Speech Enhancement Using Modified Phase Opponency Model

Om D. Deshmukh, Carol Y. Espy-Wilson

Institute of Systems Research and Department of Electrical and Computer Engineering,
University of Maryland, College Park, MD USA 20742

omdesh(espy)@umd.edu

Abstract

We previously presented a single-channel speech enhancement technique called the Modified Phase Opponency (MPO) model. The MPO model is based on a neural model called the Phase Opponency (PO) model. The efficacy of the MPO model was demonstrated on speech signals corrupted by additive white noise. In the present work, we extend the MPO model to perform efficiently on speech signals corrupted by additive colored noise with time varying spectral characteristics and amplitude levels. The MPO enhancement scheme outperforms many of the statistical and signal-theoretic speech enhancement techniques when evaluated using three different objective quality measures on a subset of the Aurora database. The superiority of the MPO speech enhancement scheme in enhancing speech signals when the amplitude level and the spectral characteristics of the background noise are fluctuating is also demonstrated.

Index Terms: speech enhancement, auditory modeling, phase opponency.

1. Introduction

A tremendous amount of research has been and continues to be done in the field of developing and implementing speech enhancement algorithms. The algorithms presented vary from signal-theoretic approaches like spectral subtraction [1] and its variations [2], to statistical methods like the Minimum Mean Square-Error Short-Time Spectral Amplitude estimator (MMSE-STSA) [3] and its variations [4, 5]. We previously presented a speech enhancement technique called the Modified Phase Opponency (MPO) model [6, 7]. The MPO model is based on a neural model for detection of tones-in-noise called the Phase Opponency (PO) model [8]. Fig. 1 shows the PO model with Center Frequency (CF) of 900 Hz. The two gammatone filters model two nerve fibers tuned to slightly different frequencies.

As 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. On the other hand when the input is a wideband signal, the output of the two filters will exhibit some degree of correlation and the cross-correlation output will be positive or very slightly negative. Thus the PO model is able to distinguish between narrowband signals and wideband noise. One of the issues with the PO model shown in Fig. 1 is 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 vary either of the two independent of the other.

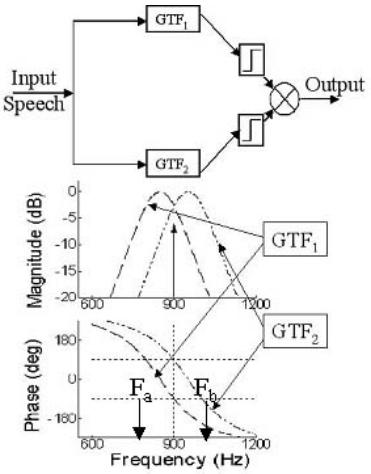

Figure 1: *PO filter pair to detect a tone at 900 Hz. GTF: gammatone filter.*

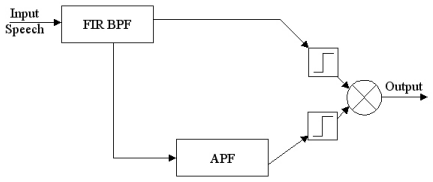

Figure 2: *MPO filter pair; BPF: Bandpass filter; APF: Allpass filter*

The MPO model used in the present work is shown in Fig. 2. A significant difference between the PO and the MPO model is that the latter allows for the control of the relative magnitude response and the relative phase response of the two paths independent of each other. The APF used in one of the paths facilitates the manipulation of the relative phase responses of the two paths without affecting the magnitude responses of the two paths. The dependence of the APF's phase response on its pole characteristics and the relation between the frequency location where the phase response is $-\pi$ (implying the outputs of the two paths are out-of-phase) and the pole characteristics of the APF were presented in [6]. Fig. 3(d) shows the phase response of the APF for a MPO structure designed to detect a narrowband signal with BW

