Speech enhancement using the modified phase-opponency model

Om D. Deshmukh[a] and Carol Y. Espy-Wilson
Department of Electrical and Computer Engineering and Institute for Systems Research, University of Maryland, College Park, Maryland 20742

Laurel H. Carney
Department of Biomedical and Chemical Engineering and Institute for Sensory Research, Syracuse University, Syracuse, New York 13244

(Received 1 February 2006; revised 11 February 2007; accepted 13 February 2007)

In this paper we present a model called the Modified Phase-Opponency (MPO) model for single-channel speech enhancement when the speech is corrupted by additive noise. The MPO model is based on the auditory PO model, proposed for detection of tones in noise. The PO model includes a physiologically realistic mechanism for processing the information in neural discharge times and exploits the frequency-dependent phase properties of the tuned filters in the auditory periphery by using a cross-auditory-nerve-fiber coincidence detection for extracting temporal cues. The MPO model alters the components of the PO model such that the basic functionality of the PO model is maintained but the properties of the model can be analyzed and modified independently. The MPO-based speech enhancement scheme does not need to estimate the noise characteristics nor does it assume that the noise satisfies any statistical model. The MPO technique leads to the lowest value of the LPC-based objective measures and the highest value of the perceptual evaluation of speech quality measure compared to other methods when the speech signals are corrupted by fluctuating noise. Combining the MPO speech enhancement technique with our aperiodicity, periodicity, and pitch detector further improves its performance. © *2007 Acoustical Society of America.* [DOI: 10.1121/1.2714913]

PACS number(s): 43.72.Ne [DOS] Pages: 3886–3898

I. INTRODUCTION

Speech signals in real-world scenarios are often corrupted by various additive noise types (e.g., computer fan noise, subway noise, car noise, and babble), convolutive noise types (e.g., change in microphone or telephone-band-limited speech), and nonlinear disturbances. Speech enhancement techniques that can attenuate the interfering noise with minimal distortions to the speech signal can be used in various speech communication applications like automatic speech recognition, hearing aids, car and mobile phones, cockpits, and multiparty conferencing devices.

The problem of speech enhancement has received a tremendous amount of research attention over the past several decades. A thorough discussion of the different speech enhancement techniques can be found in Benesty *et al.* (2005). A bulk of the speech enhancement techniques are based on modifying the short time spectral amplitude (STSA) of the noisy speech signals. The techniques based on subtractive-type algorithms assume that the background noise is locally stationary to the degree that noise characteristics computed during the speech pauses are a good approximation to the noise characteristics during the speech activity. In addition to the basic spectral subtraction algorithm (Boll, 1979), several extensions and improvements have been proposed (Beh and Ko, 2003; Berouti *et al.*, 1979; Compernolle, 1992; Gustafsson *et al.*, 2001). Virag (1999) presents a detailed analysis of the effect of variations in the subtraction parameters like the over-subtraction factor, the spectral flooring factor, and the exponent on the residual noise as well as the intelligibility of the enhanced speech. It also presents a spectral subtraction algorithm that adapts the subtraction parameters in time and frequency based on the masking properties of the human auditory system.

McAulay and Malpass (1980) have shown that, under certain assumptions about the spectral characteristics of the speech signal and the noise, the spectral subtraction method is the maximum likelihood estimator of the variance of the speech spectral components. Ephraim and Malah (1984) have proposed a system that utilizes the minimum mean square-error short-time spectral amplitude (MMSE-STSA) estimator to enhance speech signals. This method assumes that each of the Fourier expansion coefficients of the speech and of the noise process can be modeled as Gaussian random variables with zero mean. Moreover, it is also assumed that these coefficients are independent of each other. The MMSE-STSA estimator which takes into account the uncertainty of speech presence (McAulay and Malpass, 1980) is also presented. The quality of the enhanced speech is better using the MMSE estimator that takes into account the speech presence uncertainty than the one that does not. The residual noise is perceived more as white noise than as musical noise and is attributed to the smooth variation of *a priori* signal-to-noise ratio (SNR) estimates (Cappe, 1994). The MMSE-STSA al-

[a]Electronic mail: omdesh@glue.umd.edu

 0001-4966/2007/121(6)/3886/13/$23.00

gorithm is extended by Ephraim and Malah (1985) to compute the STSA estimator that minimizes the mean-square error of the log-spectral amplitude, which is a more relevant criterion for perceivable distortions in speech. Loizou (2005) replaced the squared-error cost function used in the MMSE estimator by perceptually more relevant cost functions that take into account the auditory masking effects.

All of these speech enhancement methods make various restricting assumptions about the temporal and spectral characteristics of the speech signals and the corrupting noise. It will be shown in Sec. VI D that the performance of some of these methods deteriorates when the speech signals are corrupted by fluctuating noise. In this paper we present a speech enhancement technique, called the Modified Phase Opponency (MPO) speech enhancement technique, that makes minimal assumptions about the noise characteristics. The MPO speech enhancement scheme does not assume that the noise satisfies any statistical model or any degree of stationarity, nor does it need to estimate/update the noise characteristics but takes advantage of the known nature of speech signals. The MPO speech enhancement technique is thus potentially robust to fluctuating background noise. The performance of the MPO technique on fluctuating noise is presented in Sec. VI D. The MPO speech enhancement scheme is based on the auditory Phase Opponency (PO) model (Carney *et al.*, 2002) for tone-in-noise detection that includes a physiologically realistic mechanism for processing the information in neural discharge times. Some of the other speech enhancement techniques based on models of human auditory systems include Cheng and O'Shaughnessy (1991), Hansen and Nandkumar (1995), Mesgarani and Shamma (2005), and Tsoukalas *et al.* (1997).

In the present work, the MPO enhancement scheme is also combined with our Aperiodicity, Periodicity and Pitch (APP) detector (Deshmukh *et al.*, 2005b) to develop the MPO-APP speech enhancement scheme. The APP detector was developed to estimate the proportion of periodic and aperiodic energy in a speech signal. The MPO scheme is combined with the APP detector to remove the narrowband noise that might be seen as speech-like by the MPO processing and to retain some of the wideband speech signal that might be seen as noise-like by the MPO processing. The MPO-APP speech enhancement scheme produces a binary spectrotemporal mask called the *MPO profile* that distinguishes spectrotemporal regions where the speech signal is more dominant than the regions where the noise is more dominant. The use of binary masks is fairly common in the auditory scene analysis based speech enhancement and robust speech recognition techniques (Wang, 2005) and is motivated by the phenomenon of masking in the human auditory system. Hu and Wang (2004) proposed a computational auditory scene analysis method for segregating speech signals corrupted by various additive interferences. This method segregates the low-frequency resolved harmonics of the speech signals based on temporal continuity and cross-channel correlation and the high-frequency unresolved harmonics based on amplitude modulation and temporal continuity. The segments from the resolved harmonics are grouped according to common periodicity estimates and the segments from the unresolved harmonics are grouped according to common amplitude modulation rates. The proposed MPO-APP technique also contains an algorithm that reliably estimates the proportion of periodicity in speech signals but combines it with an algorithm that detects the presence of narrowband signals in noise to separate speech signals from the background noise.

The proposed speech enhancement scheme leads to the lowest value of the linear-predictive coefficients based objective measures and the highest value of the perceptual evaluation of speech quality (PESQ) measure compared to some of the other methods when the speech signals are corrupted by fluctuating noise. The performance of the proposed speech enhancement scheme is comparable to some of the other enhancement techniques when the global SNR and the noise type of the corrupting noise are not fluctuating.

II. THE PHASE OPPONENCY MODEL

A model for detection of tone-in-noise based on processing the information in neural discharge times is presented in Carney *et al.* (2002). This model exploits the frequency-dependent phase properties of the tuned filters in the auditory periphery and uses cross-auditory-nerve-fiber coincidence detection to extract temporal cues. It is shown that the responses of some of the cross-channel coincidence detectors are reduced when a tone is present in the background noise. This reduction in response to the presence of the target is referred to as Phase Opponency (PO). The performance of the PO model in the detection of low-frequency tones embedded in fixed-level or roving-level masking noise is consistent with that of humans (Carney *et al.*, 2002), making it a suitable model to detect narrowband signals corrupted by additive noise. In the present work, the PO model is extended to develop a speech enhancement scheme by utilizing the facts that (a) much of the speech signal is voiced and can be thought of as a combination of narrowband signals (i.e., harmonics) with varying amplitudes and (b) retaining the high-amplitude harmonics near the formant frequencies is perceptually more significant than the harmonics in the valley regions. The MPO processing scheme will thus not be able to retain the obstruents in a given speech signal although it does detect the frequency onset of strident friction in high SNR situations.

Figure 1 shows the PO model with center frequency (CF) equal to 900 Hz. The two nerve fibers are modeled as two gammatone filters with slightly different CFs. The magnitude and the phase response of the two gammatone filters are also shown in Fig. 1.

