PAN ARCHIVES OF ACOUSTICS
Vol. 41, No. 3, pp. 579–590 (2016)

DOI: 10.1515/aoa-2016-0056

Wavelet Packet Transform based Speech Enhancement via Two-Dimensional SPP Estimator with Generalized Gamma Priors

Pengfei SUN, Jun QIN

Department of Electrical and Computer Engineering, Southern Illinois University Carbondale
1230 Lincoln Drive, Mail Code 6603 Carbondale, IL 62901, USA; e-mail: jqin@siu.edu

(*received January 30, 2016; accepted May 18, 2016*)

Despite various speech enhancement techniques have been developed for different applications, existing methods are limited in noisy environments with high ambient noise levels. Speech presence probability (SPP) estimation is a speech enhancement technique to reduce speech distortions, especially in low signal-to-noise ratios (SNRs) scenario. In this paper, we propose a new two-dimensional (2D) Teager-energy-operators (TEOs) improved SPP estimator for speech enhancement in time-frequency (T-F) domain. Wavelet packet transform (WPT) as a multiband decomposition technique is used to concentrate the energy distribution of speech components. A minimum mean-square error (MMSE) estimator is obtained based on the generalized gamma distribution speech model in WPT domain. In addition, the speech samples corrupted by environment and occupational noises (i.e., machine shop, factory and station) at different input SNRs are used to validate the proposed algorithm. Results suggest that the proposed method achieves a significant enhancement on perceptual quality, compared with four conventional speech enhancement algorithms (i.e., MMSE-84, MMSE-04, Wiener-96, and BTW).

Keywords: speech enhancement; speech presence probability; wavelet packet transform; two-dimensional Teager energy operator.

1. Introduction

Single-channel speech enhancement technique has been widely used for various applications, such as hearing aid devices, mobile communication, hand-free telephony, etc. However, for noisy environments with high ambient noise levels, the estimation of clean speech signals is still a great challenge with current speech enhancement methods (Martin, 2002). The high-level background noises are usually non-stationary and hard to be tracked. In addition, due to low signal to noise ratio (SNR), the estimated speech may be plagued by distortions and fluctuating with residual background noises.

Spectral estimation based on *a priori* knowledge of the probability distribution of speech and noise is a popular speech enhancement technique (Ephraim, Malah, 1984; Ephraim, Van Trees, 1995; Hu, Loizou, 2004; Park, *et al.*, 2015). This type of methods typically uses short time Fourier-transform (STFT) to obtain the spectrum within consecutive time windows of an input signal. Corresponding statistical models are developed based on optimal estimation techniques, such as minimum mean square error (MMSE) (Boll, 1979) and maximum *a posteriori* (MAP) (Hendriks, Gerkmann, Jensen, 2013). Since the spectral estimators are based on the conditional probability of that speech presents, speech presence probability (SPP) estimation can be helpful to reduce the music noise and enhance the perceptual quality of noisy speech (Fisher, Tabrikian, Dubnov, 2006; Gerkmann, Breithaupt, Martin, 2008), particularly avoiding the distortion of low SNRs speech components.

To achieve accurate estimation of SPP, different probabilistic latent components based models have been investigated (Cohen, Berdugo, 2001; Cohen, 2003). Most of these techniques are developed based on the statistical models of speech and noise signals (Cohen, 2004). Previous studies showed that Teager energy operator (TEO) was able to effectively detect speech (Kandia, Stylianou, 2006) in wavelet transform domain. Unlike those statistical methods that estimate the SPP (Loizou, 2013), TEO determines the

energy distribution of speech components in an analytic way, rather than relying on any prior knowledge of speech or noise (Bahoura, Rouat, 2006). It is considerably efficient for amplitude-modulated (AM) and frequency-modulated (FM) signal extraction (Kandia, Stylianou, 2006). Because human speech can be considered as a summation of modulated signals, TEO has been widely used in speech processing (Dunn, Quatieri, Kaiser, 1993; Bahoura, Rouat, 2001; 2006; Sanam, Shahnaz, 2013). Conventional TEO only detects speech transitions in time domain without providing the frequency distribution of speech components (Bovik, Maragos, Quatieri, 1993), and neglects the speech modulation structures. In this paper, two-dimensional (2D) TEO has been proposed to improve SPP estimator in wavelet packet transform (WPT) domain. WPT is an effective technique for multiband noise suppression (Bahoura, Rouat, 2001; Weickert, Benjaminsen, Kiencke, 2008). By applying WPT and 2D TEO, we can obtain the improved SPP estimator in the joint time-frequency (T-F) domain.

WPT based spectral estimation approaches have been developed based on the statistical models of speech and noise derived from STFT coefficients (Hu, Loizou, 2004; Ghanbari, Karami-Mollaei, 2006; Johnson, Yuan, Ren, 2007; Tasmaz, Ercelebi, 2008). Although these methods have obtained significant speech enhancement by introducing the STFT based statistical model directly, WPT coefficients with respect to speech demonstrate different probability distribution (Simoncelli, Adelson, 1996). The statistical models of speech in WPT domain have been developed to obtain accurate clean speech estimator. Several typical probability distributions, such as Gaussian, Gamma, Laplacian, and super Gaussian, have been applied to represent the spectral magnitudes of speech in STFT domain (Hendriks, Gerkmann, Jensen, 2013; Erkelens, *et al.*, 2007; Martin, 2005). Recent works reveal that the generalized gamma distribution model works better on describing speech distribution (Erkelens, *et al.*, 2007; Martin, 2005; Mohammadiha, Martin, Leijon, 2013). In this paper, considering that WPT coefficients of speech can be positive and negative values, a generalized two-side gamma distribution model is introduced to fit the WPT coefficients. The gamma distribution can be estimated from the clean speech in terms of different orders of moments (i.e., mean value, variance and kurtosis) in WPT domain. In addition, the WPT coefficients of noise are still assumed obeying Gaussian distribution.

In this paper, we propose a new algorithm, WPT-MTEO, for speech enhancement in high noisy environments. The proposed algorithm is based on the 2D TEO improved SPP estimator in the WPT domain. Two different forms of 2D TEOs are also compared with respect to accuracy of speech components detection in the T-F domain. Moreover, a MMSE estimator is obtained based on a generalized gamma prior distribution of speech. The speech samples corrupted by environmental and occupational noises (i.e., machine shop, factory and station) at different input SNRs are used to validate the proposed algorithm (Langner, Black, 2004). The performance of the proposed algorithm is compared with other four existing speech enhancement algorithms, including Wiener96 (Scalart, 1996), MMSE-84 (Ephraim, Malah, 1984), MMSE-04 (Cohen, 2004), and BTW (Chang, Yu, Vetterli, 2000).

2. Methods and materials

2.1. 2D TEO improved SPP estimator in WPT domain

TEO is useful on processing amplitude modulated (AM) or frequency modulated (FM) signals. For human speech, which can be regarded as a typical modulated signal, TEO has been used to extract energy distribution (Ying, Mitchell, Jamieson, 1993). In Ref. (Bovik, Maragos, Quatieri, 1993), TEO is proposed to obtain time-adaptive noise threshold for the extraction of the speech information based on WPT. TEO can efficiently emphasize periodic signals while depress the random signals. In this study, TEO is applied for speech components detection in the T-F domain. After applying WPT, the input noisy speech signal $y(t)$ can be described as

$$w_y(k,t) = WP_k * y(t), \qquad k = 1, ..., 2^j, \tag{1}$$

where j is the WPT level, decomposing the noisy signal $y(t)$ into 2^j bands corresponding to WPT coefficients $w_y(k,t)$. $*$ refers to convolution operation. WPT decomposes the signal into the T-F domain, and concentrates the formants' energy by its sparse representation. However, when SNR is low (e.g., SNR < -5 dB), the energy ratio between noise and speech formant decreases. TEO is introduced to detect the subtle differences, because it can efficiently extract the energy distribution of speech components. In this study, two types of 2D TEO are introduced to outline the distribution of speech components in the following sections.