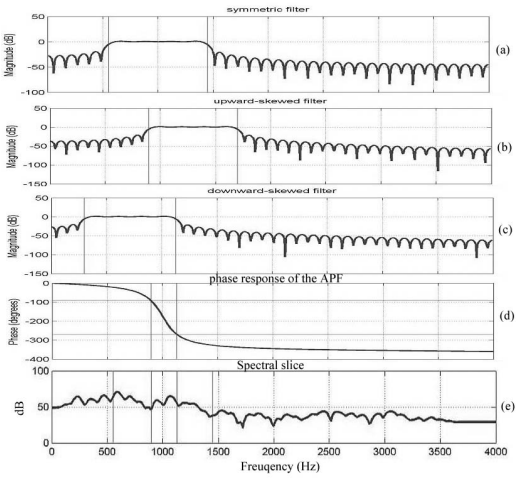

Figure 3: *Magnitude response of the symmetric (a), upward-skewed (b) and downward-skewed (c) BPF that will be used in the MPO structure with CF=1000 Hz. (d) Phase response of the APF that will be used in the MPO structure with CF=1000 Hz. (e) Spectral slice of a sonorant region in speech signal.*

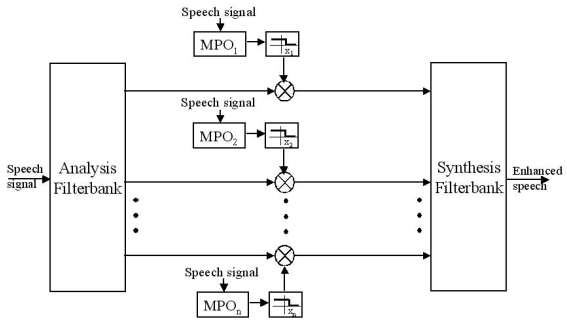

Figure 4: *Schematic of speech enhancement using the MPO model*

no more than 250 Hz and centered at 1000 Hz. The corresponding BPF is chosen such that the passband of the BPF includes some region around 1000 Hz and has a BW such that the MPO structure results in a good separation of narrowband and wideband signals even when the narrowband signal is corrupted by noise. One such BPF is shown in Fig. 3(a). The magnitude response of this BPF is symmetric about 1000 Hz (the CF) and is hence referred to as the symmetric BPF. The threshold to discriminate the presence of signal from the absence of signal was computed using the Maximum Likelihood (ML)-based Likelihood Ratio Test (LRT) with white noise and narrowband signal forming the two classes.

Much of the speech signal is voiced so that it is composed of a combination of narrow band signals (i.e. harmonics) with varying relative amplitudes. The schematic of the MPO speech enhancement scheme is shown in Fig. 4. Each MPO_i in the figure is a MPO structure with a different CF. The CFs are spaced every 50 Hz from 100 Hz to just below the maximum frequency. Only those spectro-temporal regions where the MPO output is below the threshold (indicating presence of signal) are used for reconstruction. Such a structure performs well when the input speech signal is corrupted by additive white noise which has a relatively flat spectrum. But it passes a lot more noise when the corrupting signal is colored noise with fluctuating levels. We now present the improvements made to the MPO model to improve its performance in the presence of colored noise.

2. Improvements to the MPO model

There are two main reasons that warrant deviating from the symmetric BPFs used in the previous versions of the MPO model. Consider the spectral slice shown in Fig. 3(e). The symmetric MPO structure used to detect the F2(around 1000 Hz) consists of the BPF and the APF as showin in Fig. 3(a),(d) respectively. The harmonics close to F2 fall in the *out-of-phase* frequency region of the MPO structure whereas the harmonics close to F1 (around 600 Hz) fall in the *in-phase* frequency region of the MPO structure but also in the passband of the symmetric BPF. The amplitude of F1 (and hence that of the harmonics closer 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 the MPO structure will be positive, thus missing the F2 information. The refined MPO model overcomes this problem by skewing the passband of the filter upwards in frequency with respect to the CF of the MPO as shown in Fig. 3(b). The high amplitude harmonics close to F1 will now be greatly attenuated as they fall in the stopband of the upward-skewed filter. Most of the speech information is correctly detected by the upward-skewed MPO structures as they take advantage of the spectral tilt.

Consider the case where two formants have comparable amplitudes and are close in frequency. The harmonics near these formants also have comparable amplitudes. In such cases, the upward-skewed MPO structures fail to detect the lower frequency harmonics. Downward-skewed MPO structures centered on the lower frequency harmonics, like the ones shown in Fig. 3(c), are able to successfully detect such instances because their *in-phase* region extends on the lower frequency side. In the refined MPO model, each CF is analyzed using an upward MPO structure as well as a downward MPO structure.