When the input is a tone at 900 Hz, the outputs of the two filters will be out of phase and the cross correlation will lead to a negative output. The output will remain negative as long as the input is a bandlimited signal centered at the CF (900 Hz in this case) and with bandwidth (BW) within the out-of-phase frequency region ($F_a - F_b$ in Fig. 1). We refer to the frequency region $F_a - F_b$ as the *out-of-phase* region and the rest of the frequency region as the *in-phase* region. When the input is a wideband signal, the output of the two filters will exhibit some degree of correlation and the cross-

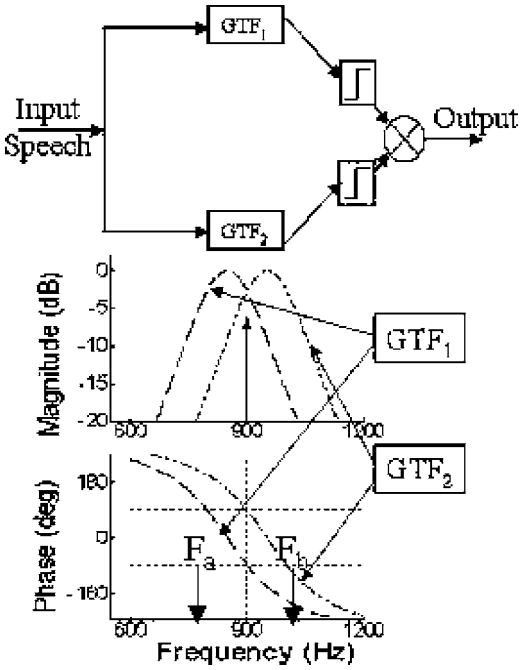

FIG. 1. PO filter pair to detect a tone at 900 Hz. GTF: Gammatone filter. The relative phase response of the two GTFs is out-of-phase in the frequency range $[F_a - F_b]$. The CFs for the two filters are 848.5 and 951.5 Hz. Adapted from Carney *et al.* (2002).

correlation output will be positive or very slightly negative. Thus the model is able to distinguish between narrowband signals and wideband noise. A saturating nonlinearity is applied to the output of each of the filters to minimize the effect of amplitude fluctuations on the overall output of the PO model.

Some of the issues with the PO model shown in Fig. 1 are that the relative magnitude response and the relative phase response of the two paths depend on the same set of parameters, making it difficult to manipulate either of the two independent of the other. It is difficult to predict the relationship between the parameters controlling the characteristics of the PO model and the width and the location of the *out-of-phase* region.

III. MODIFIED PHASE OPPONENCY MODEL

To address the concerns raised in Sec. II about the PO model, the MPO model was developed in such a way that the basic functionality of the PO model is maintained, but the various properties of the model can be analyzed and modified independent of each other. Figure 2 shows the schematic of the MPO model used in the present work. In the MPO

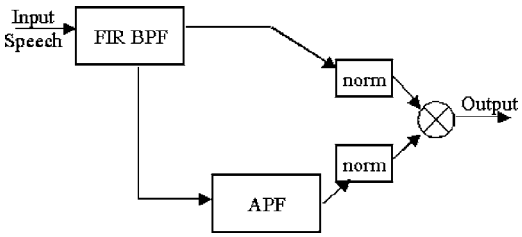

FIG. 2. Modified PO filter pair. "Norm" indicates amplitude normalization.

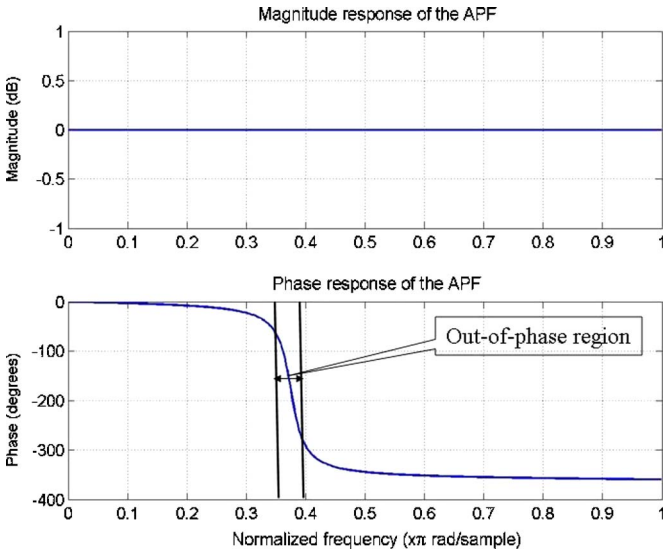

FIG. 3. (Color online) Magnitude and phase response of a typical all pass filter with one pair of complex conjugate poles. The *out-of-phase* frequency region is also shown.

model, one of the paths is modeled as a linear-phase finite impulse response (FIR) bandpass filter (BPF). The other channel is modeled as a concatenation of the same FIR BPF followed by an all pass filter (APF). The relative phase response of the two paths can be manipulated by changing the parameters of the APF which does not introduce any changes in the relative magnitude response. The magnitude response of the two paths can be manipulated by changing the parameters of the BPF which does not introduce any changes in the relative phase response. Thus the MPO model shown in Fig. 2 allows for manipulation of the relative magnitude response and the relative phase response independently of the other. The filters are followed by an amplitude normalizing scheme to minimize the effect of magnitude information in the cross-frequency coincidence. The characteristics of the BPF are mainly decided by the range of the target frequency that is to be detected. The characteristics of the APF are mainly decided by the expected frequency range and BWs of the target signals. The relation between the parameters of the APF and its phase response is explored below.

A. Mathematical formulation of the MPO model

Consider a second-order APF, $H(z)$, with one pair of complex conjugate poles,

$$H(z) = \frac{(z^{-1} - a^*)(z^{-1} - a)}{(1 - a^* z^{-1})(1 - a z^{-1})},$$

where $a = re^{j\theta}$ is the complex pole and a^* is its complex conjugate. Figure 3 shows the magnitude and the phase response of the APF for $r = 0.93$ and $\theta = 0.375\pi$. The magnitude response is uniformly 0 dB for all values of the frequency ω. The phase response, $\Phi(\omega)$, is given by

$$\Phi(\omega) = -2\omega - 2\tan^{-1}\left[\frac{2r\sin(\omega)\cos(\theta) - r^2\sin(2\omega)}{1 - 2r\cos(\omega)\cos(\theta) + r^2\cos(2\omega)}\right]. \quad (1)$$

We are interested in deriving the relationship between r and θ and the location and the width of the *out-of-phase* region. Notice from Fig. 3 that locating the *out-of-phase* region is equivalent to locating the frequency region where the phase response is the steepest. The frequency region where the phase response, $\Phi(\omega)$, is the steepest can be located by finding the frequency where the slope of the phase response has an inflection point, i.e., by finding the ω for which $d^2(\Phi(\omega))/d\omega^2 = 0$. Equating the numerator of the second-order derivative to zero and simplifying leads to

$$D(r,\omega,\theta)\cos\theta = \cos\omega,$$

where

$$D(r,\omega,\theta) = \left[\frac{1 + 2r^2 + 4r^2(\cos^2\omega + \sin^2\theta) + r^4}{4r(1+r^2)}\right]. \quad (2)$$

If we assume $\omega = \theta$, then the $\cos\theta$ term on the left-hand side of Eq. (2) is balanced by the $\cos\omega$ term on the right-hand side. Further, if we assume $r = 1$, then the sum of the coefficients in the numerator of $D(r,\omega,\theta)$ $[1+2+4+1=8]$ is exactly equal to that of the coefficients in the denominator $[4 \times (1+1) = 8]$. Thus, the equality in Eq. (2) holds for $\theta = \omega$ and $r = 1$. However, stability of the APF dictates that the magnitude of r be less than 1. It can easily be verified (Deshmukh, 2006) that $D(r,\omega,\theta)$ remains close to 1 even for various values of r less than 1. Thus, it is reasonably accurate to assume that:

The slope of the phase response, $\Phi(\omega)$, of a stable APF with a pair of complex conjugate poles at $re^{\pm j\theta}$ is the steepest at frequency $\omega = \theta$. Moreover, this frequency location is independent of r, the magnitude of the pole.

The phase response, $\Phi(\omega)$, of the APF at $\theta = \omega$, under the assumption of $r \approx 1$ simplifies to:

$$\Phi(\omega) \approx \begin{cases} \pi & \text{if } \cot\theta < 0 \\ -\pi & \text{if } \cot\theta > 0. \end{cases} \quad (3)$$

The phase response at $\omega = \theta$ can thus be approximated as $\pm\pi$.

The closer the value of r to 1, the more accurate the approximation is. The next step is to express the dependence of the width of the *out-of-phase* region on the values of r and θ. This is equivalent to expressing the slope of $\Phi(\omega)$ at $\omega = \theta$ in terms of r and θ. Evaluating the derivative of $\Phi(\omega)$ with respect to ω at $\omega = \theta$ for various values of θ and ω shows that for a given value of r, the value of $d(\Phi(\omega))/d\omega$ is not very sensitive to the value of θ. On the other hand, it is very sensitive to the choice of r. It can thus be assumed that:

$d(\Phi(\omega))/d\omega$ evaluated at $\omega = \theta$ (i.e., at the frequency where the phase response of the APF is the steepest) is independent of θ and is dependent only on the value of r.

