Exploiting Conjugate Symmetry of the Short-Time Fourier Spectrum for Speech Enhancement

Kamil Wójcicki, Mitar Milacic, Anthony Stark, James Lyons, and Kuldip Paliwal, *Member, IEEE*

***Abstract*—Typical speech enhancement algorithms operate on the short-time magnitude spectrum, while keeping the short-time phase spectrum unchanged for synthesis. We propose a novel approach where the noisy magnitude spectrum is recombined with a changed phase spectrum to produce a modified complex spectrum. During synthesis, the low energy components of the modified complex spectrum cancel out more than the high energy components, thus reducing background noise. Using objective speech quality measures, informal subjective listening tests and spectrogram analysis, we show that the proposed method results in improved speech quality.**

***Index Terms*—Magnitude spectrum, phase spectrum, speech enhancement.**

I. Introduction

In the field of speech enhancement, we are interested in the reduction of noise from noise-corrupted speech in order to improve its intelligibility and quality. Various methods have been investigated in the literature for performing speech enhancement. These can be grouped into spectral subtraction [1], MMSE estimation [2], Wiener filtering (linear MMSE) [3], Kalman filtering [4], and subspace [5] methods. Several of these methods employ the analysis-modification-synthesis (AMS) framework [6]–[9].

Let us consider an additive noise model

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

where $x(n)$, $s(n)$, and $d(n)$ denote discrete-time signals of noisy speech, clean speech, and noise, respectively. Since speech can be assumed to be quasi-stationary, it is analyzed frame-wise in the AMS framework through the short-time Fourier analysis. The discrete short-time Fourier transform (DSTFT) of the corrupted speech signal $x(n)$ is given by

$$X(n,k) = \sum_{m=-\infty}^{\infty} x(m)w(n-m)\mathrm{e}^{-j2\pi km/N} \tag{2}$$

where k denotes the kth discrete-frequency of N uniformly spaced frequencies and $w(n)$ is an analysis window function. In speech processing, the Hamming window with 20–40 ms duration is typically employed. Using DSTFT analysis, we can equally, subject to constraints described in [10], represent (1) as

$$X(n,k) = S(n,k) + D(n,k) \tag{3}$$

Manuscript received November 8, 2007; revised December 12, 2008. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Li Deng.

The authors are with the Signal Processing Laboratory, Griffith University, Nathan QLD 4111, Australia.

Digital Object Identifier 10.1109/LSP.2008.923579

where $X(n,k)$, $S(n,k)$, and $D(n,k)$ are the DSTFTs of noisy speech, clean speech, and noise, respectively. Each of these can be expressed in terms of the DSTFT magnitude spectrum and the DSTFT phase spectrum. For instance, the DSTFT of the noisy speech signal can be written in polar form as

$$X(n,k) = |X(n,k)|\,\mathrm{e}^{j\angle X(n,k)} \tag{4}$$

where $|X(n,k)|$ denotes the magnitude spectrum and $\angle X(n,k)$ denotes the phase spectrum.[1]

Traditional AMS-based speech enhancement methods modify only the magnitude spectrum while keeping the noisy phase spectrum unchanged for synthesis. In this letter, we propose a novel approach to speech enhancement, in which the noisy magnitude spectrum is recombined with a changed phase spectrum to produce a modified complex spectrum. During synthesis, low energy components of the modified complex spectrum cancel out more than the high energy components, resulting in background noise reduction. Thus, our method is well suited for scenarios where the noise energy is lower than the speech energy. Using an objective speech quality measure, informal subjective listening tests, and spectrogram analysis, we demonstrate that the proposed method results in improved speech quality.

The rest of this letter is organized as follows. Section II presents details of the proposed approach. Section III describes the experimental setup. The results and discussion are presented in Section IV. Conclusions are drawn in Section V.

