

Modified Phase Opponency Based Solution To The Speech Separation Challenge

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

In this work, we present a single-channel speech enhancement technique called the Modified Phase Opponency (MPO) model as a solution to the Speech Separation Challenge. The MPO model is based on a neural model for detection of tones-in-noise called the Phase Opponency (PO) model. Replacing the noisy speech signals by the corresponding MPO-processed signals increases the accuracy by 31% when the speech signals are corrupted by speech-shaped noise at 0 dB Signal-to-Noise Ratio (SNR). It is worth mentioning that the MPO enhancement scheme was developed using the noisy connected-digit Aurora database and was not tailored in any way to fit the Grid database used in this challenge. One of the salient features of the MPO-based speech enhancement scheme is that it does not need to estimate the noise characteristics, nor does it assume that the noise satisfies any statistical model.

Index Terms: speech separation, robust speech recognition.

1. Introduction

Various different approaches are being pursued to make Automatic Speech Recognition (ASR) systems robust to noise. One of the approaches is to enhance the speech signal by suppressing the noise as much as possible while leaving most of the actual speech content undistorted. Many of the speech enhancement techniques were originally developed for speech quality improvement, but they can also be used as a pre-processing block for ASR systems.

We previously presented a speech enhancement technique called the Modified Phase Opponency (MPO) model [1, 2]. The MPO model is based on a neural model for detection of tones-in-noise called the Phase Opponency (PO) model [3]. 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

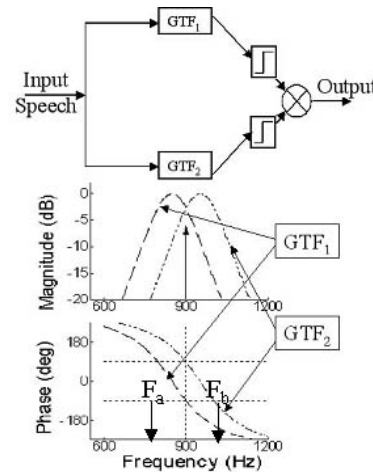


Figure 1: PO filter pair to detect a tone at 900 Hz. GTF: gammatone filter.

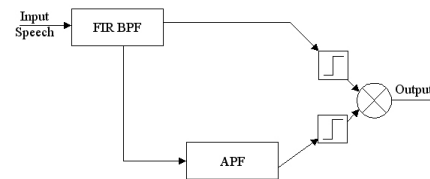


Figure 2: MPO filter pair; BPF: Bandpass filter; APF: Allpass filter

vary either of the two independent of the other.

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 All Pass Filter (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 will be out-of-phase) and the pole characteristics of the APF were presented in [1]. Fig. 3(d) shows the phase response of the APF for a MPO structure designed to detect a narrowband signal with bandwidth no more than 250 Hz and centered at CF

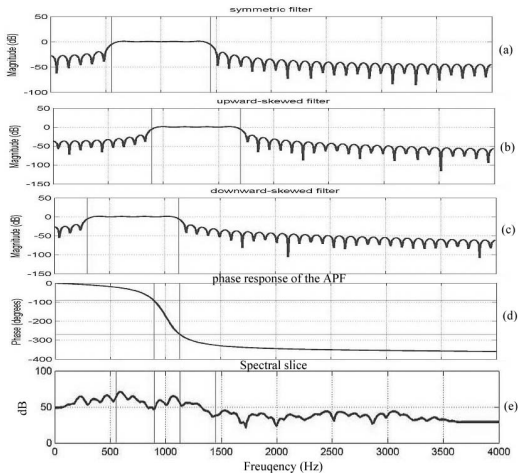


Figure 3: Magnitude response of the symmetric (a), upward-skewed (b) and downward-skewed (c) BPF that will be used in the MPO structure with $CF=1000$ Hz. (d) Phase response of the APF that will be used in the MPO structure with $CF=1000$ Hz. (e) Spectral slice of a sonorant region in speech signal.

= 1000 Hz. The corresponding BPF is chosen such that the pass-band of the BPF includes some region around the CF 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.

Fig. 3(a-c) show three possible BPFs. The magnitude response of the BPF in Fig. 3(a) is symmetric about the CF and is referred to as the symmetric BPF. The magnitude response of the BPFs in Fig. 3(b,c) are skewed upward and downward in frequency with respect to the CF and are hence referred to as the upward-skewed and downward-skewed BPF, respectively. The MPO structures using the upward-skewed BPFs are more reliable in detecting speech information as they take advantage of the spectral tilt present in sonorant speech regions. For example, the F2 information (around 1000 Hz) in the spectral slice shown in Fig. 3(e) can be detected only using the upward-skewed MPO structures.