In summary, for an APF with poles at $a = re^{j\theta}$ and a^*, the *out-of-phase* region is centered around $\omega = \theta$ irrespective of the value of r (the phase response at $\omega = \theta$ is approximately equal to $\pm\pi$) and the width of the *out-of-phase* region is controlled only by the value of r, irrespective of the value of θ.

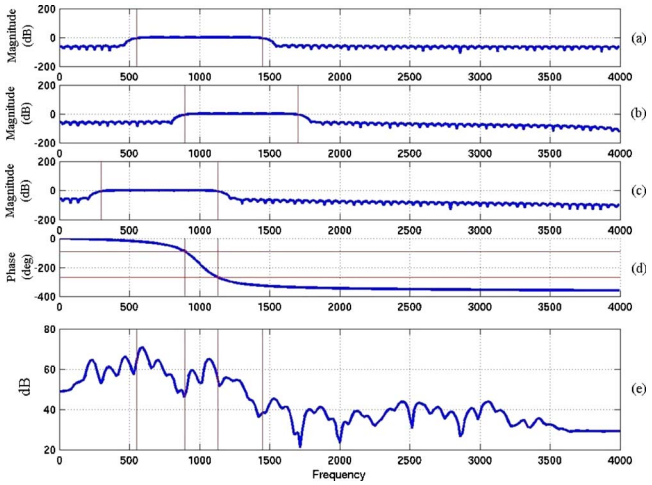

FIG. 4. (Color online) Magnitude response of the (a) symmetric BPF; (b) upward-skewed BFP; and (c) downward-skewed BPF that will be used in the MPO structure with CF = 1000 Hz. (d) Phase response of the APF that will be used in the MPO structure with CF = 1000 Hz. (e) Spectral slice of a sonorant region in speech signal.

B. MPO design

Our aim is to design a MPO structure to detect signals centered at ω_c and of bandwidths less than or equal to $\Delta\omega$. The first requirement is to choose the APF such that the phase response is about $-\pi$ at ω_c. Analysis in Sec. III A shows that this requirement is satisfied by choosing the phase of the pole of the APF as $\theta = \omega_c$ [see Eq. (3)], irrespective of what the value of r, the magnitude of the pole of the APF, is. The expected bandwidth of the target signal, $\Delta\omega$, dictates the value of r. The value of r should be such that the phase response, $\Phi(\omega)$, of the APF spans $-\pi/2$ to $-3\pi/2$ in $\Delta\omega$ rad centered around ω_c (i.e., the *out-of-phase* region corresponds to the expected bandwidth of the input signal). Unfortunately, there is no closed form relation between r and the BW of the *out-of-phase* region. For a given expected bandwidth, $\Delta\omega$, the value of r has to be computed using multiple trials. But, as is shown in Sec. III A, for a given bandwidth, the value of r is dependent only on the bandwidth and is independent of the center frequency of the signal. Assume that the optimal value of r for the expected bandwidth of $\Delta\omega$ is $r = r_c$. The APF is completely defined by specifying the parameters r and θ.

The next step is to choose the FIR BPF. The BPF has to satisfy two constraints:

(1) The passband of the BPF should include the *out-of-phase* region.
(2) The passband should be such that the MPO output is negative for narrowband signals (with bandwidth less than or equal to $\Delta\omega$ and centered at the CF) and positive for wideband signals.

Several BPFs can be designed that satisfy the two above-mentioned constraints. Figure 4 shows three such BPFs for a

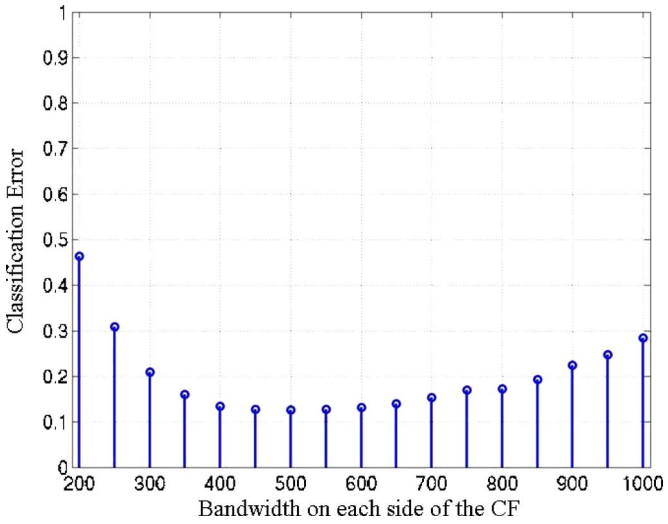

FIG. 5. (Color online) Variation in the binary classification error as the bandwidth of the BPF is varied. The two classes are: (a) presence of narrowband signal in broadband noise at 0 dB SNR and (b) broadband noise.

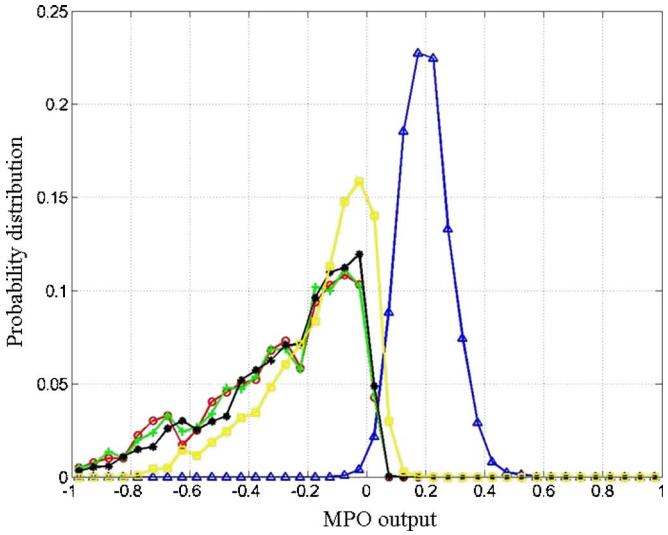

FIG. 6. (Color online) Distribution of the output of MPO model when the input is white noise (△); band-limited signal at ∞ dB SNR (○); at 20 dB SNR (+); at 10 dB SNR (black curve: *); and at 0 dB-SNR (□).

MPO structure that can be used to distinguish narrowband signals centered at 1000 Hz with bandwidth less than 250 Hz from wideband noise signals. Initial versions of the MPO-based speech enhancement scheme (Deshmukh and Espy-Wilson, 2005; Deshmukh *et al.*, 2005a) used BPFs with passband symmetry across the CF of the MPO structure. Figure 4(a) shows the magnitude response of such a symmetric BPF. Figure 4(d) shows the phase response of the APF used in the corresponding MPO structure. The optimal bandwidth of the BPF is computed by calculating the two-class (narrowband-signal-in-noise versus noise-only) classification error for different choices of bandwidths and choosing the one that gives the least error. For low values of bandwidth the output for the presence-of-signal situations as well as for the absence-of-signal situations will be negative leading to many false-positive errors whereas for high values of bandwidth the output for the absence-of-signal situations as well as for the presence-of-signal situations will be positive leading to many correct-miss errors. Figure 5 plots the total classification error for a MPO structure that uses the APF shown in Fig. 4(d) and for different bandwidths of the corresponding symmetric BPF. The optimal BPF is $450\times2=900$ Hz. Note that the classification error is close to the minimum value over a wide range of the BPF bandwidths (800–1200 Hz) and is thus not very sensitive to the exact choice of the BPF bandwidth. In all the simulations, the BPF length was kept constant at 76 samples, and was constructed using the standard MATLAB routine *fir1*.

Figure 6 shows the distribution of the output of the MPO model shown in Figs. 4(a) and 4(d) for 5000 frames (a frame is 30 ms long with a frame rate of 5 ms) each of white noise and a bandlimited signal centered at 1000 Hz and of bandwidth 250 Hz corrupted with white noise at ∞, 20, 10, and 0 dB SNR. Notice that the distribution of the output for white noise is well separated from that for the bandlimited signal at ∞ dB SNR. Moreover, the distribution of the bandlimited signal corrupted by white noise remains quite similar over a wide range of SNRs (∞ to 0 dB). The threshold to discriminate the presence of the signal from the absence of signal is computed using the maximum likelihood (ML)-based likelihood ratio test (LRT). The optimal threshold value is −0.0215. Figure 7 shows the receiver operating characteristic (ROC) curve for MPO detectors at three different CFs: 950 Hz (dash curve), 1000 Hz (dotted curve), and 1050 Hz (solid curve). The optimal threshold values are: −0.0183, −0.0215, and −0.0197, respectively. The ROC curves in Fig. 7 were obtained by varying the threshold over the range: [opt_thresh −0.05 : −0.005] where opt_thresh is the optimal threshold for the corresponding MPO detector. In general, it is observed that the probability of false alarm is below 3% for threshold values below 0 (the theoretical threshold) and the probability of detection remains above 96% for threshold values as low as opt_thresh −0.05 indicating that the exact value of the threshold is not critical for the overall operation of the MPO detectors. It is worth pointing out that the thresholds for the MPO detectors at different CFs are computed using the two extremes of (a) narrowband sig-