2.1.1. Independent 2D TEO

The generalized form of 1D TEO can be described as

$$T(t,s) = w(t)^{2/s} - \left(w(t-t_0)w(t+t_0)\right)^{1/s}, \tag{2}$$

where $w(t)$ is the observation and $T(t,s)$ is the TEO kernel, reflecting the instantaneous energy of $w(t)$. t_0, as a constant window width of samples, can be called as the lag parameter (Kaiser, 1993). In this study, we use s as the parameter to adjust the local mean

value, as a result to control the energy contrast. Two types of 2D TEOs, independent and intersectional 2D TEOs, are proposed to develop the improved SPP estimator. For the independent 2D TEO, the time TEO kernel $T_k^1(t,s)$ and frequency TEO kernel $T_t^1(k,s)$ are independently obtained by

$$\begin{aligned} T_k^1(t,s) &= w(k,t)^{2/s} \\ &\quad - \left(w(k,t-\Delta t)w(k,t+\Delta t)\right)^{1/s}, \end{aligned} \tag{3}$$

$$\begin{aligned} T_t^1(k,s) &= w(k,t)^{2/s} \\ &\quad - \left(w(k-\Delta k,t)w(k+\Delta k,t)\right)^{1/s}, \end{aligned} \tag{4}$$

where $w(k,t)$ is the WPT coefficient. k and t are the frequency and time indexes, respectively. Therefore, Δk and Δt are corresponding frequency and time lag window parameters. The outline of the independent TEO can be obtained as

$$S_k(t,s) = \frac{\left|h(t) * T_k^1(t,s)\right|}{\max\left(|h(t) * T_k^1(t,s)|\right)}, \tag{5}$$

$$S_t(k,s) = \frac{\left|h_1(k) * T_t^1(k,s)\right|}{\max\left(|h_1(k) * T_t^1(k,s)|\right)}, \tag{6}$$

$$S^1(k,t,s) = S_k(t,s)S_t(k,s). \tag{7}$$

After applying low pass filters $h(t)$ and $h_1(k)$ to TEO kernels and normalization, $S_k(t,s)$ and $S_t(k,s)$ represent the energy outline of k-th WPT-band and the frequency distribution at time t, respectively. $S^1(k,t,s)$ refers to the independent 2D outline of the energy distribution of the independent TEO.

2.1.2. Intersectional 2D TEO

The intersectional 2D TEOs, with respect to horizontal-vertical direction and diagonal direction, are expressed as

$$\begin{aligned} TH\{w(k,t)\} &= \left\{\frac{\partial w}{\partial k}\right\}^2 + \left\{\frac{\partial w}{\partial t}\right\}^2 \\ &\quad - w\left\{\frac{\partial^2 w}{\partial k^2} + \frac{\partial^2 w}{\partial t^2}\right\}, \end{aligned} \tag{8}$$

$$\begin{aligned} TD\{w(k,t)\} &= 2\left\{\frac{\partial w}{\partial k}\right\}\left\{\frac{\partial w}{\partial t}\right\} \\ &\quad - w\left\{\frac{\partial^2 w}{\partial k \partial t} + \frac{\partial^2 w}{\partial t \partial k}\right\}, \end{aligned} \tag{9}$$

where $TH\{w(k,t)\}$ and $TD\{w(k,t)\}$ are horizontal-vertical and diagonal 2D TEO kernels. With a discrete form, a contrast parameter s incorporated nonlinear 2D version can be given by

$$\begin{aligned} T^{2,H}(k,t,s) &= 2w(k,t)^{2/s} \\ &\quad - \left(w(k-\Delta k,t)w(k+\Delta k,t)\right)^{1/s} \\ &\quad - \left(w(k,t-\Delta t)w(k,t+\Delta t)\right)^{1/s}, \end{aligned} \tag{10}$$

$$\begin{aligned} T^{2,D}(k,t,s) &= 2w(k,t)^{2/s} \\ &\quad - \left(w(k-\Delta k,t+\Delta t)w(k+\Delta k,t-\Delta t)\right)^{1/s} \\ &\quad - \left(w(k-\Delta k,t-\Delta t)w(k+\Delta k,t+\Delta t)\right)^{1/s}. \end{aligned} \tag{11}$$

Following the same procedures in (5)–(7), one can obtain the 2D outline of the energy distribution of the intersectional 2D TEO as

$$S^{2,1}(k,t,s) = \frac{\left|H(k,t) * T^{2,H}(k,t,s)\right|}{\max\left(\|H(k,t) * T^{2,H}(k,t,s)\|\right)}, \tag{12}$$

$$S^{2,2}(k,t,s) = \frac{\left|H(k,t) * T^{2,D}(k,t,s)\right|}{\max\left(\|H(k,t) * T^{2,D}(k,t,s)\|\right)}, \tag{13}$$

where 2D low pass filters $H(k,t)$ is applied to TEO kernel $T^2(k,t,s)$, '*' is convolution operation.

2.1.3. 2D TEO improved SPP estimator

Considering that TEO demonstrates higher energy density for harmonic signals and lower energy density for random noise, the energy density obtained by TEO is frequently applied to representing the existence of speech components or not. In this study, two outlines of energy distribution for two different TEOs after the normalization procedures as (5)–(7) and (12)–(13) can be applied as the SPP estimator, which is defined as

$$\mathrm{SPPT}(k,t,s) = S^i(k,t,s), \tag{14}$$

where i refers to the independent (type 1) or intersectional (type 2) 2D TEO. By introducing the proposed 2D TEOs to detect the speech components, SPP estimation can be obtained without prior knowledge of speech and noise signals. The proposed 2D TEO improved SPP estimator can be very sensitive to noise. To overcome this problem and obtain more accurate SPP estimation, two groups of lag window parameter (Δk, Δt) are used to derive the SPP values, which represent local SPP and global SPP, respectively. Therefore, a more robust SPP estimator is derived as

$$\begin{aligned} \mathrm{SPP}(k,t,s) &= \mathrm{SPPT}_l(k,t,\Delta k_1,\Delta t_1,s) \\ &\quad \cdot \mathrm{SPPT}_g(k,t,\Delta k_2,\Delta t_2,s), \end{aligned} \tag{15}$$

where SPPT_l refers to the local SPP. Δk_1, and Δt_1 are set as unit values, representing high window resolution. Comparatively, SPPT_g refers to the global SPP. Δk_2, and Δt_2 are selected as larger values, representing low window resolution but more smooth transition. In this study, due to the 64 subbands of WPT, Δk_2 is selected as 4, and Δt_2 is 8. In addition, the contrast parameter

s was chosen with different values: for SPPT_l, s is 1; for SPPT_g, s is 2.

WPT coefficients in T-F domain of the clean speech and the noisy speech are shown in Figs. 1a and 1b, respectively. Figures 1c and 1d illustrate the detected speech in the T-F domain by applying the proposed SPP estimators, improved by independent and intersectional 2D TEOs. One can see that the intersectional 2D TEO improved SPP estimator displayed a better detection result than the independent 2D TEO improved SPP estimator. Results indicate that the intersectional 2D TEO improved SPP estimator can more effectively suppressed the noise under low SNR scenarios (SNR < -5 dB). In this study, we focus on speech enhancement in high noise environments. Therefore, the intersectional 2D TEO is selected for the development of the proposed SPP estimator.