Downward-skewed MPO structures are useful in detecting the low frequency harmonics corresponding to lower formants when two formants are of comparable amplitudes and are close in frequency. 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.

2. MPO based speech enhancement scheme

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-based speech enhancement scheme is shown in Fig. 4. The CFs are spaced every 50 Hz from 100 Hz to just below the maximum frequency. Each MPO_i in Fig. 4 consists of five upward-skewed and five downward-skewed MPO structures, all tuned to the same CF, but with slightly different *out-of-phase* regions. Noise can be wrongly seen as speech by one or more of the five upward/downward-

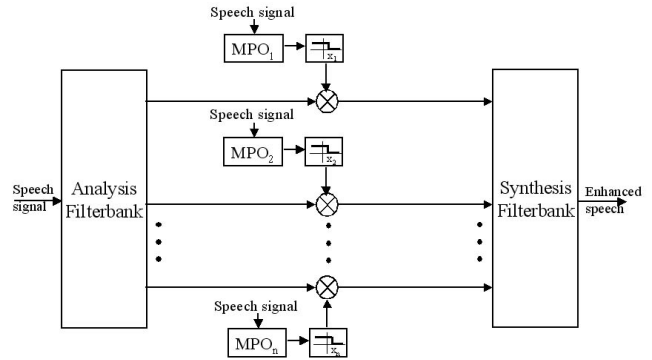


Figure 4: Schematic of speech enhancement using the MPO model

skewed MPO structures, but it is rarely seen as speech by all the five upward/downward-skewed structures. Similarly, speech is almost always seen as speech by *all* of the five structures.

The speech enhancement scheme can 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 signal and (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.

The output of the MPO enhancement scheme can be interpreted as a binary mask [5] 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 noise is thought to be more dominant than the speech signal. We refer to such a binary mask as the *MPO profile*. Fig. 5 show the spectrogram of the clean speech and the MPO-processed clean speech respectively. Fig. 5(b,d-g) show the spectrograms of the utterance 'bin white at S 1 soon' in clean as well as the noisy versions at 6, 0, -6 and -12 dB SNR, respectively, overlaid with the corresponding *MPO profiles*. The blue/dark regions are the spectro-temporal channels where speech signal is thought to be present. Note that the *MPO profile* of the clean utterance has a value of one in almost all the high energy sonorant regions thus indicating that very little information is lost by MPO processing when the input signal is clean. This is also confirmed by the ASR results as is discussed in the next session. The *MPO profile* at 6 dB and 0 dB SNR (Fig. 5(d),e)) show that a lot of the speech information is retained and that as the SNR reduces the low energy speech information is gradually lost. It can also be deduced that the MPO processing is very conservative in what gets called as speech (i.e. MPO processing makes more false-rejection errors).

The noisy speech signal from the channels where the *MPO profile* is one is used 'as-is' to construct the enhanced speech signal. The contribution of the rest of the channels is decided by a weighing scheme 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

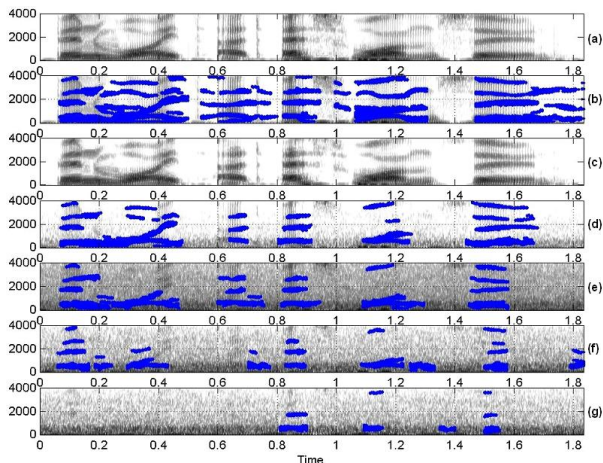


Figure 5: Spectrograms of (a) clean speech (b) clean speech overlaid with the MPO profile. (c) MPO-processed clean speech. (d)-(g) show the spectrograms of noisy speech at 6, 0, -6, -12 dB SNR respectively overlaid with the corresponding MPO profile. X-axis is time in seconds. Y-axis is frequency in Hz.