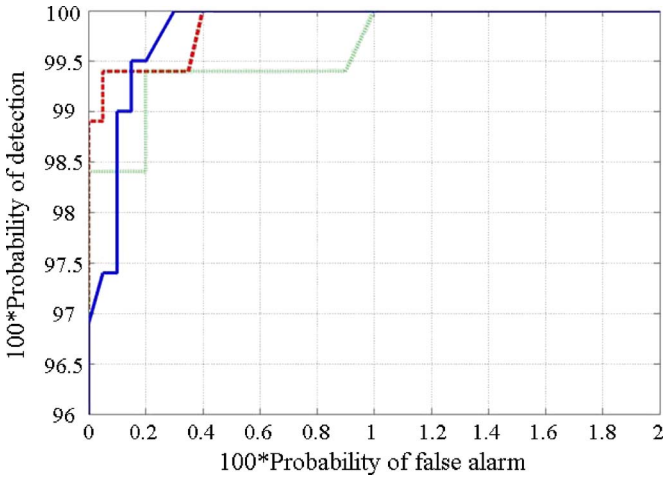

FIG. 7. (Color online) ROC curves for MPO detectors at three different CFs: 950 Hz (dash curve); 1000 Hz (dotted curve); and 1050 Hz (solid curve).

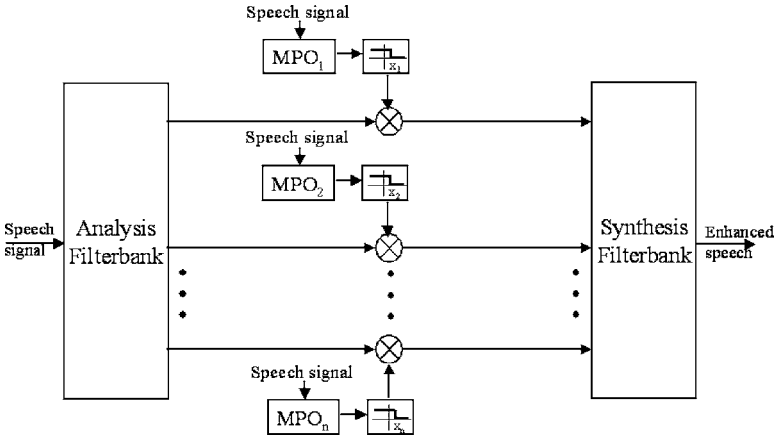

FIG. 8. Schematic of the MPO-based speech enhancement scheme. The threshold, x_i, is trained using the ML-LRT technique and all the regions with output above this threshold are suppressed.

nals centered at the CF and (b) white noise. These thresholds are not retrained when the background conditions change. It is shown in Sec. VI that the MPO speech enhancement scheme is robust to various noise types at different levels with no additional noise-specific training.

The other two BPFs shown in Fig. 4 have passbands skewed upward or downward in frequency with respect to the CF of the MPO structure [Figs. 4(b) and 4(c), respectively]. Both of these BPFs offer some advantages over the symmetric BPF and will be discussed in Sec. IV A.

IV. MPO-BASED SPEECH ENHANCEMENT

Speech signals, for the most part, are composed of narrowband signals (i.e., harmonics) with varying amplitudes. The MPO-based speech enhancement scheme attempts to detect and maintain these time-varying narrowband signals while attenuating the other spectro-temporal regions. Figure 8 shows the schematic of the MPO-based speech enhancement scheme. The analysis-synthesis filterbank can be any near-perfect reconstruction (PR) filterbank. The overall performance of the MPO enhancement scheme is insensitive to the choice of the analysis-synthesis filterbank. In the present work, a DFT-based PR filterbank is used. In a related work (Anzalone *et al.*, submitted), a near-PR analysis-synthesis gammatone filterbank proposed by Hohmann (2002) was used. The input speech signal is split into overlapping frames of length 30 ms at a frame rate of 5 ms. Each MPO_i in the figure is a MPO structure (Fig. 2) with a different CF. The CFs are spaced every 50 Hz from 100 Hz to just below the maximum frequency. The threshold, x_i, to discriminate the presence of signal from the absence of signal is trained separately for each MPO structure as described in Sec. III B. The MPO structures act as switches allowing the speech frame to either pass as it is for reconstruction if the corresponding MPO output is less than the threshold (indicating presence of signal) or be attenuated by 10 dB if the output is greater than or equal to the threshold (indicating absence of signal). Attenuating the signal-absent regions by 10 dB, instead of zeroing them out completely, reduces the perceptual effect of the residual noise. Higher attenuation of the speech-absent regions leads to an overall increase in the objective distortion measures as well as a lower PESQ measure.

In the initial version of the MPO-based speech enhancement scheme (Deshmukh and Espy-Wilson, 2005; Deshmukh *et al.*, 2005a), each of the MPO_i in Fig. 8 consisted of a symmetric BPF and the APF was configured so that signals with bandwidths less than or equal to 150 Hz would lead to negative outputs. Such a scheme performs well when the input speech signal is corrupted by additive white noise which has a relatively flat spectrum with minimal level fluctuations over time. But it passes a lot more noise when the corrupting signal is colored noise with fluctuating levels. To overcome this problem, the present version of the speech enhancement scheme uses two sets of MPO structures at each CF. Each set has five different MPO structures such that each one of them has a different *out-of-phase* region ranging from 120 to 250 Hz. Noise can be wrongly seen as speech by one or more of the five different structures in either set, but it is rarely seen as a narrowband speech signal by all five structures. Similarly, narrowband speech signals are almost always seen as speech signals by *all* five MPO structures. Using five structures in each set strikes a better balance between computational cost and the amount of residual noise as compared to a higher or lower number of structures per set.

A. Choosing the BPF for speech signals

Consider the spectral slice shown in Fig. 4(e). The harmonics close to F2 (around 1050 Hz) fall in the out-of-phase frequency region of the MPO structure whereas the harmonics close to F1 (around 550 Hz) fall in the in-phase frequency region of the MPO structure with symmetric BPF. The amplitude of F1 (and hence that of the harmonics close to F1) is greater than that of F2 due to the known spectral tilt in sonorant regions of speech signals. As a result, although there is a narrowband signal at the CF of the MPO, the output of this MPO structure will be positive and therefore the speech information present in that frequency region will be missed. The upward skewed BPF shown in Fig. 4(b) will attenuate the F1 region and thus the output of the upward-skewed MPO structure will be driven mainly by the frequency content near the CF and in the frequency region above the CF. Most of the time, such upward skewed MPO structures are able to correctly detect the speech information as they inherently take advantage of the spectral tilt present in sonorant speech regions. The F2 information in Fig. 4(e)

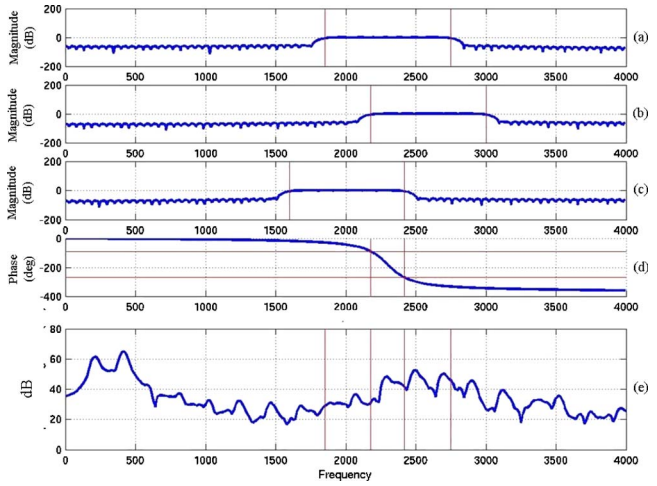

FIG. 9. (Color online) Magnitude response of the (a) symmetric BPF; (b) upward-skewed BFP; and (c) downward-skewed BPF that will be used in the MPO structure with CF=2300 Hz. (d) Phase response of the APF that will be used in the MPO structure with CF=2300 Hz. (e) Spectral slice of a sonorant region in speech signal.

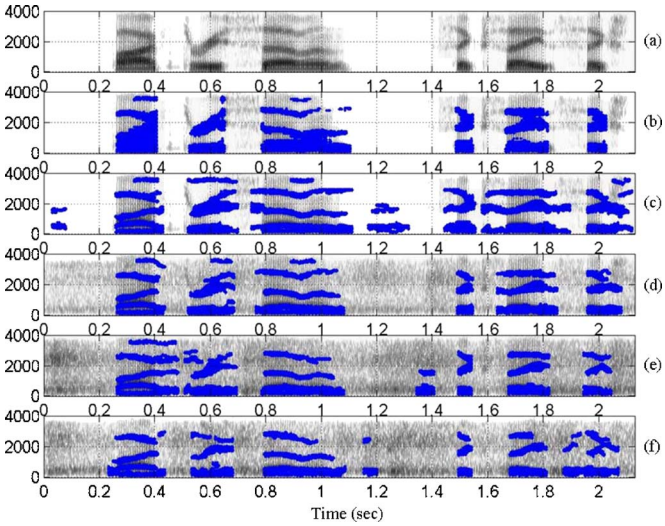

FIG. 10. (Color online) (a) Spectrogram of the utterance "five three seven six eight six;" (b) the energy-based *maximal mask*; and (c), (d), (e), and (f) MPO profile at ∞, 20, 10, and 5 dB SNR, respectively.

that was missed by the symmetric MPO structure will be detected by the upward-skewed MPO structure.