II. Proposed Method

A. Enhancement Procedure

The proposed speech enhancement method is based on the AMS framework commonly employed in speech processing. The AMS framework consists of three stages: 1) the analysis stage, where the input speech is processed using DSTFT analysis [see (2)]; 2) the modification stage, where the noisy complex spectrum undergoes some kind of modification; and 3) the synthesis stage, where the inverse discrete short-time Fourier transform (IDSTFT) operation is followed by the overlap-add (OLA) synthesis to reconstruct the output signal. A block diagram of the proposed approach is shown in Fig. 1.

The noisy speech signal, used in the analysis stage of the AMS framework, is a real-valued signal, and therefore, its DSTFT is conjugate symmetric, i.e., $X(n,k) = X^*(n,N-k)$. In our approach, we control the degree to which the conjugates reinforce or cancel by altering their angles. That is, we compute the changed phase spectrum as follows. First, the

[1]In our discussions, when referring to the magnitude spectrum, phase spectrum, and complex spectrum, the DSTFT modifier is implied unless otherwise stated.

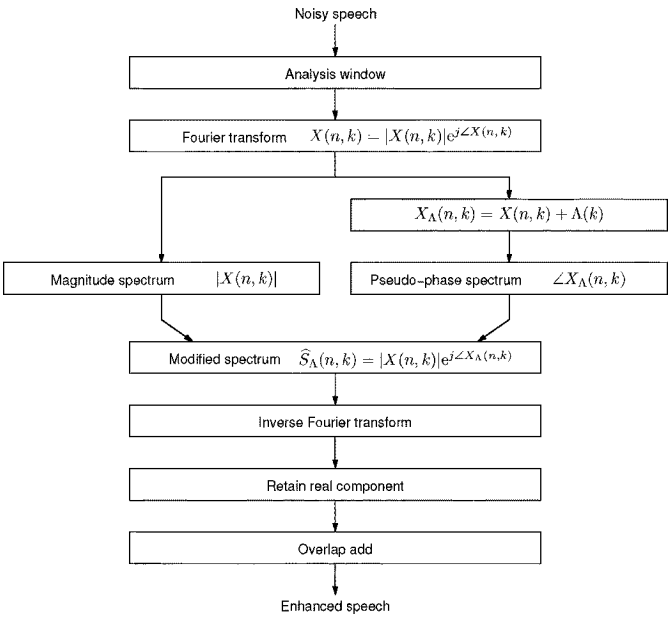

Fig. 1. Block diagram of the proposed speech enhancement method.

noisy complex spectrum is offset by an additive real-valued frequency-dependent $\Lambda(k)$ function

$$X_\Lambda(n,k) = X(n,k) + \Lambda(k) \tag{5}$$

where $\Lambda(k)$ should be made anti-symmetric about $F_s/2$ (half the sampling rate) to achieve the cancellation effect. In this letter, we employ a simple anti-symmetric $\Lambda(k)$ function given by

$$\Lambda(k) = \begin{cases} +\lambda & 0 \le k < N/2 \\ -\lambda & N/2 \le k \le N-1 \end{cases} \tag{6}$$

where λ is a real-valued constant and N is the length of frequency analysis assumed to be even. Second, $X_\Lambda(n,k)$ is used to compute the changed phase spectrum through a four-quadrant version of the arctangent function

$$\angle X_\Lambda(n,k) = \arctan\left(\frac{\mathrm{Im}\{X_\Lambda(n,k)\}}{\mathrm{Re}\{X_\Lambda(n,k)\}}\right) \tag{7}$$

where $\mathrm{Im}\{\cdot\}$ and $\mathrm{Re}\{\cdot\}$ denote imaginary and real operators, respectively. We shall refer to the changed phase spectrum as the pseudo-phase spectrum, since it need not possess the properties of a true[2] phase spectrum. The pseudo-phase spectrum is recombined with the noisy magnitude spectrum to produce a modified complex spectrum

$$\widehat{S}_\Lambda(n,k) = |X(n,k)|\, e^{j\angle X_\Lambda(n,k)}. \tag{8}$$

In the synthesis stage, the IDSTFT is used to convert frequency-domain frames, $\widehat{S}_\Lambda(n,k)$, to time-domain representation. Note that due to the additive offset introduced in (5), the resulting time-domain frames may be complex. In the proposed method, the imaginary component is discarded. The enhanced time-domain signal, $\hat{s}(n)$, is produced by employing the OLA procedure.