As mentioned earlier, the MPO configuration shown in Fig. 2 and Fig. 3(a,d) passes a lot of noise when the noise is colored and/or with fluctuating levels. To overcome this problem, the refined MPO model uses five different MPO structures at each CF. Each of the five MPO structures has a slightly different *out-of-phase* region. Noise can be wrongly seen as speech by one or more of the five different MPO structures, but it is rarely seen as speech by all of the five structures. Similarly, speech signals are almost always seen as speech by *all* of the five structures.

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) at least one frequency channel from all the five different upward skewed or all the five different downward skewed MPO structures should be at least four times more negative than the threshold for that particular channel, indicating a strong presence of speech (b) the temporal region should be at least 50 ms long. The second step detects the frequency channels within the *speech-present* temporal regions where speech information is present by finding the channels where the output from all the five upward skewed or all the five downward skewed MPO structures is below the threshold.

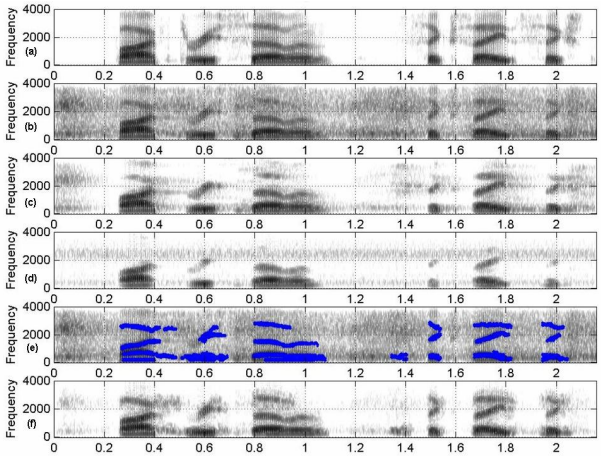

Figure 5: *Spectrogram of: (a) clean speech (b) speech corrupted by subway noise at 10 dB SNR (c) MMSE-STSA-enhanced speech (d) GSS-enhanced speech (e) noisy speech overlaid with the MPO profile. (f) MPO-enhanced speech. X-axis is time in seconds.*

The output of the MPO enhancement scheme can be interpreted as a binary mask [10] with a value of one in the spectro-temporal channels where speech is thought to be more dominant than the noise and a value of zero where the reverse is true. We refer to such a binary mask as the *MPO profile*. Fig. 5(e) shows the spectrogram of the digit sequence 'five three seven six eight six' corrupted by subway noise at 10 dB SNR overlaid with the *MPO profile*. The blue/dark regions are the spectro-temporal channels where the *MPO profile* is one. The noisy speech signal from these channels is used 'as-is' to construct the enhanced speech signal. In the previous versions of the MPO enhancement scheme, the noisy speech signal from the rest of the channels was attenuated by 20 dB before using it for reconstruction. Such a scheme leads to sharp spectral discontinuities by increasing the depth of the spectral valleys, especially at higher SNRs. In the present MPO enhancement technique, the weighing scheme is based on the transfer function associated with a conjugate pair of poles corresponding to the centroid of the frequencies of the contiguous speech-present channels. The transfer function is similar to the general form of the vocal tract transfer function derived in [9]:

$$T_n(s) = \frac{s_n s_n^*}{(s - s_n)(s - s_n^*)} \tag{1}$$

where $s = j2\pi f$, s_n is the complex frequency of the pole, and $s_n = \sigma_n + j2\pi F_n$. The value of σ_n is chosen such that the BW of the pole is 100 Hz. The weighing scheme corresponding to the frame centered at 825 ms of the utterance shown in Fig. 5 is displayed in Fig. 6. The F_n values for this frame are: 550, 1500 and 2750 Hz. The signal in the *speech absent* frames is uniformly attenuated by 20 dB.

3. Results

The salient features of the MPO-based speech enhancement are that (1) it makes minimal assumptions about the noise characteristics (the only assumption is that noise is broader than the harmonics of the speech signal) and (2) the noise estimates from the previous frames do not play any role in the noise removal techniques in the later frames.