The downward-skewed filter shown in Fig. 4(c) is the exact opposite of the upward-skewed filter. The passband of the downward-skewed filter extends downward in frequency with respect to the CF of the MPO structure. Consider the case shown in Fig. 9(e) when two formants are of comparable amplitudes and are in close proximity in frequency (hence, the harmonics near these formant frequencies have comparable amplitudes). In such cases, the upward skewed MPO structures will detect the higher frequency harmonics, but will fail to detect the lower frequency harmonics. The downward skewed MPO structures centered on the lower frequency harmonics will be able to successfully detect such instances because their *in-phase* region extends on the lower frequency side. Each CF is thus analyzed using an upward MPO structure as well as a downward MPO structure.

B. Speech enhancement scheme

As explained in Sec. IV A each MPO_i in Fig. 8 consists of five upward-skewed MPO structures (one set) and five downward-skewed MPO structures (second set) all tuned to the same CF, but with the width of the *out-of-phase region* ranging from 120 to 250 Hz. The speech enhancement scheme can now be described as a two-step process. In the first step, the temporal regions where speech is present are computed. For a temporal region to be voted as *speech present*, it has to satisfy two conditions: (a) The MPO output of at least one frequency channel from all five different upward-skewed or all five different downward-skewed MPO structures should be at least four times more negative than the threshold for that particular channel and (b) the temporal region should be at least 50 ms long.

In the second step, the frequency channels within the *speech-present* temporal regions where speech information is present are computed by finding the channels where the MPO output from all five upward skewed or all five downward skewed MPO structures is below the corresponding threshold. The noisy speech signal from only these channels is used for reconstruction.

The first step is to ensure that wrongful insertions in the decision of temporal *speech-present* regions are kept to a minimum and the second step is to ensure that all of the valid *speech-present* spectral channels in a given *speech-present* temporal region are detected.

The MPO speech enhancement scheme can thus be thought of as applying a time-frequency two-dimensional binary mask to the input speech signal. The binary mask has a value of one in *speech-present* spectrotemporal regions where the speech signal is dominant and has a value of zero in the *speech-absent* spectrotemporal regions where the noise signal is more dominant. The binary mask is referred to as the "MPO profile." Figure 10(a) shows the spectrogram of the utterance "five three seven six eight six" in clean. Figures 10(c)–10(f) shows the corresponding MPO profile when the utterance is corrupted by subway noise at ∞, 20, 10, and 5 dB SNR, respectively. The dark (blue) regions are the spectrotemporal channels where the MPO profile is one. The speech signal from these channels is used "as-is" to construct the enhanced speech signal. The MPO processing retains almost all of the perceptually significant speech information when the input signal is clean. Some of the formant transitions through the fricative regions as well as the frequency onset of strident frication are also tracked (e.g., around 1.6 and around 2.2 s). As the SNR is reduced, most of the strong sonorant information is detected by the MPO processing while very little noise is mistaken as speech signal [e.g., around 1.4 s in Fig. 10(e) and around 1.2 and 1.9 s in Fig. 10(f)].

Figures 11(a) and 11(c) compare the MPO profiles of two utterances "four three six four six three" (left) and "one five" (right) in clean and when they are corrupted by car noise and subway noise, respectively, at 10 dB SNR. The MPO processing retains almost all of the perceptually significant speech information when the input signal is clean.

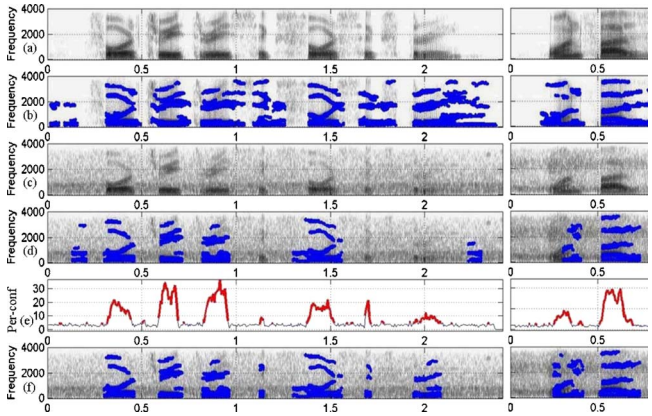

FIG. 11. (Color online) (a) Spectrograms of clean signals "four three three six four six three" (left) and "one five" (right). (b) The MPO profiles of the corresponding signals. (c) Spectrograms of the signals corrupted by car noise (left) and subway noise (right) at 10 dB SNR. (d) The MPO profiles of the corresponding noisy signals. (e) Periodicity confidence. Frames with periodicity confidence greater than per_thresh are shown with a thicker line. (f) The combined MPO-APP profiles of the noisy signals.

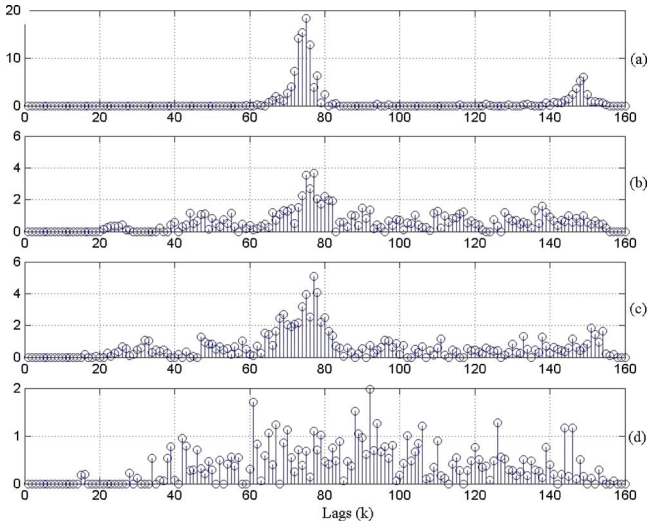

FIG. 12. AMDF clusters for a periodic frame at different SNRs and for an aperiodic frame.

When the input signal is noisy, the MPO processing, while detecting most of the strong harmonics, fails to detect the short vowel /I/ in both the instances of "six" [1.12–1.16 and 1.68–1.73 s in Fig. 11(d)] as well as completely misses the last "three" (1.9–2.3 s). Also notice that the /w/ in the noisy "one" [0.22–0.3 s Fig. 11(d) on the right] is not detected by the MPO processing. F1 and F2 for /w/ are very close and thus look like a wideband signal which is not detected by the MPO processing. Also note that in all of these regions the temporal signal has strong periodicity which distinguishes it from the temporal signal in the noise-only regions. On the other hand, some of the noise is wrongly seen as narrowband signal by the MPO processing and is passed for reconstruction [e.g., 0.12–0.2 and 2.23–2.3 s in Fig. 11(d)]. But these noise regions are not as periodic as the speech-present regions.

The number of noise-insertions and the number of speech-deletions can be reduced by combining the MPO processing with an algorithm that reliably estimates the periodicity information in speech signals. In the present work, the MPO processing is combined with our APP detector (Deshmukh *et al.*, 2005b). The APP detector estimates the proportion of periodic and aperiodic energy in each spectrotemporal channel as well as the confidence of periodicity in each time frame. Such a time-frequency analysis by the APP detector makes it convenient to combine the APP detector with the MPO processing. The narrowband noise that is inserted in the reconstructed speech signal by the MPO processing does not have a harmonic structure across the frequency channels similar to that of the periodic regions in a speech signal. On the other hand, the locally wide-band regions of speech signals formed due to the proximity of two or more formants retain a coherent harmonic structure across the frequency channels. The APP detector captures this coherence of across-frequency-channel periodicity and can reduce such speech-deletions in the MPO speech enhancement scheme.

Section V presents a brief overview of the APP detector, the use of the APP detector as a separate speech enhancement technique, and the way in which the MPO processing is combined with the APP detector.

V. APERIODICITY, PERIODICITY, AND PITCH DETECTOR

The processing in the APP detector begins by splitting the speech signal into 60 frequency channels that are equispaced on the ERB scale. The average magnitude difference function (AMDF) is computed on the envelope of each of the frequency channels at a frame rate of 2.5 ms and a frame length of 20 ms. The AMDF is given by

$$\gamma_n(k) = \sum_{m=-\infty}^{m=\infty} |x(n+m)w(m) - x(n+m-k)w(m-k)|,$$

where $x(n)$ is the envelope signal, k is the lag value in samples, and $w(m)$ is a 20-ms-long rectangular window.