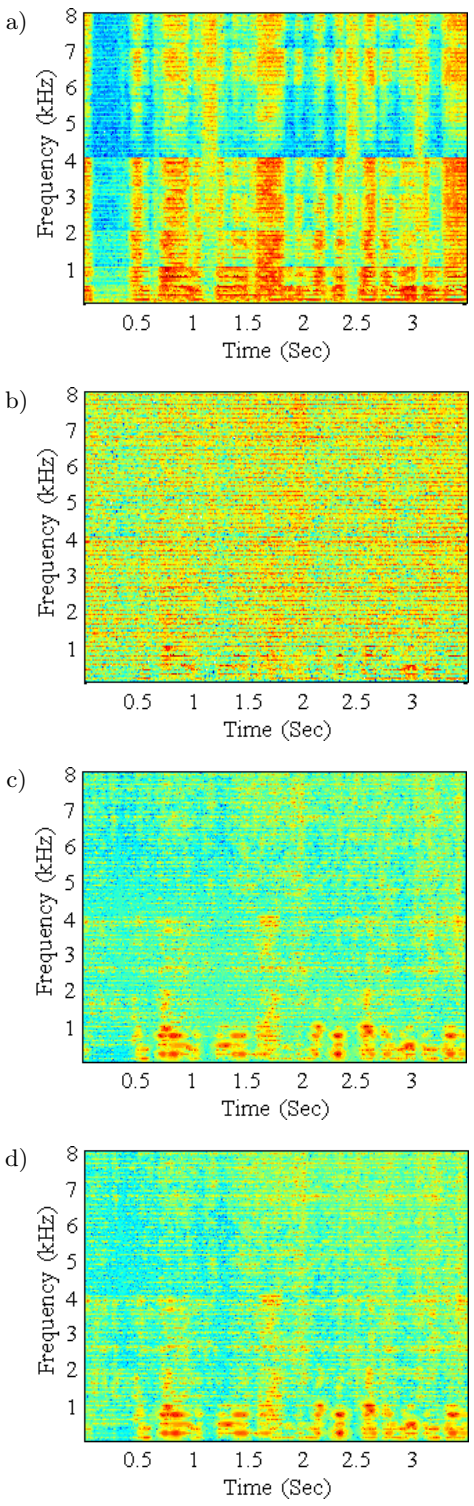

Fig. 1. The T-F distribution for: a) clean speech, b) noisy speech (SNR $= -5$ dB factory noise), and applied proposed SPP estimators improved by c) the independent and d) intersectional 2D TEOs.

2.2. Generalized speech model and clean speech estimator in WPT domain

Several statistical models, including Gamma, Laplacian and super Gaussian functions have been used to describe the probability density of speech in the STFT domain (Erkelens, *et al.*, 2007). In this study, noise signals in WPT domain are assumed to obey Gaussian distribution. The statistical model of speech signals in WPT domain has been obtained by introducing a two-side generalized Gamma model (Erkelens, *et al.*, 2007). This generalized Gamma model achieves high accuracy on predicting speech spectrum distribution, and accordingly can be defined as (Erkelens, *et al.*, 2007)

$$p(w) = \frac{\gamma\beta^{\nu}}{2\Gamma(\nu)}|w|^{\gamma\nu-1}\exp(-\beta|w|^{\gamma}), \qquad (16)$$

where $\Gamma(\cdot)$ is gamma function, β is scale parameter that also related with prior SNRs, and ν is shape parameter for the generalized Gamma function, and w represents WPT coefficient. Two-side form of gamma model is used because speech coefficients in WPT domain display a symmetrical probability distribution in $[-\infty\ 0]$ and $[0\ +\infty]$.

2.2.1. Optimization of parameters of the generalized speech model

In (16), three parameters (i.e., γ, β, and ν) significantly affect the shape of probability distribution with respect to the WPT coefficients. γ is usually chosen to be 1 or 2. β and ν are estimated based on input speech samples, and relationships among the three parameters can be found in (Erkelens, *et al.*, 2007). In terms of different γ values, the other two shape parameters can be estimated in WPT domain. When $\gamma = 1$, the parameters β and ν can be obtained by solving (17)

$$\frac{1}{\beta}\frac{\Gamma(\nu+1)}{\Gamma(\nu)} = \overline{w}_{\gamma=1}\,, \qquad \frac{\nu(\nu+1)}{\beta^2} = \sigma^2_{\gamma=1}\,, \qquad (17)$$

where σ^2 is the speech spectral variance, and $\overline{w}$ is the mean value of speech coefficients. When $\gamma = 2$, there is no explicit solution (close form) for ν based on first

and second moment. Thus, kurtosis K as a high order moment parameter is introduced to estimate ν:

$$K = \frac{\mu_4}{\mu_2} = \frac{\int_0^\infty w_{k,t}^4 p(w_{k,t})\,\mathrm{d}w_{k,t}}{\int_0^\infty w_{k,t}^2 p(w_{k,t})\,\mathrm{d}w_{k,t}}, \tag{18}$$

where $p(w_{k,t})$ refers to the probability of speech coefficients in (16). Then β and ν can be derived through (19)

$$\frac{\nu+1}{\nu} = K_{\gamma=2}\,, \qquad \frac{\nu}{\beta} = \sigma^2_{\gamma=2}\,. \tag{19}$$

One arbitrarily selected speech sample is used to subjectively evaluate the parameter γ. As shown in Fig. 2, the histogram of the WPT coefficients of clean speech sample in the second subband $w_{2,t}$ is compared with the estimated statistical models when $\gamma = 1$ and $\gamma = 2$, respectively. $p(w)$ is the normalized histogram value. The parameters for each statistical model are obtained according to (17) and (19). It can be found that the model with $\gamma = 1$ in (17) shows a better fitting on the histogram of WPT coefficients than that with $\gamma = 2$ in (19).

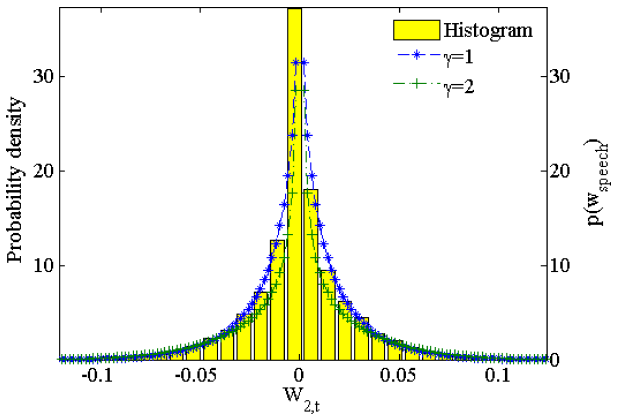

Fig. 2. The histogram of the-second-subband WPT coefficients of clean speech (bar), and the speech probability distributions in terms of the model in (10) when $\gamma = 1$ and $\gamma = 2$.

To generally compare the models with parameter $\gamma = 1$ and $\gamma = 2$, 30 speech samples from CMU database (Langner, Black, 2004) are used to compute the normalized fitting errors in 64 subbands. In each subband, the lowest normalized fitting error of different models for each speech sample is selected. The mean values and standard deviations of these lowest normalized fitting errors are calculated as well. As shown in Fig. 3, in each subband, the model in (16) with $\gamma = 1$ shows lower minimal normalized fitting errors than speech model with $\gamma = 2$ at all subbands.