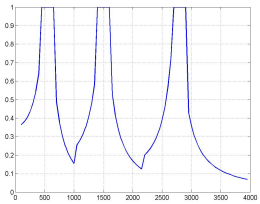

Figure 6: *Spectral weighing scheme. X-axis is frequency in Hz.*

The performance of the MPO enhancement scheme was evaluated on a subset of the Aurora database [11] and compared with some of the other enhancement schemes like: the MMSE STSA [3] and two of its variations [4, 5], and the Generalized Spectral Subtraction (GSS) [2] method. Fig. 5 shows the spectrograms of a connected digits utterance 'five three seven six eight six' (a) in clean, (b) corrupted by subway noise at 10 dB SNR, (c) MMSE-STSA-enhanced (d) GSS-enhanced and (f) MPO-enhanced speech signal. It is evident from the figure that the MPO scheme strikes a better balance between the amount of noise removed and the amount of speech information retained compared to the other methods. For example, the MPO scheme is able to retain the (weak) F3 information at the end of the vowel near 0.6 sec, near 1.55 sec and again near 2 sec, while passing very little noise.

The performance of the different enhancement schemes was also compared using the Linear Predictive (LP) coefficients based objective quality measures like (1) the Itakura-Saito distortion measure (IS), (2) the Log-Area-Ratio (LAR) measure and (3) the Log-Likelihood Ratio (LLR) measure. Tables 1, 2 and 3 compare the increase in the IS, LAR and LLR measures as the SNR drops from ∞ dB to 20, 10 and 5 dB. Note that as the SNR reduces, the MPO-enhanced speech shows the lowest increase in the distortion values especially at lower SNRs. The distortion values for MPO-processed speech are relatively high in clean (i.e. when clean unprocessed speech is compared with MPO-processed clean speech) as the MPO processing attenuates the spectral valleys in the speech signal thus increasing the dissimilarities between the LP coefficients computed on clean speech and the LP coefficients computed on the MPO-processed clean speech. The second reason for the dissimilarities is that the MPO processing does not maintain most of the obstruent information. Fig. 7(a,b) shows that the MPO processing retains almost all of the sonorant speech information when operating on clean speech.

Table 1: Increase in the IS distortion measure as the SNR reduces

type	clean	20 dB	10 dB	5 dB
MMSE[3]	0.353	0.597	2.001	3.473
MMSE-logSTSA[4]	0.721	1.416	5.776	14.839
MMSE-logSTSA-SNR[5]	0.285	0.820	4.690	18.747
GSS[2]	0.959	3.446	3.993	3.010
MPO	3.056	2.697	1.157	2.759

Fig. 7 demonstrates the efficiency of the MPO method in enhancing the speech signal when the level and the type of the background noise are varying over time. Fig. 7(a) and 7(b) show the spectrogram of the clean utterance and MPO-processed clean ut-

Table 2: Increase in the LAR measure as the SNR reduces

type	clean	20 dB	10 dB	5 dB
MMSE[3]	0.923	1.656	3.446	4.549
MMSE-logSTSA[4]	1.089	1.913	3.880	4.920
MMSE-logSTSA-SNR[5]	0.760	1.516	3.559	4.987
GSS[2]	2.186	2.156	3.294	3.905
MPO	3.164	1.535	1.937	3.226

Table 3: Increase in the LLR measure as the SNR reduces

type	clean	20 dB	10 dB	5 dB
MMSE[3]	0.071	0.211	0.551	0.855
MMSE-logSTSA[4]	0.094	0.243	0.601	0.912
MMSE-logSTSA-SNR[5]	0.054	0.205	0.579	0.923
GSS[2]	0.116	0.318	0.635	0.898
MPO	0.425	0.333	0.430	0.713

terance respectively. As is obvious from the figure, MPO processing loses very little of the sonorant information when the input is clean speech. Listening tests confirm that there is very little difference between the clean speech and MPO-processed clean speech. Fig. 7(c) shows the spectrogram of the noisy speech where each digit is corrupted by one of the three different noise types: (a) subway (b) exhibition hall and (c) car noise at different SNRs. The digit sequence is 'nine four two eight five six' and the SNR order is 5, 20, 0, 15, -5, 10 dB. Fig. 7(d-g) show the spectrograms of the speech signal enhanced by the MMSE-STSA method, the MMSE-logSTSA-SNR method, the GSS method and the MPO method respectively. Notice that the MPO method is able to enhance the weak higher formants (around 700 ms and 2000 ms) while passing very little noise. It is also worth pointing out that only the MPO method was able to process the digit corrupted by the most noise (digit 'five' at -5 dB SNR) whereas the other methods pass it with little or no enhancement. Informal listening tests reveal that the MPO enhanced speech is of better quality than the output of the other methods.