[2]In other words, one that is computed from a real-valued signal.

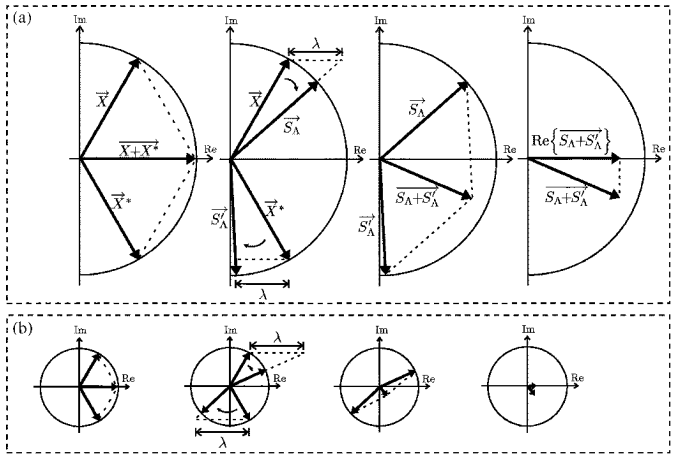

Fig. 2. Vector diagrams: modification of DSTFT conjugate symmetry. Top row (a): $|\vec{X}| > \lambda$. Bottom row (b): $|\vec{X}| < \lambda$. Column one: conjugate vectors, $\vec{X}$ and $\vec{X^*}$, as well as their addition vector, $\overrightarrow{X+X^*}$. Column two: the real parts of the conjugate vectors are offset by λ and $-\lambda$, respectively. Thus, the angles of vectors $\vec{X}$ and $\vec{X^*}$ are altered, while their magnitudes are kept unchanged to produce vectors $\vec{S}_\Lambda$ and $\vec{S}'_\Lambda$, respectively [see (8)]. Column three: the resulting vectors are added to produce the $\overrightarrow{S_\Lambda + S'_\Lambda}$ vector. Column four: the imaginary part of the $\overrightarrow{S_\Lambda + S'_\Lambda}$ addition vector is discarded.

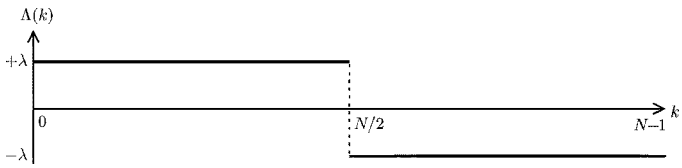

Fig. 3. Anti-symmetric $\Lambda(k)$ function employed in our evaluations, where λ is a constant and k is the kth discrete-frequency component of N uniformly spaced frequencies.

B. Explanation Using Vector Diagrams

The noise cancellation can be described by viewing the frequency-domain signal as groupings of conjugates. The conjugates arise naturally from the symmetry of the magnitude spectrum and anti-symmetry of the phase spectrum and are a result of applying the DSTFT to a real-valued signal. During the IDSTFT operation, the conjugates sum together to produce a larger real-valued signal. By modifying the conjugates, we can influence the degree to which they sum together and thus contribute to the reconstructed time-domain signal. We achieve this by modifying their angles. Two cases of conjugate vector phase modification are presented in Fig. 2. In the first case, Fig. 2(a), the magnitudes of the conjugates, i.e., $|\vec{X}|$ and $|\vec{X^*}|$, are larger than λ (the magnitude of the $\Lambda(k)$ function).[3] This results in limited change of the original signal. In the second case, Fig. 2(b), the vector magnitudes are smaller than λ. A significant change occurs, as the two conjugate vectors, $\vec{X}$ and $\vec{X^*}$, are pushed toward the real-axis facing 0 and π radians, respectively. Summation produces a significant cancellation, leaving little or no real-valued component. As can be seen in Fig. 2, the strength of cancellation for a given λ is dependent on the DSTFT magnitude of the noisy speech, $|\vec{X}| = |\vec{X^*}|$, with larger magnitude components being less attenuated and smaller magnitude components being more attenuated. Typically, noise frequency components

[3]For the purposes of this section, we adopt a vector notation. Also, for clarity, we drop both time and frequency indexes.