Moreover, the ν value is also optimized. Instead of estimating from the WPT coefficients, ν is incremen-

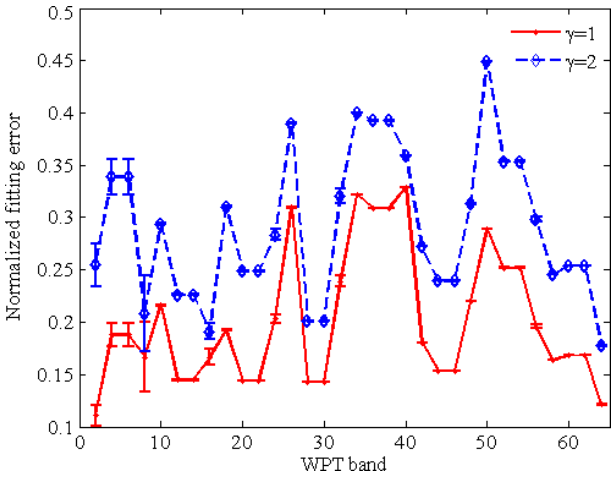

Fig. 3. The mean value and standard deviation values for the minimal normalized fitting errors of speech corpus in each WPT band. The statistical models are fitting to the WPT coefficients of speech corpus in each subband with respect to $\gamma = 1$ and $\gamma = 2$.

tally selected in the range [0, 2], and β is still estimated according to (17) and (19). Normalized fitting error, defined as $\|p(w_{k,t}) - h(w_{k,t})\|$, is used to evaluate how

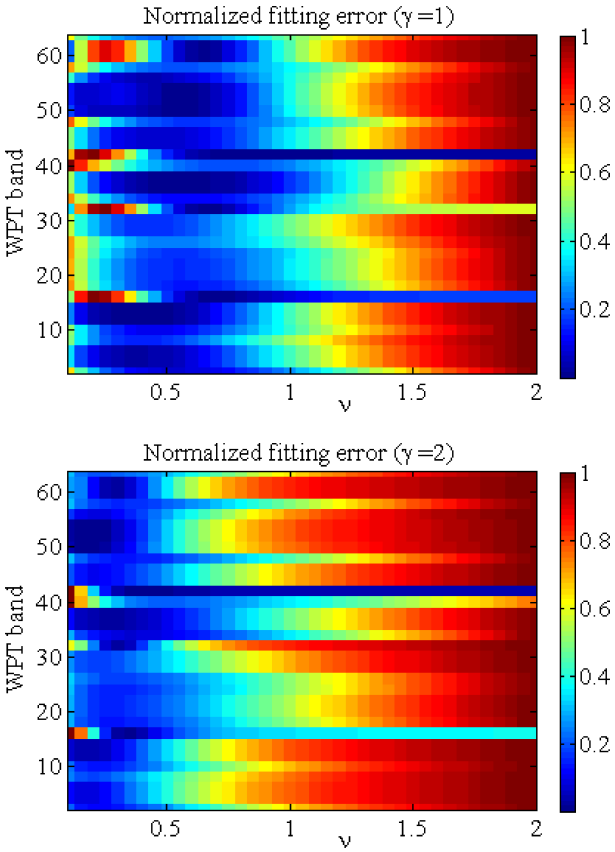

Fig. 4. The distribution of normalized fitting error for speech statistical models with different values in each WPT band with respect to $\gamma = 1$ and $\gamma = 2$. The color bar on the right show that bottom color represents small error values and the top color represents large error values.

well each statistical model explains the distribution of WPT coefficients. Here 6-levels WPT decomposes the signal into 64 subbands, in which the normalized fitting error between the estimated probability $p(w_{k,t})$ and the histogram $h(w_{k,t})$ is calculated when ν is changing. Figure 4 reveals that for $\gamma = 1$, the lowest fitting errors are achieved when ν is in the range [0.4, 0.6]; for $\gamma = 2$, the lowest fitting errors are achieved when ν is in the range [0.1, 0.3]. Therefore, $\gamma = 1$ and $\nu = 0.4$ are selected as the speech statistical model parameters in WPT domain in this study.

2.2.2. MMSE clean speech estimator

Based on the estimated generalized speech model in WPT domain, a clean speech estimator can be derived (Erkelens, *et al.*, 2007). Considering a signal model with the form

$$w_y(k,t) = w_x(k,t) + w_r(k,t), \tag{20}$$

where $w_y(k,t)$, $w_x(k,t)$ and $w_r(k,t)$ are WPT coefficients in k-th subband at time t obtained from the noisy speech, clean speech, and noise, respectively. Assuming that $w_x(k,t)$ and $w_r(k,t)$ are statistically independent across time and frequency, X and Y are used to represent the coefficients, then the following MMSE estimator can be obtained:

$$\begin{aligned} E(X|Y) &= \frac{\int\limits_{-\infty}^{+\infty} X p(Y|X)p(X)\,\mathrm{d}X}{\int\limits_{-\infty}^{+\infty} p(Y|X)p(X)\,\mathrm{d}X} \\ &= \frac{\int\limits_{-\infty}^{+\infty} X p_r(Y-X)p_x(X)\,\mathrm{d}X}{\int\limits_{-\infty}^{+\infty} p_r(Y-X)p_x(X)\,\mathrm{d}X}, \end{aligned} \tag{21}$$

where $p_x(X)$ obeys the generalized gamma distribution in (16), and $p_r(Y-X)$ obeys the Gaussian distribution.

When $\gamma = 1$, the estimator is defined as (Erkelens, *et al.*, 2007):

$$\begin{aligned} E(X|Y) = \sigma_r \nu \Bigg[&\exp\left(\frac{1}{4}Y_-^2\right) D_{-(\nu+1)}(Y_-) \\ &- \exp\left(\frac{1}{4}Y_+^2\right) D_{-(\nu+1)}(Y_+) \Bigg] \Big/ \Bigg[\exp\left(\frac{1}{4}Y_-^2\right) D_{-\nu}(Y_-) \\ &+ \exp\left(\frac{1}{4}Y_+^2\right) D_{-\nu}(Y_+) \Bigg], \end{aligned} \tag{22}$$

where $D_{-\nu}(\cdot)$ is a special function, called as the parabolic cylinder function of order ν, and

$$Y_\pm = \beta\sigma_r \pm \frac{Y}{\sigma_r}, \tag{23}$$

σ_r is the estimated variance of noise. For $\nu = 0.4$ in this study, β can be calculated by (17), where the priori SNR is estimated by the Decision-Directed approach (Ephraim, Malah, 1984).

2.3. Implementation

As shown in Fig. 5, in the proposed algorithm, WPT was initially applied to noisy speech, and based on the WPT coefficients the intersectional 2D TEO was obtained to yield the 2D SPP estimator. In parallel, the WPT coefficients of clean speech samples were used to develop the pre-learned statistical model. Second, both the pre-learned speech model and SPP were fed into the MMSE estimator to estimate the clean speech from noisy speech. Finally, the estimated clean speech components in T-F domain were transformed by inverse WPT to obtain the enhanced speech.

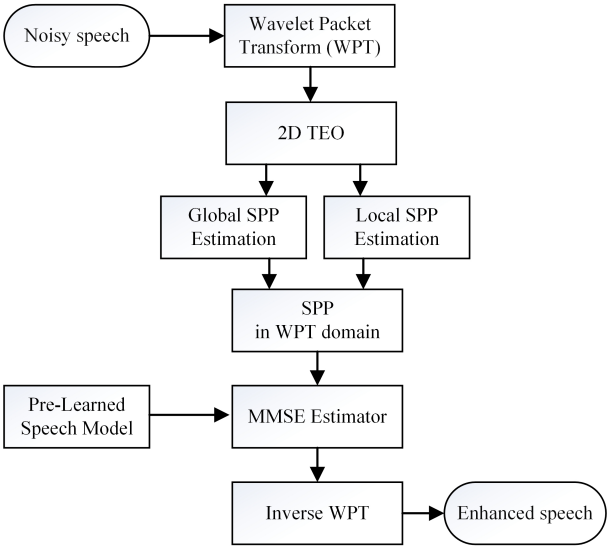

Fig. 5. The flow chart of implementation of the proposed algorithm.