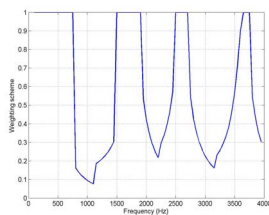


Figure 6: Spectral weighing scheme.

function is similar to the general form of the vocal tract transfer function derived in [4]:

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

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 bandwidth of the pole is 100 Hz. Such an attenuation scheme reduces the perceptual artifacts introduced by the enhancement technique. The weighing scheme corresponding to the frame centered at 100 ms of the clean utterance shown in Fig. 5(b) is displayed in Fig. 6. The F_n values for this frame are: 400, 1800, 2600 and 3650 Hz.

3. Evaluations

The performance of the MPO speech enhancement scheme is evaluated by conducting robust speech recognition experiments on the Grid database [6]. The database consists of sentences of the form: “<command:4> <color:4> <preposition:4> <letter:25> <digit:10> <adverb:4>”, where the number indicates the number of choices at each point. The training subset consists of 17000 sentences (500 each from 34 talkers) recorded in clean environment at a sampling rate of 25 kHz. The speech signals in the test subset are corrupted either by speech-shaped noise at ∞ dB, 6 dB, 0 dB, -6 dB or -12 dB SNR or by a different sentence at ∞ dB, 6

dB, 3 dB, 0 dB, -3 dB, -6 dB or -12 dB Target-to-Masker Ratios (TMRs). The test data consists of 900 utterances in each condition, 300 of which form the development test set. The same talkers are used for training, development and test sets but no training utterance occurs in either the development or the test set. Finally, as per the rules of the challenge, the recognition results on only the color, letter and digit keywords are used to compute the final recognition accuracy.

It is worth mentioning that the MPO enhancement scheme was developed using the noisy connected-digits Aurora database [7] and was not tailored in any way to fit the Grid database. One of the salient features of the MPO-based speech enhancement scheme is that it does not need to estimate the noise characteristics nor does it assume that the noise satisfies any statistical model.

The aurora database is sampled at 8 kHz and hence the *MPO profile* ranges from 0–4 kHz. The *MPO profile* computed on 0–4 kHz can be applied to the 25 kHz-sampled Grid database in one of the following three ways: (a) downsample the Grid database to 8 kHz and apply the *MPO profile* as-is, (b) apply the *MPO profile* from 0–4 kHz and pass the high frequency information as-is (i.e. set the *MPO profile* to one for all spectro-temporal channels with CF > 4 kHz) or (c) apply the *MPO profile* from 0–4 kHz and suppress the high frequency information (i.e. set the *MPO profile* to zero for all spectro-temporal channels with CF > 4 kHz). These three different methods are referred to as MPO_{4k} , MPO_{hon} and MPO_{hoff} respectively and results are presented for each of these methods as well as for the ‘no-processing’ case where the noise test data is used without any processing. For each of the three methods, the recognizer was trained using the same parameters and topology as the default recognizer had but the training utterances were processed using the respective method to obtain matched conditions for training and testing. For example, the training utterances used to train the MPO_{hon} recognizer were MPO-processed with the corresponding *MPO profile* applied from 0–4 kHz and the high frequency information passed as-is.

Table 1: Recognition accuracy for speech-shaped-noise condition

Type	clean	6 dB	0 dB	-6 dB	-12 dB
no proc	98.56	56.67	18.94	11.78	11.67
MPO_{hon}	97.89	73.67	40.67	19.11	13.28
MPO_{hoff}	96.44	71.06	41.94	18.72	14.50
MPO_{4k}	96.00	73.83	50.06	26.00	14.33

Table 1 shows the recognition accuracy on the test set when the speech signals are corrupted by speech shaped noise at various SNRs. The row corresponding to ‘no-processing’ shows the baseline results obtained using the noisy test utterances. These results are also plotted in Fig. 7. It is evident from the figure that all of the three ways in which the *MPO profile* is applied to the test set result in an improvement in accuracy. The results obtained in the clean condition with either of the three methods are very similar to the ones obtained in ‘no-processing’ condition implying that the MPO-processing retains most of the speech information when the input is clean speech. The slight drop in accuracy (from 98.56% to about 97%) could be because of the fact that MPO-processing removes most of the obstruent information since it consists of turbulent noise. The MPO_{4k} processing leads to an increase in the accuracy of about 31% at 0 dB SNR. Similar results were obtained for the development set. This proves that the MPO enhancement scheme is effective in improving the performance of robust auto-