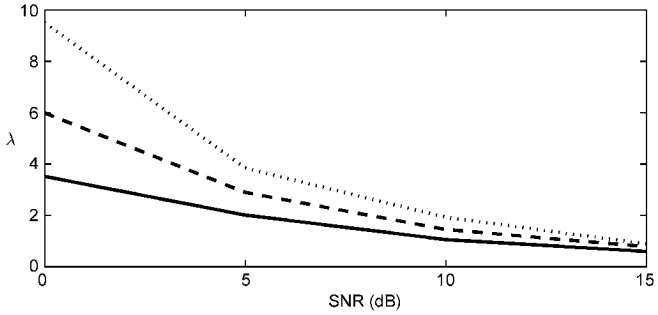

Fig. 4. Empirically determined λ as a function of input speech SNR for white Gaussian noise (solid line), train noise (broken line), and babble noise (dotted line).

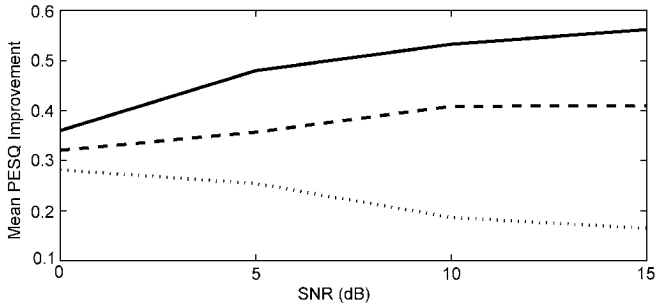

Fig. 5. Mean PESQ improvement scores as a function of input speech SNR for white Gaussian noise (solid line), train noise (broken line), and babble noise (dotted line).

are assumed to have much lower magnitudes than signal components. This assumption is the basis for many noise cancellation and noise estimation algorithms. Using this assumption, we can tune the additive $\Lambda(k)$ function so as to induce significant phase modification that results in cancellation among noise vectors while limiting the modification effect upon signal carrying vectors.

III. Enhancement Experiments

A. Speech Corpus and Noise Types

In our evaluations, we use the NOIZEUS speech corpus [11]. NOIZEUS is composed of 30 phonetically-balanced sentences belonging to six speakers (three males and three females). The corpus is sampled at 8 kHz and filtered to simulate receiving frequency characteristics of telephone handsets. The NOIZEUS corpus comes with nonstationary noises at different SNRs. In our evaluation, we use the train and babble noises. We also generate a corresponding stimuli set corrupted by additive white Gaussian noise at four SNR levels: 0 dB, 5 dB, 10 dB, and 15 dB.

B. Evaluation Methods

For evaluation purposes, we employ an objective speech quality measure, namely, the perceptual estimation of speech quality (PESQ). The PESQ algorithm [12] is a fusion of two other perceptually motivated objective speech quality measures: PAMS and PSQM99. PESQ produces robust estimates of speech quality in the presence of a wide range of noise types. PESQ prediction maps mean opinion score (MOS) estimates to a range between -0.5 and 4.5, where 1.0 corresponds to