For a periodic signal, the AMDF attains minima (referred to as dips) values close to one at lags equivalent to the pitch period and its integer multiples. Moreover, for a frame in a periodic speech region, the pitch period is quite similar across the different frequency channels. Thus, lag-wise addition of the AMDF dips across the frequency channels leads to clustering at integer multiples of the pitch period as shown in Fig. 12(a) for a frame during the /w/ centered at 0.27 s in clean condition (Fig. 11, right-hand side). Figures 12(b) and 12(c) show that the clusters are retained even as the SNR is reduced to 10 and 5 dB, respectively. For an aperiodic frame, the lag-wise addition of the AMDF dips across the frequency channels results in dips that are randomly scattered over the range of the possible lag values. For example, Fig. 12(d) shows the lag-wise addition of the AMDF dips for an aperiodic frame centered at 0.46 s in the utterance shown on the right-hand side in Fig. 11 and corresponds to the phoneme /f/ in "five."

The periodicity confidence of a given temporal frame is computed as the strength of the dips close to the cluster peaks relative to the strength of the rest of the dips. The

locations of the cluster peaks is the estimate of the pitch frequency. For aperiodic frames, where no strong clusters exist, a cluster is formed around the lag with the maximum strength. The periodicity confidence values for the plots in Figs. 12(a)–12(d) are: 50.8, 9.7, 9.3, and 0.4, respectively. The periodicity confidence can thus distinguish a periodic frame from an aperiodic frame even when the speech signal is corrupted by noise. The optimal threshold of the periodicity confidence, per_thresh, to distinguish periodic frames from aperiodic frames is computed using periodic and aperiodic frames from clean speech signals and is not altered as the background conditions change.

Frequency channels where the AMDF dips close to the cluster peaks are stronger than the AMDF dips away from the cluster peaks are classified as periodic channels. The rest of the channels are classified as aperiodic channels. This leads to a spectrotemporal binary mask, called the "APP profile," which has a value of 1 in frequency channels which are estimated to be periodic and a value of 0 in channels which are estimated to be aperiodic. Also, note that if any of the frequency channels has periodic noise then the periodic noise will be classified as periodic speech signal only if the corresponding period of the noise is very close to the estimated pitch period of the speech signal. [Please refer to Deshmukh *et al.* (2005b) for more details on the various stages of the APP detector.]

A. APP-based speech enhancement technique

The APP speech enhancement technique consists of the APP detector sandwiched between a near-PR analysis-synthesis gammatone filterbank proposed by Hohmann (2002). (The setup is similar to the one shown in Fig. 8 for the MPO speech enhancement scheme.) The filters of the near-PR filterbank are equi-spaced on the ERB scale and the CFs and the bandwidths are chosen such that they match the filters used in the analysis of the APP detector. The spectrotemporal channels where the APP profile is one (indicating presence of periodic signal) are passed as-is to the synthesis filterbank and the rest of the spectrotemporal channels are attenuated by 10 dB before being passed to the synthesis filterbank.

B. Combining MPO processing with the APP detector

As mentioned earlier, some of the main shortcomings of the MPO processing are: (1) Noise insertions: Where some of the narrowband noise is detected as speech-like although it lacks the harmonicity typical of the sonorant speech regions and (2) speech deletions: Where locally wideband regions of speech signals are not retained although they have a coherent harmonic structure across the frequency channels.

For each of the frequency channels analyzed by the MPO processing, the AMDF dips are computed on the envelope of the channel signal. The periodicity confidence of each temporal frame is computed by the lag-wise addition of the AMDF dips across the frequency channels as mentioned in Sec. V. A given *speech-present* region in the MPO-processing is classified as *speech-absent* (and thus not used for reconstruction) if the maximum value of the periodicity confidence in the corresponding region is below per_thresh. For example, consider the MPO-estimated speech-present region between 0.13 and 0.20 s [Fig. 11(d), left-hand side]. The maximum value of the periodicity confidence in this region is below per_thresh [frames with periodicity confidence greater than per_thresh are shown with a thicker line in Fig. 11(e)]. Thus, this region will not be used for reconstruction. Such a strategy leads to a reduction in noise insertions.

All the frames that are classified as *speech absent* by the MPO processing but have the corresponding periodicity confidence greater than $2 \times$ per_thresh (indicating strong periodicity) are classified as *speech present* and reintroduced for synthesis. Figure 11(f) shows the combined MPO-APP profile for the noisy utterances shown in Fig. 11(c). Notice that the noise insertions are removed and most of the speech deletions are reintroduced.

VI. EVALUATION

A. Database

The Aurora database (Hirsch and Pearce, 2000) was used to evaluate the MPO-based speech enhancement scheme. This database is a derivative of the TIdigits database resampled at 8 kHz. The database has three different subsets for testing. In the present study, only the test subset A was used for evaluation. Subset A consists of utterances corrupted by four different noise types at seven different SNRs from ∞ to -5 dB. Each utterance in the Aurora database is corrupted by one of the noise types at a given SNR. The four different noise types are: Subway noise, babble, car noise, and exhibition hall noise. These are referred to as N1, N2, N3, and N4, respectively.

A database of speech signals corrupted by fluctuating noise (F-DB) was formed from this subset of the Aurora database. Each utterance in F-DB database consists of seven digits. Each of the seven digits is corrupted by a different noise type at a different SNR. The F-DB database consists of 1120 such utterances.

The overall MPO speech enhancement strategy (i.e., number of MPO structures at each CF, combining the output of these MPO structures) was developed using a subset of the TIMIT database corrupted by white noise and a subset of the Aurora database.

The performance of the MPO and the combined MPO-APP speech enhancement technique was compared with some of the speech enhancement techniques presented in the literature: (a) MMSE-STSA (Ephraim and Malah, 1984); (b) MMSE-STSA with noncausal SNR estimation (NC-MMSE) (Cohen, 2004); (c) logMMSE-STSA (Ephraim and Malah, 1985); (d) Generalized Spectral Subtraction (GSS) (Compernolle, 1992); (e) Hu-Wang method (2004); and (f) APP-based speech enhancement technique. The code for the Hu-Wang method was downloaded from their lab website. The code was used as-is except the sampling rate was changed from 16 to 8 kHz.

The evaluations presented here only compare the quality of the enhanced speech signals and not their intelligibility.

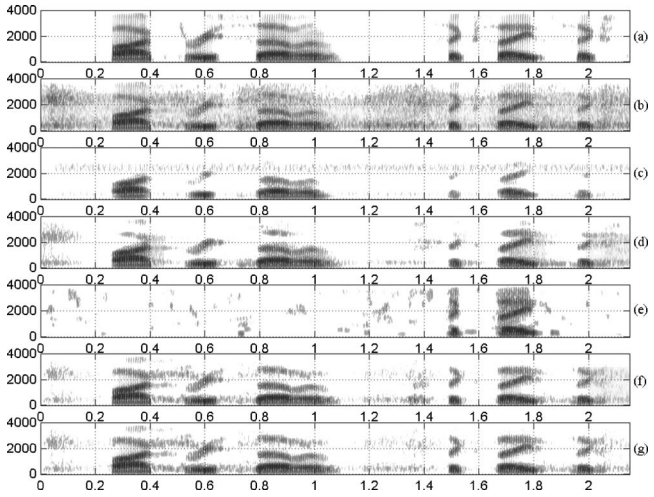

FIG. 13. Spectrogram of (a) the clean speech signal "five three seven six eight six." (b) The speech signal corrupted by subway noise at 10 dB SNR. The speech signal enhanced using the (c) GSS technique, (d) MMSE technique, (e) Hu-Wang technique, (f) MPO technique, and (g) MPO-APP technique.

Detailed evaluations comparing the intelligibility of the speech signals enhanced using the different techniques will be reported in the future.

B. Spectrograms displays

We begin the evaluations by comparing the spectrograms of the speech signals enhanced using different techniques. Figure 13 shows the spectrograms of a clean speech signal, the speech signal corrupted by subway noise at 10 dB SNR, and the speech signals enhanced by the GSS method, the MMSE-STSA method, the Hu-Wang method, the proposed MPO method, and the combined MPO-APP method. The GSS method has a relatively less amount of residual noise but suppresses alot of high-frequency low-energy speech information. The MMSE-STSA method is able to retain most of the speech information but passes alot more noise compared to the other methods. The MPO method, on the other hand, attenuates most of the residual noise while retaining much of the speech signal including the high frequency low energy speech information. For example, the MPO method is able to retain the weak F3 information around 2500 Hz near 0.65 s and again around 2700 Hz near 1.5 and 1.95 s while passing very little noise. Also notice that the combined MPO-APP method removes the noise around 1.4 s which was passed by the MPO method.

C. Evaluation using objective measures

The performance of the MPO speech enhancement scheme was evaluated using three objective quality assessment measures that have a high degree of correlation with subjective quality. These three measures are based on the dissimilarity of the linear predictive (LP) coefficients between the original and the enhanced speech signals (Hansen and Pellom, 1998).