In a companion paper [12] submitted as part of the speech separation challenge, we show that the accuracy of automatic speech recognition increases when the noisy speech signals are replaced by MPO-enhanced speech signals.

Work is in progress to study the phenomenon of musical noise and to propose algorithms to eliminate or minimize the musical noise effect and to evaluate the performance of the MPO-processed speech on robust speech recognition using the Aurora database.

4. Acknowledgments

This work was supported by NSF grant BCS0236707. The authors thank E. Zavarehei for making the source code for the MMSE-STSA, MMSE-logSTSA and MMSE-logSTSA-SNR publicly available, A. George for help in coding the SS-based enhancement method, and J. Hansen and B. Pellom for making the source code for the objective quality evaluations publicly available.

5. References

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction", IEEE Trans. on Acous, Speech and Signal Proc., ASSP-27(2), 113–120, 1979.

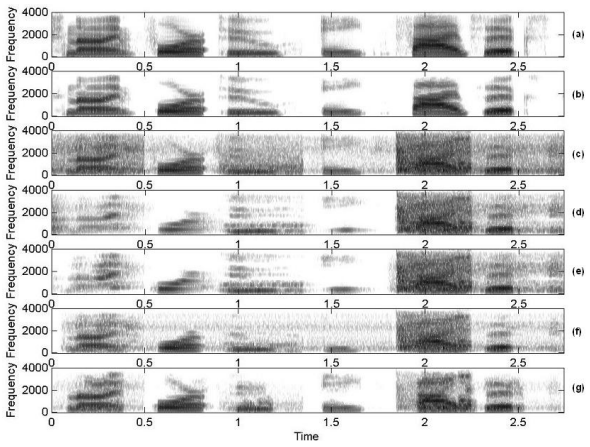

Figure 7: *Spectrograms of (a) clean speech (b) MPO-processed clean speech (c) speech corrupted by noise with time varying amplitude and spectral characteristics (d) MMSE-STSA-enhanced speech (e) MMSE-logSTSA-SNR-enhanced speech (f) GSS-enhanced speech (g) MPO-enhanced speech*

[2] Compernolle D., "DSP techniques for speech enhancement", ETRW, 10–13, 1992.

[3] Ephraim Y. and Malah D., "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator", IEEE Trans. on Acous. Speech and Signal Proc., ASSP-32(6), 1109–1121, 1984.

[4] Ephraim Y. and Malah D., "Speech enhancement using a minimum mean-square log-spectral amplitude estimator", IEEE Trans. Acous. Speech and Signal Proc., ASSP-33(2), 443–445, 1985.

[5] Cohen I., "Speech enhancement using a noncausal a priori SNR estimator", IEEE Sig. Proc. Let., 11(9),725–728 ,2004.

[6] Deshmukh, O., Espy-Wilson C., "Speech Enhancement Using Auditory Phase Opponency Model", Proc. Eurospeech, 2117–2120, 2005.

[7] Deshmukh O., Anzalone M., Espy-Wilson C., Carney L., "A noise reduction strategy for speech based on phase-opponency detectors", 149th Meeting of the ASA, 2005.

[8] Carney et. al., 'Auditory phase opponency: A temporal model for masked detection at low frequencies', Acta Acustica (88), 334–347, 2002

[9] Stevens K., "Acoustic Phonetics", M.I.T. Press, Cambridge, 1999

[10] Wang D.L., "On ideal binary mask as the computational goal of auditory scene analysis.", in Divenyi P. (ed.), Speech Separation by Humans and Machines, Kluwer Academic, Norwell, 181–197, 2005

[11] Hirsch H., and Pearce D., "The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions", ISCA ITRW, 18–20, 2000.

[12] Deshmukh O., Espy-Wilson C., "Modified phase opponency based solution to the speech separation challenge", Proc. Interspeech 2006.