3. Results and evaluation

In our study, the proposed algorithm is employed in a speech enhancement framework. The noisy speech signals were synthesized by adding different background noise samples to randomly selected speech samples at different input SNRs. The background noise signals were selected from industrial noise database (AudioMiCro, 2015) and environmental noise database (Hu, Loizou, 2007), including machine, factory, and station. 30 adult Enginsh speech samples were randomly selected from CMU database (Langner, Black, 2004). The noisy speech signals were synthesized with 16 kHz sampling rate and at various input SNRs from −10 dB to 10 dB. Moreover, the performance of the proposed WPT-MTEO algorithm was compared with four speech enhancement algorithms, including MMSE-84, MMSE-04, Bayesian estimation based thresholding and the improved Wiener filter. MMSE-04 (Cohen, 2004) and MMSE-84 (Ephraim, Malah, 1984) are compared in terms of the amplitude estimation approach in the STFT domain (Ephraim, Malah, 1984). Bayesian thresholding is one typical

algorithm for Bayesian estimation in wavelet domain (Chang, Yu, Vetterli, 2000). Wiener-96 filter is a very classical algorithm for speech enhancement in many applications (Scalart, 1996).

3.1. Algorithm assessment based on PESQ and SegSNR

Two objective metrics, perceptual evaluation of speech quality (PESQ) and Segmental SNRs (SegSNR) as implemented in (Hu, Loizou, 2007), are used to quantitatively evaluate the performance of the speech enhancement algorithms in this study. PESQ is originally developed for assessing perceived quality of coded speech. It demonstrates high correlation with speech quality in the speech enhancement context. The maximum PESQ and improved SegSNR for five algorithms are summarized in Table 1. At all input SNRs (-10 dB $<$ SNRs $<$ 10 dB), the proposed algorithm shows the best performance compared with other four algorithms. Specifically, at low SNRs (-5 dB and -10 dB), the proposed WPT-MTEO algorithm achieves remarkable higher PESQ than the other four algorithms as well as obtains highest SNR improvement for all three background noises. Results indicate that the proposed algorithm has the capability to enhance the speech quality in high noise environment (low SNRs).

Figure 6 shows the averaged improvements of PESQ and SegSNR of noisy speech by applying five

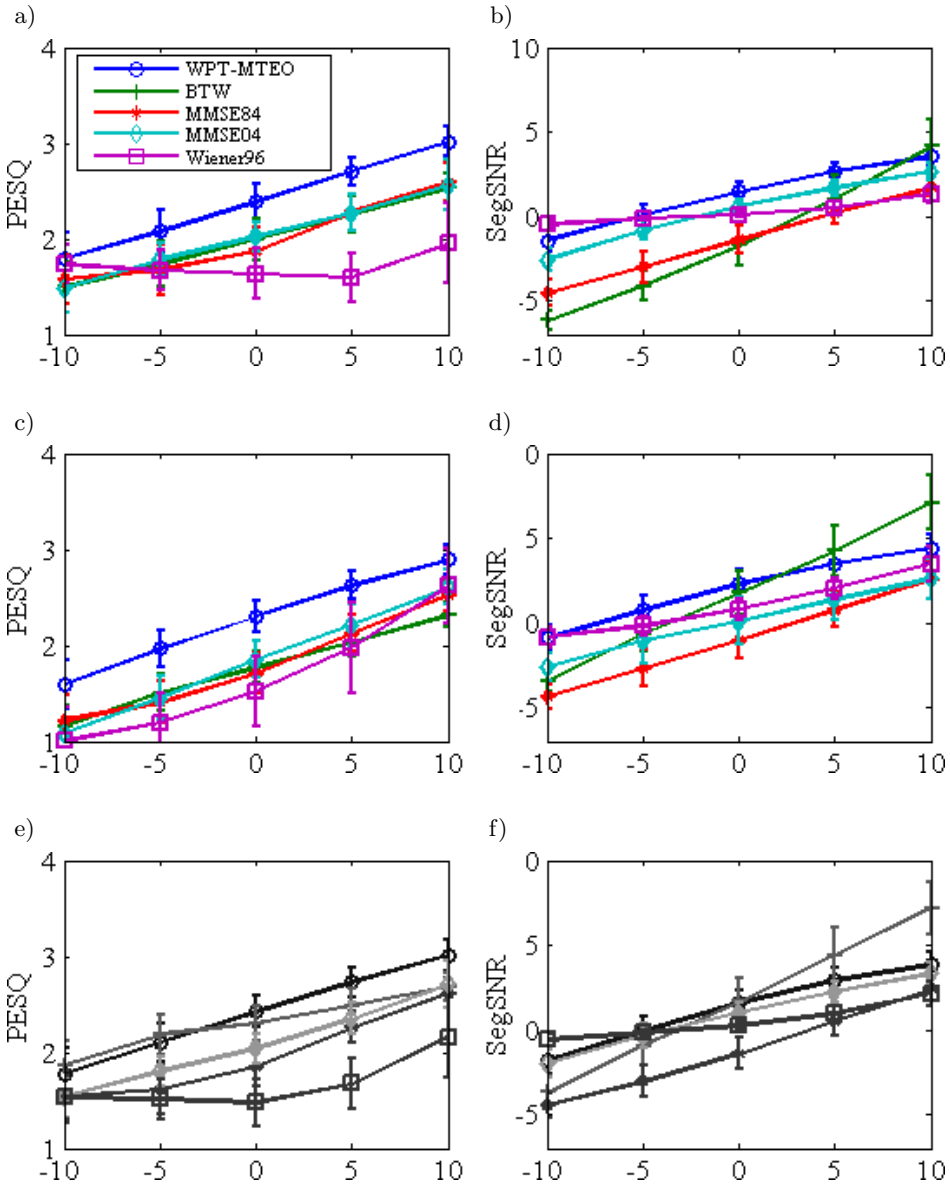

Fig. 6. Averaged PESQ scores and SegSNRs with standard deviations obtained from 30 speech corpus corrupted by different noises (i.e., factory noise in (a), (b), machine shop noise in (c), (d), and station noise in (e), (f)) for five algorithms at five input SNR levels (i.e., [-10 dB 10 dB]).

Table 1. The maximum PESQ and improved SegSNR obtained by applying the proposed WPT-MTEO and other four existing algorithms for three different background noises at various input SNRs.

		−10 dB		−5 dB		0 dB		5 dB		10 dB	
		Δ_{SNR}	PESQ	Δ_{SNR}	PESQ	Δ_{SNR}	PESQ	Δ_{SNR}	PESQ	Δ_{SNR}	PESQ
Machine Shop	Wiener	8.70	1.90	7.52	1.82	6.45	2.29	5.35	2.84	2.63	3.16
	BTW	6.43	1.66	7.33	1.93	5.66	2.13	3.80	2.37	1.65	2.72
	MMSE-04	7.36	1.67	7.07	1.89	5.83	2.18	4.03	2.76	1.25	2.99
	MMSE-84	5.55	1.92	5.54	1.90	4.46	2.07	3.29	2.58	1.05	2.88
	WPT-MTEO	**8.77**	**2.04**	**7.65**	**2.12**	**7.35**	**2.39**	**6.56**	**2.91**	**5.58**	**3.21**
Factory	Wiener	8.68	1.91	7.57	1.97	6.40	1.83	5.11	2.05	5.08	3.26
	BTW	7.28	2.29	7.77	2.51	6.36	2.63	4.94	2.83	3.15	3.09
	MMSE-04	8.31	2.03	7.50	2.13	6.77	2.35	5.69	2.68	4.61	3.24
	MMSE-84	5.93	2.09	5.58	2.20	4.44	2.32	4.12	2.56	3.25	3.07
	WPT-MTEO	**8.76**	**2.30**	**8.21**	**2.56**	**7.33**	**2.75**	**6.08**	**2.97**	**5.84**	**3.35**
Station	Wiener	8.45	1.92	7.56	1.96	5.61	2.05	4.85	2.32	4.79	3.20
	BTW	6.19	2.00	4.18	2.30	4.67	2.46	4.53	2.79	2.77	3.03
	MMSE-04	8.37	2.11	7.71	2.31	5.67	2.53	5.34	2.82	4.22	3.19
	MMSE-84	5.56	2.16	5.30	2.22	4.45	2.43	4.01	2.71	3.34	3.22
	WPT-MTEO	**8.54**	**2.39**	**7.78**	**2.56**	**6.01**	**2.77**	**5.46**	**2.95**	**5.64**	**3.25**

speech enhancement algorithms for three different types of background noises at various SNRs (-10 dB < SNRs < 10 dB). As shown in Figs. 6(a), (c), and (e), the proposed WPT-MTEO algorithm demonstrates significant enhancement on PESQ, compared with other four algorithms. In Fig. 6(b), (d) and (f), the SegSNR improvement results show that our developed algorithm is comparable with other four algorithms.