(1) *Itakura-Saito (IS) distortion measure*: The IS distortion measure between a frame of clean speech signal and the corresponding frame of the enhanced speech signal is computed by

$$d_{\mathrm{IS}} = \left[\frac{\sigma_c^2}{\sigma_p^2}\right]\left[\frac{L_p R_c L_p^T}{L_c R_c L_c^T}\right] + \log\left[\frac{\sigma_p^2}{\sigma_c^2}\right] - 1,$$

where L_c and L_p are the LPC vectors for the clean frame and the processed frame, respectively, σ_c^2 and σ_p^2 are the all-pole gains for the clean frame and the processed frame, respectively, and R_c is the autocorrelation matrix of the clean frame.

(2) *Log-likelihood ratio (LLR) measure*: The LLR measure, unlike the IS measure, does not compare the all-pole gains of the clean frame and the processed frame and thus lays more emphasis on the difference in the overall spectral envelopes of the two frames. The LLR measure is computed using

$$d_{\mathrm{LLR}} = \log\left[\frac{L_p R_c L_p^T}{L_c R_c L_c^T}\right].$$

(3) *Log-area-ratio (LAR) measure*: The LAR measure is computed using the Pth order LP reflection coefficients of the clean frame and the processed frame in the following way:

$$d_{\mathrm{LAR}} = \left[\frac{1}{P}\sum_{j=1}^{P}\left[\log\frac{1+r_c(j)}{1-r_c(j)} - \log\frac{1+r_p(j)}{1-r_p(j)}\right]^2\right]^{1/2},$$

where r_c and r_p are the reflection coefficients of the clean frame and the processed frame, respectively.

All three measures return frame-level scores for a given processed speech signal. An overall score is computed by calculating the mean of the frame-level scores of the frames with the lowest 95% scores. Such a scheme removes a fixed number of frames which may have unrealistically high scores (Hansen and Pellom, 1998). An overall score of zero implies the processed speech signal is exactly identical to the original clean speech signal. Higher values indicate a greater degree of distortion in the processed speech signal.

The performance was also evaluated using the objective perceptual quality measure called the PESQ measure (Rix *et al.*, 2001). The PESQ measure is the ITU-T standard to evaluate the perceptual quality of processed speech signal. The PESQ evaluation includes aligning the clean (reference) signal and the enhanced signal in time and processing them through an auditory transform. The auditory transform includes models of various stages of the human auditory appa-

TABLE I. IS distortion measure at different SNRs. The results are averaged across the four different noise types used in this study.

Type	Clean	20 dB	10 dB	5 dB
GSS	1.407	3.839	3.774	3.427
MMSE	0.526	1.448	2.571	3.453
logMMSE	1.642	3.584	5.230	7.468
NC-MMSE	0.623	2.221	3.712	8.765
Hu-Wang	NaN	NaN	NaN	NaN
APP	15.084	4.771	3.183	3.257
MPO	1.034	1.425	2.018	2.511
MPO-APP	1.083	1.399	1.981	2.469

TABLE II. LAR measure at different SNRs. The results are averaged across the four different noise types used in this study.

Type	Clean	20 dB	10 dB	5 dB
GSS	1.712	3.987	5.372	5.758
MMSE	0.994	3.215	4.871	5.496
logMMSE	1.160	3.218	5.211	5.867
NC-MMSE	0.779	2.867	4.881	6.013
Hu-Wang	7.761	17.841	32.533	40.827
APP	2.590	4.538	5.433	5.923
MPO	1.939	4.070	5.003	5.577
MPO-APP	1.919	4.022	4.941	5.507

ratus. The outcome of the PESQ measure is an estimate of the subjective mean opinion score (MOS), which has values between 0 (poor quality) and 4.5 (no perceptual distortion).

Table I compares the IS distortion measure at different SNRs for the output of different enhancement techniques. The MPO processing leads to the lowest IS measure when the input speech signal is noisy. Combining the MPO processing with the APP detector (MPO-APP) leads to a further drop in the IS distortion measure in noisy conditions. The IS distortion measure computed on MPO-processed clean speech signals is higher than that computed on clean speech signals processed by other enhancement techniques. One of the reasons for this higher value could be that the spectral valleys in clean speech signal are further attenuated by the MPO processing. The IS distortion values for the Hu-Wang method were quite high and are hence replaced by NaNs (not-a-number). One of the reasons for the drop in the performance of the Hu-Wang method could be the change in the sampling rate. (Some of the parameters in the algorithm could be optimized for the default sampling rate of 16 kHz.).

Tables II and III compare the LAR and LLR measures, respectively, at different SNRs for the output of different enhancement techniques. The LAR and LLR measures obtained for the proposed MPO enhancement scheme are comparable with those obtained for some of the other enhancement schemes although the values are consistently higher (indicating more distortion) than those obtained for the MMSE-STSA enhancement scheme. Combining the MPO enhancement scheme with the APP detector (MPO-APP) consistently leads to a drop in the distortion values. Table IV compares the PESQ measure at different SNRs for the output of different enhancement techniques. The results are similar to those obtained for the LAR and LLR measures. The combined MPO-APP enhancement technique is an improvement over the MPO enhancement technique and its performance is comparable to that of some of the other enhancement techniques.

TABLE III. LLR measure at different SNRs. The results are averaged across the four different noise types used in this study.

Type	Clean	20 dB	10 dB	5 dB
GSS	0.103	0.453	0.782	0.958
MMSE	0.081	0.397	0.748	0.942
logMMSE	0.111	0.392	0.784	0.998
NC-MMSE	0.064	0.366	0.765	1.032
Hu-Wang	2.746	11.362	26.807	35.710
APP	0.241	0.556	0.775	0.923
MPO	0.159	0.519	0.761	0.943
MPO-APP	0.158	0.510	0.750	0.929

TABLE IV. PESQ measure at different SNRs. The results are averaged across the four different noise types used in this study.

Type	Clean	20 dB	10 dB	5 dB
GSS	3.718	3.042	2.361	1.928
MMSE	4.086	3.060	2.464	2.072
logMMSE	4.054	3.133	2.502	2.075
NC-MMSE	4.136	3.045	2.440	2.020
Hu-Wang	1.402	1.012	0.780	0.625
APP	3.002	2.768	2.377	2.059
MPO	4.116	2.955	2.331	1.958
MPO-APP	3.815	2.994	2.420	2.051

D. Robustness to fluctuating noise types and noise levels

The salient features of the MPO-based speech enhancement scheme are as follows: (a) It makes minimal assumptions about the noise characteristics (the only assumption is that noise is broader than the harmonics of the speech signal), (b) it does not need to estimate the noise characteristics nor does it assume the noise satisfies any particular statistical model, and (c) the noise removal performance on a given frame is independent of the performance on the adjoining frames. This scheme can thus be potentially robust when the level and the type of the background noise are fluctuating. The performance of the proposed MPO enhancement scheme was evaluated on the F-DB database of speech signals corrupted by fluctuating noise. Figure 14(a) shows the spectrogram of one of the clean signals "oh three zero six zero two four" from the F-DB database. The noisy signal is shown in Fig. 14(b). Figures 14(c)–14(f) compare the spectrograms of

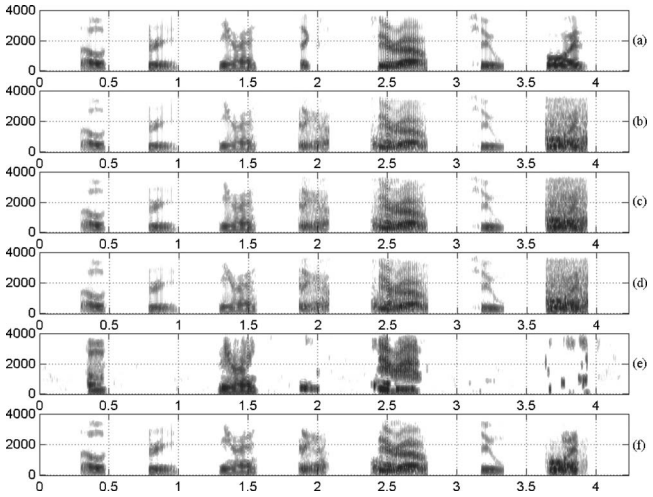

FIG. 14. The efficiency of the MPO method in enhancing the speech signal when the background noise is fluctuating is demonstrated. The digit sequence is "oh three zero six zero two four." Spectrogram of (a) the clean signal, (b) the noisy signal, the speech signal enhanced using the (c) GSS technique, (d) logMMSE technique, (e) Hu-Wang technique, and (f) MPO technique.

TABLE V. Performance of different speech enhancement techniques using various objective measures when the speech signals are corrupted by fluctuating noise.