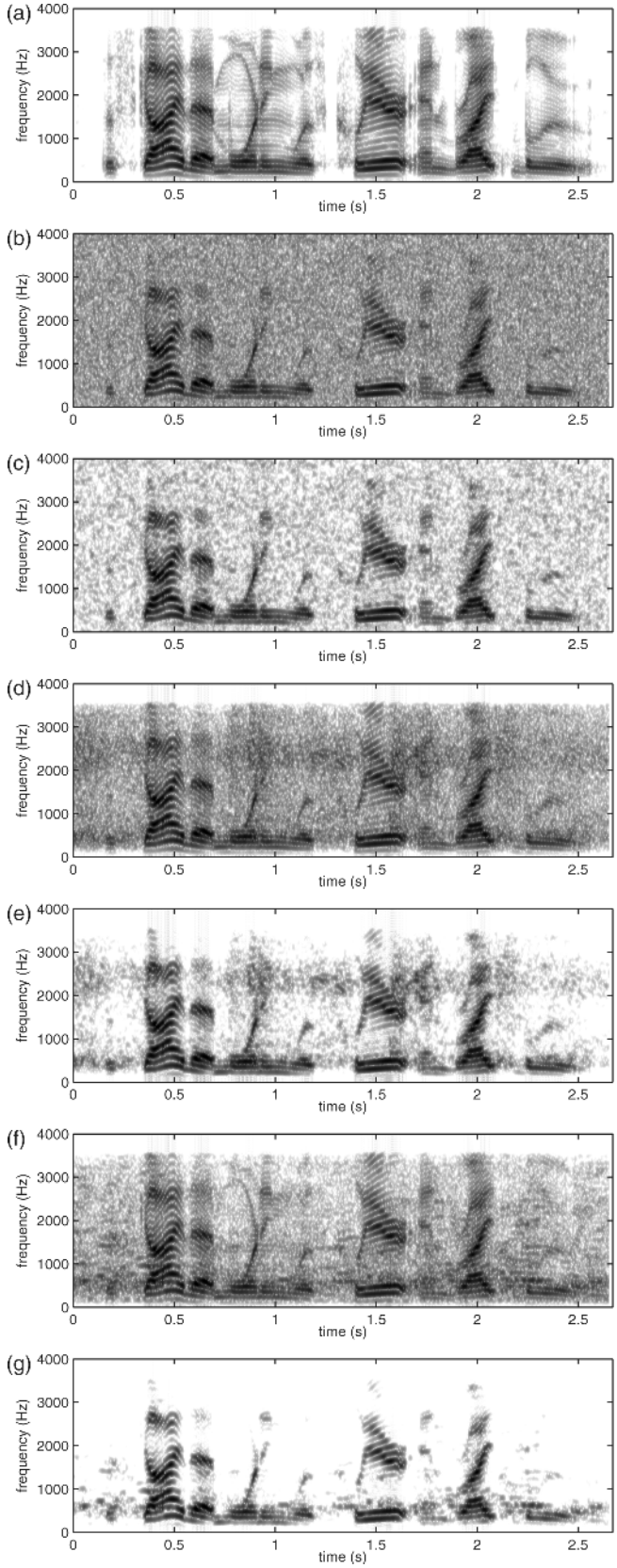

Fig. 6. Spectrograms of *sp10.wav* utterance, "The sky that morning was clear and bright blue," by a male speaker from the NOIZEUS corpus: (a) clean speech; (b, d, f) speech degraded by white noise, train noise, and babble noise, respectively (10 dB SNR); (c, e, g) corresponding enhanced speech.

TABLE I
MEAN PESQ SCORES FOR THE WHITE NOISE CASE FOR THE PROPOSED, SPECTRAL SUBTRACTION (SSUB), AND MMSE METHODS

INPUT SPEECH			METHODS		
SNR (dB)	CLEAN	NOISY	PROPOSED	SSUB	MMSE
0	4.50	1.60	1.97	1.76	1.97
5	4.50	1.87	2.35	2.21	2.35
10	4.50	2.16	2.70	2.65	2.65
15	4.50	2.50	3.06	3.06	2.96

bad and 4.5 corresponds to *distortionless*. In our evaluation, we compute mean PESQ scores over a subset of the NOIZEUS corpus. In addition, we employ informal subjective listening tests and spectrogram analysis.