3.2. Algorithm assessment based on three composite objective measures

In this study, three composite objective measures have been used to evaluate the performance of our developed speech enhancement algorithm (WPT-MTEO). These three composite objective measures are introduced to predict the quality of noisy speech enhanced by noise suppression algorithms (Hu, Loizou,

Table 2. The maximum C_{sig}, C_{bak} and C_{ovl} for Wiener, BTW, MMSE84, MMSE04, and proposed WPT-TEO at 30 speech samples.

		−10 dB			−5 dB			0 dB			5 dB			10 dB		
		C_{sig}	C_{bak}	C_{ovl}	C_{sig}	C_{bak}	C_{ovl}	C_{sig}	C_{bak}	C_{ovl}	C_{sig}	C_{bak}	C_{ovl}	C_{sig}	C_{bak}	C_{ovl}
Machine Shop	Wiener	−0.05	1.90	0.21	−0.24	1.94	0.36	−0.33	1.96	0.62	−0.50	2.12	0.98	0.53	2.93	2.41
	BTW	0.56	1.80	0.30	0.87	2.03	0.83	1.08	2.30	1.43	1.05	2.65	1.94	0.93	2.96	2.53
	MMSE-04	0.68	1.61	0.29	0.92	1.87	0.80	0.99	2.17	1.37	1.08	2.57	2.14	1.23	2.84	2.69
	MMSE-84	0.34	1.93	0.44	0.43	2.12	0.83	0.51	2.35	1.31	0.67	2.66	1.98	0.74	2.93	2.56
	WPT-MTEO	**1.06**	**1.99**	**1.23**	**1.39**	**2.20**	**1.53**	**1.58**	**2.38**	**1.84**	**1.79**	**2.68**	**2.22**	**1.87**	**3.02**	**2.75**
Factory	Wiener	−0.17	1.81	−0.21	−0.36	1.85	−0.07	−0.45	1.93	0.35	−0.13	2.21	1.03	0.68	3.07	2.43
	BTW	0.76	1.39	0.25	0.93	1.99	0.66	1.11	2.21	1.17	0.99	2.43	1.80	0.90	2.68	2.49
	MMSE-04	0.79	1.59	0.12	0.99	1.91	0.65	0.97	2.26	1.29	1.18	2.64	2.08	1.21	2.91	2.62
	MMSE-84	0.44	1.92	0.31	0.53	2.08	0.61	0.53	2.31	1.10	0.80	2.67	1.83	0.85	2.93	2.45
	WPT-MTEO	**1.23**	**1.97**	**1.43**	**1.94**	**2.18**	**2.14**	**2.73**	**2.51**	**2.25**	**2.81**	**3.05**	**2.51**	**3.22**	**3.35**	**2.65**
Station	Wiener	0.05	1.80	−0.36	0.25	1.81	0.14	0.56	2.30	0.97	0.69	2.64	1.70	1.10	2.98	2.46
	BTW	0.88	1.66	−0.06	1.08	1.88	0.52	1.17	2.14	1.17	1.17	2.42	1.79	1.03	2.64	2.43
	MMSE-04	0.86	1.56	−0.24	0.93	1.81	0.44	1.00	2.18	1.27	1.03	2.49	1.90	0.91	2.89	2.42
	MMSE-84	0.52	1.88	0.00	0.62	2.00	0.35	0.66	2.19	1.11	0.71	2.54	1.84	0.91	2.96	2.45
	WPT-MTEO	**1.69**	**1.92**	**1.16**	**2.17**	**2.16**	**1.63**	**2.55**	**2.45**	**1.95**	**2.57**	**2.78**	**2.28**	**2.71**	**3.13**	**2.63**

2007). They can be described as follows: (a) C_{sig} is the measurement of signal distortion (SIG), which is a linear combination of log-likelihood ratio (LLR), PESQ, and weighted–slope spectral distance (WSS); (b) C_{bak} is the measurement of noise distortion (BAK), which linearly combines the SegSNR, PESQ, and WSS; and (c) C_{ovl} is defined as the overall quality, and it is formed by linearly combining PESQ, LLR, and WSS (Ying, *et al.*, 1993).

Figure 7 shows the improvements of three objective measures by five speech enhancement algorithms. The proposed WPT-MTEO algorithm shows the highest improvements for all three metrics (C_{sig}, C_{bak}, and C_{ovl}). Compared to other four algorithms, the WPT-METO algorithm gains averagely about 0.3 higher point on noise distortion measure C_{bak}, and it is averagely about 1 higher point on signal distortion measure C_{sig}. For the overall speech enhancement quality measure C_{ovl}, the WPT-MTEO algorithm also obtains the best performance. At low SNRs one can found that the WPT-MTEO algorithm obtains significant improvements over all three metrics. Specifically, the WPT-MTEO algorithm demonstrates remarkable improvements on C_{sig} and C_{ovl} at low SNRs. It indicates that the proposed algorithm can not only enhance speech in high noise environments, but also can keep high quality of enhanced speech.

Moreover, the maximum values of C_{sig}, C_{bak}, and C_{ovl}, obtained by applying five speech enhancement algorithms are summarized in Table 2. Same as the results shown in Fig. 7, the WPT-MTEO algorithm achieves advantages over the other four algorithms.