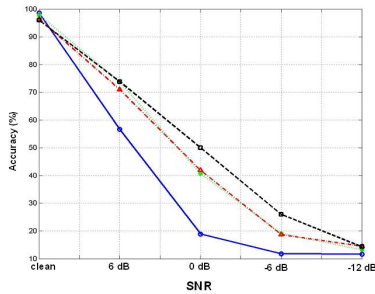
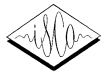


Figure 7: Recognition accuracy for speech corrupted by the speech-shaped-noise. blue solid curve with o : no processing, green dotted curve with $*$: MPO_{hon} , red dash-dotted curve with \triangle : MPO_{hoff} , black dashed curve with \square : MPO_{4k}

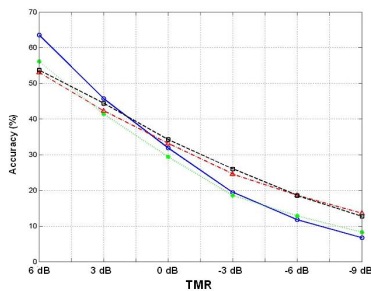


Figure 8: Recognition accuracy for speech corrupted by the two-talker scenario. blue solid curve with o : no processing, green dotted curve with $*$: MPO_{hon} , red dash-dotted curve with \triangle : MPO_{hoff} , black dashed curve with \square : MPO_{4k}

matic speech recognition when the speech signals are corrupted by speech shaped noise.

Table 2: Recognition accuracy for two-talker condition

Type	6 dB	3 dB	0 dB	-3 dB	-6 dB	-12 dB
no proc	63.58	45.75	31.92	19.42	11.75	6.75
MPO_{hon}	56.17	41.42	29.42	18.58	12.83	8.25
MPO_{hoff}	53.08	42.33	33.17	24.58	18.67	13.58
MPO_{4k}	53.75	44.42	34.25	26.00	18.58	12.75

Table 2 shows the recognition accuracy on the test set when the speech signals are corrupted by other competing utterances at various TMRs. The row corresponding to 'no-processing' shows the baseline results obtained using the noisy test utterances. These results are also plotted in Fig. 8. The figure shows that MPO-processing leads to a slight drop in the accuracy at positive SNRs and a slight increase in the accuracy at negative SNRs. These results are not surprising as the corrupting noise in this case is a competing speech signal which is also narrowband. In this case the MPO-processing will retain both the target speech signal as well as the masking signal. The results for the clean two-talker condition are very similar to the results for the clean speech shaped noise condition. The results for the two-talker case can be categorized further based on whether the talker and the masker are the same,

have the same gender or have different genders. These results are tabulated for the MPO_{4k} case in table 3. MPO-processing does not favor any one category over the others as the interfering noise in all the categories is still narrowband.

Table 3: Categorized recognition results for two-talker condition

SNR	same talker	same gender	diff gender	average
6dB	52.94	55.59	53.00	53.75
3dB	44.34	45.81	43.25	44.42
0dB	30.54	35.75	37.00	34.25
-3dB	24.43	27.37	26.50	26.00
-6dB	16.29	20.11	19.75	18.58
-9dB	11.31	14.53	12.75	12.75

MPO-processing has to be combined with other speech separation methods (e.g. computational auditory scene analysis based methods) to improve the overall performance when the corrupting noise is narrow-band. Work is in progress to study and propose different methods than can help retain the obstruent information and to evaluate the performance of the MPO-processed speech on robust speech recognition using the Aurora database.

In a companion paper [8] we demonstrate that 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.

4. Acknowledgments

This work was supported by NSF grant BCS0236707.

5. References

- [1] Deshmukh, O., Espy-Wilson C., "Speech Enhancement Using Auditory Phase Opponency Model", Proc. Eurospeech, 2117–2120, 2005.
- [2] 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.
- [3] Carney et. al., 'Auditory phase opponency: A temporal model for masked detection at low frequencies', Acta Acustica (88), 334–347, 2002
- [4] Stevens K., "Acoustic Phonetics", M.I.T. Press, Cambridge, 1999
- [5] 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
- [6] Cooke, M., Barker, J., Cunningham, S. and Shao, X., "An audio-visual corpus for speech perception and automatic speech recognition", submitted to J. Acoust. Soc. Amer.
- [7] 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.
- [8] Deshmukh O., Espy-Wilson C., "Speech Enhancement Using Modified Phase Opponency Model", Proc. Interspeech 2006.