C. Experimental Procedure

To evaluate the approach described in Section II, we employ the modified AMS procedure shown in Fig. 1. We zero-mean and normalize samples of each of the sentence files to be between -1.0 and $+1.0$. In our experiments, the frame duration is set to 32 ms and the frame shift to 4 ms. The Hamming window is used as the analysis window. We employ FFT length of 1024 samples. We use an anti-symmetric $\Lambda(k)$ function given in (6) and shown in Fig. 3. The value of λ for use in our evaluations was determined empirically in such a way as to maximize both PESQ and SNR scores. The empirical mappings were performed using a gender balanced subset of the NOIZEUS corpus. The subset consisted of 14 utterances. The remainder of the corpus was used for testing. Fig. 4 shows the resulting mappings as a function of input speech SNR for white, train, and babble noises. Note that in practice, a voice activity detection (VAD) or adaptive noise estimation algorithms [13] could be employed to control λ.

IV. RESULTS AND DISCUSSION

Mean PESQ improvement scores for the three noise cases investigated in our experiments are shown in Fig. 5. These scores show that the proposed method performs best in the case of white noise, with lower improvements attained for the train and babble noise cases.

The results of spectrogram analysis are shown in Fig. 6. The enhanced signal for the white noise case does not exhibit speech distortion, while the background noise has been attenuated. In the train and babble noise cases, though the background noise is suppressed, a small amount of signal distortion is also introduced. This can be attributed to the use of a simple $\Lambda(k)$ function given in (6), which is constant across frequency k. Tuning the $\Lambda(k)$ function may improve the results for the nonwhite noise cases and will be investigated in the future.

We have also conducted informal listening experiments where the listeners were provided with the clean signal, the noisy signal, and the enhanced signal. For the white noise case in particular, we found the residual noise present in the enhanced speech to be *nondistracting* and *easy to ignore*.

In this letter, we give indicative performance of the proposed method against two popular speech enhancement techniques, namely, the spectral subtraction [1] and minimum mean squared error (MMSE) [2] methods, for the white noise case. The mean PESQ results are shown in Table I. The proposed method achieves results comparable to these two methods. A more detailed comparison with other techniques reported in the literature will be presented in future work.

V. CONCLUSION

In this letter, we presented a novel approach to speech enhancement. In the proposed method, the noisy short-time magnitude spectrum is recombined with a changed short-time phase spectrum to produce a modified short-time complex spectrum. During synthesis, the low energy components of the modified complex spectrum cancel out more than the high energy components, thus reducing background noise. Using an objective speech quality measure, informal subjective listening tests, as well as spectrogram analysis, we showed that the proposed method results in improved speech quality.

REFERENCES

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.
[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.
[3] N. Wiener, *The Extrapolation, Interpolation, and Smoothing of Stationary Time Series With Engineering Applications*. New York: Wiley, 1949.
[4] K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'87)*, Apr. 1987, vol. 12, pp. 297–300.
[5] Y. Ephraim and H. V. Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 251–266, Jul. 1995.
[6] J. Allen and L. Rabiner, "A unified approach to short-time Fourier analysis and synthesis," *Proc. IEEE*, vol. 65, no. 11, pp. 1558–1564, Nov. 1977.
[7] R. Crochiere, "A weighted overlap-add method of short-time Fourier analysis/synthesis," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 2, pp. 99–102, Apr. 1980.
[8] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 2, pp. 236–243, Apr. 1984.
[9] M. Portnoff, "Short-time Fourier analysis of sampled speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-29, no. 3, pp. 364–373, Jun. 1981.
[10] L. Alsteris and K. Paliwal, "Short-time phase spectrum in speech processing: A review and some experimental results," *Digit. Signal Process.*, vol. 17, pp. 578–616, May 2007.
[11] Y. Hu and P. Loizou, "Subjective comparison of speech enhancement algorithms," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'06)*, Toulouse, France, 2006, pp. 153–156.
[12] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual Evaluation of Speech Quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'01)*, Salt Lake City, UT, 2001, vol. 2, pp. 749–752.
[13] P. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL: CRC , 2007.