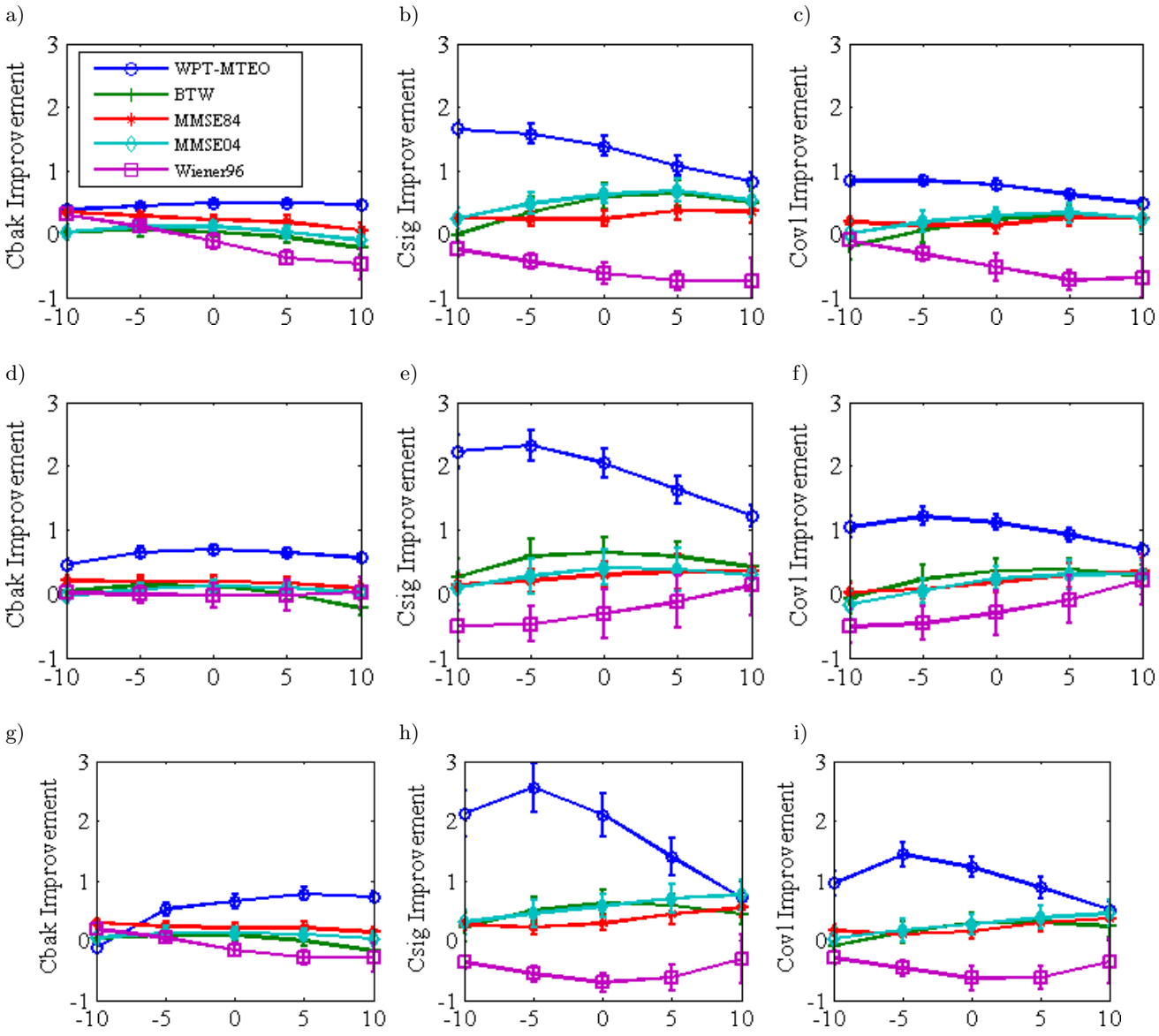

Fig. 7. Improvements of C_{bak}, C_{sig}, and C_{ovl} obtained from 30 noisy speech signals with three different background noises (i.e., factory noise in (a), (b), (c), machine shop noise in (d), (e), (f), and station noise in (g), (h), (i)) applied five speech enhancement algorithms at various input SNRs (-10 dB $<$ SNRs $<$ 10 dB).

With all three background noises, the WPT-MTEO algorithm demonstrates the highest improvements of three metrics among five speech enhancement algorithms.

In addition, Fig. 8 shows the spectrograms of clean speech, noisy speech (SNR = −5 dB) with factory background noise, and the enhanced speech by applying five speech enhancement algorithms,

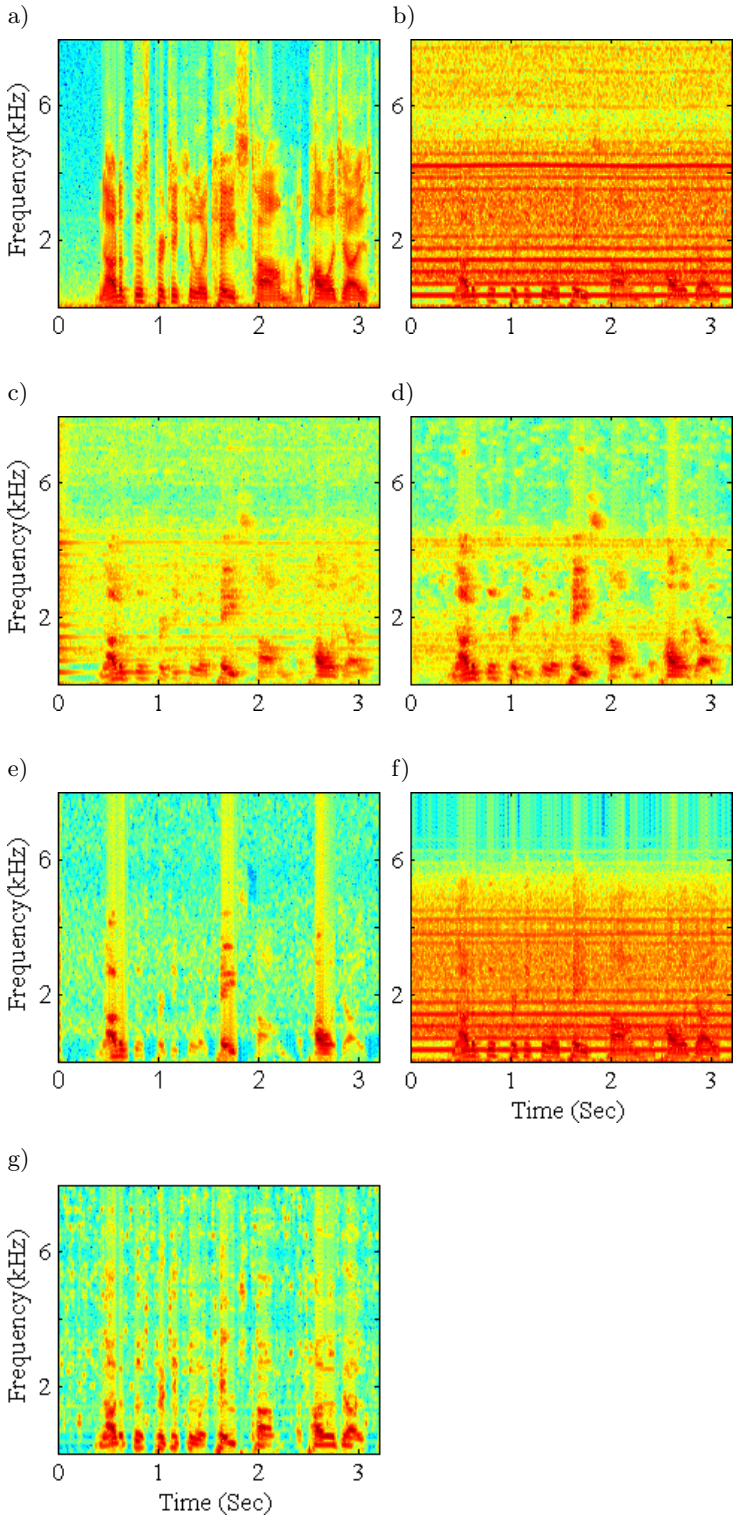

Fig. 8. Spectrums of (a) clean speech, (b) noisy speech with factory background noise (SNR = −5 dB), and enhanced speech by applying five algorithms, including (c) MMSE84, (d) MMSE04, (e) Wiener96, (f) BTW, and (g) WPT-MTEO, respectively.

respectively. It can be subjectively found that the proposed WPT-MTEO algorithm (as show in Fig. 8g) achieves high noise cancellation whereas retains high quality of enhanced speech. In contrast, three algorithms: MMSE84, MMSE04, and BTW, (as shown in Figs. 8c, 8d and 8f, respectively) cannot effectively eliminate the noise components in the frequency range around 1.8 kHz – 4.5 kHz. As shown in Fig. 8e, another algorithm (Wiener96) suppresses noise components but also significantly distorts speech components. Results suggest that the proposed algorithm is able to successfully separate speech from high-level industrial noise, and can achieve high quality of enhanced speech.

