, pp. 334–341, Jul. 2003.

[14] F. Jabloun and B. Champagne, "Incorporating the human hearing properties in the signal subspace approach for speech enhancement," *IEEE Trans. Speech and Audio Process.*, vol. 11, no. 6, pp. 700–708, Nov. 2003.

[15] P. C. Loizou, *Speech enhancement: theory and practice*, CRC press, 2013.

[16] J. Benesty, S. Makino, and J. Chen, *Speech enhancement*, Springer, 2005.

[17] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'10)*. IEEE, 2010, pp. 4266–4269.

[18] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, 2012.

[19] R. Talmon, I. Cohen, and S. Gannot, "Single-channel transient interference suppression with diffusion maps," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 21, no. 1, pp. 132–144, 2013.

[20] A.-T. Yu and H.-C. Wang, "New speech harmonic structure measure and it application to post speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'04)*. IEEE, 2004, vol. 1, pp. I–729.

[21] C. Plapous, C. Marro, and P. Scalart, "Improved signal-to-noise ratio estimation for speech enhancement," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 14, no. 6, pp. 2098–2108, 2006.

[22] J. Wen, X. Liu, M. S. Scordilis, and H. Lu, "Speech enhancement using harmonic emphasis and adaptive comb filtering," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 18, no. 2, pp. 356–368, 2010.

[23] J. R. Jensen, J. Benesty, M. G. Christensen, and S. H. Jensen, "Enhancement of single-channel periodic signals in the time-domain," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 20, no. 7, pp. 1948–1963, 2012.

[24] K. Wu, S. T. Goh, and A. W. H. Khong, "Speaker localization and tracking in the presence of sound interference by exploiting speech harmonicity," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'13)*, 2013, pp. 365–369.

[25] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust. Speech and Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[26] D. W. Griffin and J. S. Lim, "Multiband excitation vocoder," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 36, no. 8, pp. 1223–1235, 1988.

[27] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgrena, and V. Zue, *TIMIT Acoustic-Phonetic Continuous Speech Corpus*, Philadelphia, PA, 1993.

[28] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, 1993.