Type	IS	LAR	LLR	PESQ
GSS	41.301	2.972	0.377	2.030
MMSE	23.473	2.542	0.351	1.974
NC-MMSE	29.975	3.061	0.474	2.024
logMMSE	34.946	2.646	0.370	1.911
Hu-Wang	NaN	16.997	8.250	0.800
APP	43.148	2.952	0.344	2.271
MPO	5.241	2.167	0.276	2.282
MPO-APP	4.971	2.175	0.277	2.320

the speech signals obtained using the GSS enhancement scheme, the logMMSE method, the Hu-Wang method, and the proposed MPO enhancement method, respectively. Notice that the MPO method is able to retain most of the speech information while passing very little noise. The MPO method attenuates the noise in between the spectral peaks of "four" (3.6–3.9 s, local SNR −5 dB) and "zero" (2.5–2.8 s, local SNR 5 dB) while retaining most the spectral peaks. The performance of the different enhancement techniques in terms of the various objective measures on the F-DB database is tabulated in Table V. The IS, LAR, and LLR measures show lower distortion values on MPO-processed speech signals compared to the output of the other enhancement techniques and the PESQ measure shows that the speech signals enhanced using the MPO enhancement scheme have a better perceptual quality than those obtained using the other enhancement schemes. Also, the combined MPO-APP enhancement scheme introduces further improvements in the MPO-enhanced speech signals.

VII. CONCLUSIONS AND FUTURE WORK

We have presented an algorithm for enhancing speech signals corrupted by additive noise. The proposed MPO enhancement scheme alters the components of the PO model in such a way that the basic functionality of the PO model is maintained but the various properties of the model can be analyzed and modified independently of each other. The MPO speech enhancement scheme is based on the fact that speech signals, for the most part, are composed of narrowband signals (i.e., harmonics) with varying amplitudes and that the harmonics that are higher in amplitude are perceptually more significant. Combining the MPO speech enhancement technique with the APP detector further improves its performance by reducing the number of speech deletions and noise insertions. The speech enhancement scheme presented here does not need to estimate the noise characteristics, nor does it assume that the noise satisfies any particular statistical model. The performance of the proposed enhancement scheme, evaluated using different objective measures, is comparable to that of some of the other speech enhancement schemes when the characteristics of the background noise are not fluctuating. The proposed MPO-APP enhancement scheme outperforms other speech enhancement schemes when the speech signals are corrupted by fluctuating noise.

The proposed enhancement scheme is implemented in MATLAB and is about 30 times real time on a typical PC with 2.99 GHz CPU and 2.0 Gbytes of RAM. One of the main factors contributing to the computational cost is the frequency spacing of the MPO structures. In the present work, the MPO structures are spaced every 50 Hz. Preliminary evaluations show that increasing the spacing to 100 Hz drastically reduces the computational cost while degrading the performance only slightly. Several other modules in the implementation of the MPO method can be optimized to reduce the computational cost and will be addressed in the near future. The residual noise passed by the proposed enhancement scheme is usually narrowband in nature and is perceived as musical noise. Work is in progress to propose algorithms to reduce the insertions of the musical noise. The main limitations of the proposed MPO enhancement scheme are its inability to retain turbulent speech sounds and its inability to separate target speech signals from competing speech signals. Evaluations of the MPO enhancement scheme on the task of recognizing speech from a target speaker in the presence of speech from competing speakers show only a slight improvement in the recognition rate especially at low SNRs (Deshmukh and Espy-Wilson, 2006). The MPO-APP processing has to be combined with other speech separation methods to improve the overall performance in such cases. Work is in progress to evaluate the subjective quality of the speech signals enhanced using the proposed MPO-APP scheme. Work is also in progress to evaluate the effectiveness of MPO processing as a preprocessing block for robust speech recognition systems using large databases like the Aurora database and to compare its performance with that of some of the other enhancement schemes.

ACKNOWLEDGMENTS

This work was supported by NSF BCS0236707. The authors would like to thank Dr. Michael Anzalone for helpful discussions on the PO model; Ayanah George for her help in coding the spectral subtraction speech enhancement method proposed in Boll (1979); E. Zavarehei for making the source code for the MMSE, logMMSE, and NC-MMSE publicly available; and J. Hansen and B. Pellom for making the source code for the objective quality evaluations publicly available.

Anzalone, M. C. (**2006**). "Time-frequency gain manipulation for noise-reduction in hearing aids: Ideal and phase-opponency detectors," Ph.D. thesis, Syracuse University.

Beh, J., and Ko, H. (**2003**). "A novel spectral subtraction scheme for robust speech recognition: Spectral subtraction using spectral harmonics of speech," in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Hongkong, pp. 648–651.

Benesty, J., Makino, S., and Chen, J. (**2005**). *Speech Enhancement* (Springer, The Netherlands).

Berouti, M., Schwartz, R., and Makhoul, J. (**1979**). "Enhancement of speech corrupted by additive noise," in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Washington, DC, pp. 208–211.

Boll, S. F. (**1979**). "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust., Speech, Signal Process. **ASSP-27**, 113–120.

Cappe, O. (**1994**). "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," IEEE Trans. Speech Audio Pro-

cess. **2**, 345–349.
Carney, L., Heinz, M. G., Evilsizer, M. E., Gilkey, R. H., and Colburn, H. S. (**2002**). "Auditory phase opponency: A temporal model for masked detection at low frequencies," Acta Acust. **88**, 334–347.
Cheng, Y. M., and O'Shaughnessy, D. (**1991**). "Speech enhancement based conceptually on auditory evidence," IEEE Trans. Signal Process. **39**, 1943–1954.
Cohen, I. (**2004**). "Speech enhancement using a noncausal a-priori SNR estimator," IEEE Signal Process. Lett. **11**, 725–728.
Compernolle, D. V. (**1992**). "DSP techniques for speech enhancement," ESCA tutorial and research workshop on speech processing in adverse conditions, Cannes, France, pp. 21–30.
Deshmukh, O., and Espy-Wilson, C. (**2005**). "Speech enhancement using auditory phase opponency model," in Proceedings of the Eurospeech, pp. 2117–2120, Lisbon, Portugal.
Deshmukh, O., Espy-Wilson, C., Azalone, M., and Carney, L. (**2005a**). "A noise reduction strategy for speech based on phase-opponency detectors," in 149th Meeting of the Acoustical Society of America, Vancouver, Canada.
Deshmukh, O. D. (**2006**). "Synergy of acoustic phonetics and auditory modeling towards robust speech recognition," Ph.D. thesis, University of Maryland, College Park, MD.
Deshmukh, O. D., and Espy-Wilson, C. Y. (**2006**). "Modified phase opponency based solution to the speech separation challenge," in International Conference on Spoken Language Processing, Pittsburgh, PA, pp. 101–104.
Deshmukh, O. D., Espy-Wilson, C. Y., Salomon, A., and Singh, J. (**2005b**). "Use of temporal information: Detection of periodicity, aperiodicity, and pitch in speech," IEEE Trans. Speech Audio Process. **13**, 776–786.
Ephraim, Y., and Malah, D. (**1984**). "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," IEEE Trans. Acoust., Speech, Signal Process. **32**, 1109–1121.
Ephraim, Y., and Malah, D. (**1985**). "Speech enhancement using a minimum mean-square log-spectral amplitude estimator," IEEE Trans. Acoust., Speech, Signal Process. **33**, 443–445.
Gustafsson, H., Nordholm, S. E., and Claesson, I. (**2001**). "Spectral subtraction using reduced delay convolution and adaptive averaging," IEEE Trans. Speech Audio Process. **9**, 799–807.
Hansen, J., and Pellom, B. (**1998**). "An effective quality evaluation protocol for speech enhancements algorithms," in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 2819–2822.
Hansen, J. H., and Nandkumar, S. (**1995**). "Robust estimation of speech in noisy backgrounds based on aspects of the auditory process," J. Acoust. Soc. Am. **97**, 3833–3849.
Hirsch, H. G., and Pearce, D. (**2000**). "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions," in ISCA ITRW ASR2000 Automatic Speech Recognition: Challenges for the Next Millennium, Paris, France, pp. 18–20.
Hohmann, V. (**2002**). "Frequency analysis and synthesis using a gammatone filterbank," Acta Acust. **88**, 334–347.
Hu, G., and Wang, D. L. (**2004**). "Monaural speech separation based on pitch tracking and amplitude modulation," IEEE Trans. Neural Netw. **15**, 1135–1150.
Loizou, P. C. (**2005**). "Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum," IEEE Trans. Speech Audio Process. **13**, 857–869.
McAulay, R. J., and Malpass, M. L. (**1980**). "Speech enhancement using a soft-decision noise suppression filter," IEEE Trans. Acoust., Speech, Signal Process. **ASSP-28**, 137–145.
Mesgarani, N., and Shamma, S. A. (**2005**). "Speech enhancement based on filtering the spectrotemporal modulations," in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, PA, pp. 1105–1108.
Rix, A. W., Beerends, J. G., Hollier, M. P., and Hekstra, A. P. (**2001**). "Perceptual evaluation of Speech Quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," Technical Rep., ITU-T recommendation, P. 862.
Tsoukalas, D. E., Mourjopoulos, J. N., and Kokkinakis, G. (**1997**). "Speech enhancement based on audible noise suppression," IEEE Trans. Speech Audio Process. **5**, 497–514.
Virag, N. (**1999**). "Single channel speech enhancement based on masking properties of the human auditory system," IEEE Trans. Speech Audio Process. **7**, 126–137.
Wang, D. L. (**2005**). *Speech Separation by Humans and Machines*, Chap. 12 (Kluwer Academic, Norwell, MA).