4. Conclusions

In this paper, we have developed a new algorithm, WPT-MTEO, for speech enhancement in high noise environments. The WPT-MTEO combines a 2D TEO improved SPP estimator in WPT domain with a MMSE estimator based on a generalized gamma prior of speech. Two different types of 2D TEOs, independent and intersectional 2D TEOs, have been introduced for the development of the energy-density based SPP estimator. By utilizing the statistic characteristics of speech samples, parameters of the generalized speech model in WPT domain are optimized. The corresponding MMSE amplitude estimator is applied as well. Selected speech samples corrupted with different types of background noises (i.e., machine shop, factory, and station) at various SNRs, are used to validate our developed algorithm. The performance of the developed algorithm is compared with other four existing speech enhancement algorithms, including Wiener96 (Scalart, 1996), MMSE-84 (Ephraim, Malah, 1984), MMSE-04 (Cohen, 2004), and BTW (Chang, Yu, Vetterli, 2000). Results show that our developed algorithm achieves remarkable improvements on speech perceptional quality improvement with respect to various metrics. Particularly, the performance at low SNR is in great advantage, compared with four other existing algorithms. It indicates that the proposed algorithm can successfully enhance speech at low SNRs with high quality of enhanced speech. The proposed algorithm is promising for speech enhancement applications in high noise environments.

References

1. AudioMiCro, *Free Industrial and Machinery Sound Effects*, Retrived November 29th, 2015, from http://www.audiomicro.com/free-sound-effects/free-industrial-and-machinery/.
2. Bahoura M., Rouat J. (2006), *Wavelet speech enhancement based on time-scale adaptation*, Speech Communication, **48**, 12, 1620–1637.
3. Bahoura M., Rouat J. (2001), *Wavelet speech enhancement based on the teager energy operator*, Signal Processing Letters, IEEE, **8**, 1, 10–12.
4. Boll S.F. (1979), *Suppression of acoustic noise in speech using spectral subtraction*, Acoustics, Speech and Signal Processing, IEEE Transactions on, **27**, 2, 113–120.
5. Bovik A., Maragos C.P., Quatieri T.F. (1993), *Am-fm energy detection and separation in noise using multiband energy operators*, Signal Processing, IEEE Transactions on, **41**, 12, 3245–3265.
6. Chang S.G., Yu B., Vetterli M. (2000), *Adaptive wavelet thresholding for image denoising and compression*, Image Processing, IEEE Transactions on, **9**, 9, 1532–1546.
7. Cohen I., Berdugo B. (2001), *Speech enhancement for non-stationary noise environments*, Signal processing, **81**, 11, 2403–2418.
8. Cohen I. (2003), *Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging*, Speech and Audio Processing, IEEE Transactions on, **11**, 5, 466–475.
9. Cohen I. (2004), *Speech enhancement using a noncausal a priori snr estimator*, Signal Processing Letters, IEEE, **11**, 9, 725–728.
10. Dunn R.B., Quatieri T.F., Kaiser J.F. (1993), *Detection of transient signals using the energy operator*, Acoustics, Speech, and Signal Processing, ICASSP., 1993 IEEE International Conference on, pp. 145–148.
11. Ephraim Y., Malah D. (1984), *Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator*, Acoustics, Speech and Signal Processing, IEEE Transactions on, **32**, 6, 1109–1121.
12. Ephraim Y., Van Trees H.L. (1995), *A signal subspace approach for speech enhancement*, Acoustics, Speech and Signal Processing, IEEE Transactions on, **3**, 4, 251–266.
13. Erkelens J.S., Hendriks R.C., Heusdens R., Jensen J. (2007), *Minimum mean-square error estimation of discrete fourier coeficients with generalized gamma priors*, Audio, Speech, and Language Processing, IEEE Transactions on, **15**, 6, 1741–1752.
14. Fisher E., Tabrikian J., Dubnov S. (2006), *Generalized likelihood ratio test for voiced-unvoiced decision in noisy speech using the harmonic model*, Audio, Speech, and Language Processing, IEEE Transactions on, **14**, 2, 502–510.
15. Gerkmann T., Breithaupt C., Martin R. (2008), *Improved a posteriori speech presence probability estimation based on a likelihood ratio with fixed priors*, Audio, Speech, and Language Processing, IEEE Transactions on, **16**, 5, 910–919.
16. Ghanbari Y., Karami-Mollaei M.R. (2006), *A new approach for speech enhancement based on the adaptive thresholding of the wavelet packets*, Speech communication, **48**, 8, 927–940.
17. Hendriks R.C., Gerkmann T., Jensen J. (2013), *Dft-domain based single-microphone noise reduction*

for speech enhancement: a survey of the state of the art, Synthesis Lectures on Speech and Audio Processing, **9**, 1, 80–84.

18. Hu Y., Loizou P.C. (2004), *Speech enhancement based on wavelet thresholding the multitaper spectrum*, Speech and Audio Processing, IEEE Transactions on, **12**, 1, 59–67.
19. Hu Y., Loizou P.C. (2007), *Subjective comparison and evaluation of speech enhancement algorithms*, Speech communication, **49**, 7, 588–601.
20. Johnson M.T., Yuan X., Ren Y. (2007), *Speech signal enhancement through adaptive wavelet thresholding*, Speech Communication, **49**, 2, 123–133.
21. Kaiser J.F. (1993), *Some useful properties of teager's energy operators*, Acoustics, Speech, and Signal Processing, ICASSP-93, IEEE International Conference on, pp. 149–152.
22. Kandia V., Stylianou Y. (2006), *Detection of sperm whale clicks based on the teager-kaiser energy operator*, Applied Acoustics, **67**, 11, 1144–1163.
23. Langner B., Black A.W. (2004), *Creating a database of speech in noise for unit selection synthesis*, Fifth ISCA Workshop on Speech Synthesis, 229–230.
24. Loizou P.C. (20130, *Speech enhancement: theory and practice*, CRC press.
25. Martin R. (2002), *Speech enhancement using mmse short time spectral estimation with gamma distributed speech priors*, Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference, pp. 253–256.
26. Martin R. (2005), *Speech enhancement based on minimum mean-square error estimation and supergaussian priors*, Speech and Audio Processing, IEEE Transactions on, **13**, 5, 845–856.
27. Mohammadiha N., Martin R., Leijon A. (2013), *Spectral domain speech enhancement using hmm state-dependent super-gaussian priors*, Signal Processing Letters, IEEE, **20**, 3, 253–256.
28. Park J., Kim J.-W., Chang J.-H., Jin Y. G., Kim N.S. (2015), *Estimation of speech absence uncertainty based on multiple linear regression analysis for speech enhancement*, Applied Acoustics, **87**, 2015, 205–211.
29. Sanam T.F., Shahnaz C. (2013), *Noisy speech enhancement based on an adaptive threshold and a modified hard thresholding function in wavelet packet domain*, Digital Signal Processing, **23**, 3, 941–951.
30. Scalart P. (1996), *Speech enhancement based on a priori signal to noise estimation*, Acoustics, Speech, and Signal Processing, ICASSP Conference Proceedings, IEEE International Conference on, pp. 629–632.
31. Simoncelli E.P., Adelson E.H. (1996), *Noise removal via bayesian wavelet coring*, Image Processing Proceedings., International Conference on, pp. 379–382.
32. Tasmaz H., Ercelebi E. (2008), *Speech enhancement based on undecimated wavelet packet-perceptual fiterbanks and mmse-stsa estimation in various noise environments*, Digital Signal Processing, **18**, 5, 797–812.
33. Weickert T., Benjaminsen C., Kiencke U. (2008), *Analytic complex wavelet packets for speech enhancement*, Acoustics, Speech and Signal Processing, ICASSP 2008. IEEE International Conference, pp. 3269–3272.
34. Ying G., Mitchell C., Jamieson L. (1993), *Endpoint detection of isolated utterances based on a modified teager energy measurement*, Acoustics, Speech, and Signal Processing, ICASSP-93, IEEE International Conference on, pp. 732–735.