# Speech Enhancement Using Auditory Phase Opponency Model

Om Deshmukh, Carol Espy-Wilson

Electrical and Computer Engineering Department and Institute for Systems Research
University of Maryland,College Park, USA
omdesh(espy)@Glue.umd.edu

## Abstract

In this work we address the problem of single-channel speech enhancement when the speech is corrupted by additive noise. The model presented here, called the Modified Phase Opponency (MPO) model, is based on the auditory PO model, proposed by Carney et. al., 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. Initial evaluation of the MPO model on speech corrupted by white noise at different SNRs shows that the MPO model is able to enhance the spectral peaks while suppressing the noise-only regions.

## 1. Introduction

Several studies in the past have compared the performance of human speech perception in noise with that of the ASR systems. A detailed comparison of performance of human speech perception to that of the ASR can be found in [1]. These studies show that there is a wide gap between human performance and machine performance, especially in degraded conditions. Moreover, ASR systems that perform well in one kind of background noise typically fail to maintain the performance when tested on a different kind of disturbance. This difference in performance has fueled a variety of research to develop and implement algorithms for speech enhancement and robust speech recognition.

Some of the signal-theoretic approaches to speech enhancement include spectral subtraction [2], Ephraim-Mallah Minimum Mean Square-Error Short-Time Spectral Amplitude estimator (MMSE-STSA) [3], adaptive Kalman filtering [4], signal subspace approaches [5], wavelet-packet transforms [6] and multi-taper spectrum estimator [7].

Speech enhancement techniques using models of human auditory systems have also been proposed. In [8, 9], models of Lateral Inhibition Network (LIN) [10], a biological neural network thought to exist in the auditory system, are used for speech enhancement. A speech enhancement technique using a model of multiscale representation of speech modulations is proposed in [11]. The multiscale model is inspired by psychoacoustical and neurophysiological findings in the early and central stages of the auditory pathway.

The auditory model used in this work is motivated by the auditory Phase Opponency (PO) model [12] and is described briefly in the next section.

## 2. Phase Opponency model

PO model is a model for detection of tone-in-noise based on processing the information in neural discharge times . 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 responses of some of the cross-channel coincidence detectors are reduced when a tone is added to a noise. This reduction in response in the presence of the target is referred to as phase opponency.

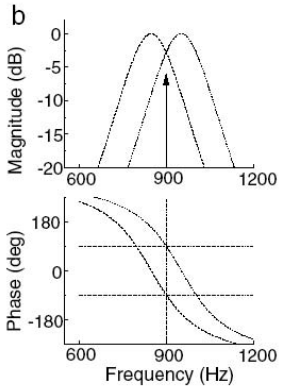


Figure 1: *PO filter pair to detect a tone at 900 Hz.*

Consider a case where the aim is to detect a tone at frequency $\omega_0$. The PO model for this case will consist of models of two auditory neurons, modeled as gammatone filters, that have considerable overlap in their passbands and the frequency of interest, $\omega_0$, lies in their common passbands. The order and the Center Frequencies (CFs) of the two filters are chosen such that the phase response of the two filters differ by about $180^\circ$ around the frequency $\omega_0$ (see Fig. 1). The response of both the neurons will be synchronized to the tone. However, the responses will be out-of-phase and a cross-frequency coincidence will result in a negative output. When broad band noise is used as input, the output of the two neurons will be partially correlated and the cross-frequency coincidence will result in a positive output.

## 3. Modified PO model

In the Modified PO (MPO) model, one channel (i.e. neuron) of the PO filter pair is modeled as a linear-phase Finite Impulse Response (FIR) Band Pass Filter (BPF). The other channel is modeled as a concatenation of the same FIR BPF followed by an All Pass Filter (APF). The filters are followed by a signum non-linearity to minimize the effect of magnitude information in the cross-frequency coincidence. This structure is shown in Fig. 2. The characteristics of the BPF are mainly decided by the range of the target frequency that is to be detected. The

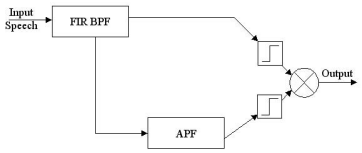


Figure 2: *Modified PO filter pair*

characteristics of the APF are mainly decided by the expected frequency range and bandwidths of the target signals. Consider an 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. The magnitude response is 1 for all values of $\omega$ and the phase response, $\Phi(\omega)$, is given by:

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

The frequency region where the phase response,$\Phi(\omega)$, is steepest can be located by finding the frequency where the slope of the phase response has an inflection point, i.e. by finding the $\omega$ for which $d^2(\Phi(\omega))/d\omega^2 = 0$. For simplicity, let us first compute $d^2(\Phi(\omega))/d\omega^2$ for just one pole, $a$, and then account for the complex conjugate pole $a^*$.

Taking the derivative w.r.t to $\omega$ on both sides of eq (1) and simplifying, we have:

$$\frac{d(\Phi(\omega)}{d\omega} = -1 - 2\left[\frac{rcos(\omega - \theta) - r^2}{1 - 2rcos(\omega - \theta) + r^2}\right] \quad (2)$$

The second order derivative is then given by (after simplification):

$$\frac{d^2(\Phi(\omega)}{d\omega^2} = -2\left[\frac{(r^3 - r)sin(\omega - \theta)}{(1 - 2rcos(\omega - \theta) + r^2)^2}\right] \quad (3)$$

If we factor in the effect of the complex conjugate pole in eq (3), we have:

$$\begin{aligned}\frac{d^2(\Phi(\omega)}{d\omega^2} &= -2\left[\frac{(r^3 - r)sin(\omega - \theta)}{(1 - 2rcos(\omega - \theta) + r^2)^2}\right] \\ &\quad -2\left[\frac{(r^3 - r)sin(\omega + \theta)}{(1 - 2rcos(\omega + \theta) + r^2)^2}\right] \quad (4)\end{aligned}$$

Since we are only interested in finding the value of $\omega$ for which the above equation (4) becomes zero, the denominator can be ignored. The numerator can be simplified as:

$$\begin{aligned}N(\omega) &= -2(r^3 - r)[(2 + 4r^2 + 2r^4)sin\omega cos\theta \\ &\quad -(8r + 8r^3)sin\omega cos\omega \\ &\quad +8r^2(cos^2\omega + sin^2\theta)sin\omega cos\theta]\end{aligned}$$

Equating $N(\omega)$ to zero and simplifying implies,

$$\begin{aligned}N(\omega) &= 0; \\ &\Longrightarrow D(r,\omega)cos\theta = cos\omega \quad (5)\end{aligned}$$

$$\begin{aligned}Where,& \\ D(r,\omega) &= \left[\frac{1 + 2r^2 + 4r^2(cos^2\omega + sin^2\theta) + r^4}{4r(1 + r^2)}\right]\end{aligned}$$

Table 1: *Dependence of $D(r,\omega)$ on $r$*

| $r$ | $D(r,\omega)$ |
|---|---|
| 0.750 | 1.0008 |
| 0.800 | 1.0003 |
| 0.850 | 1.0001 |
| 0.900 | 1.0000 |
| 0.950 | 1.0000 |
| 1.000 | 1.0000 |

Notice, from equation (5), that the sum of the coefficients in the numerator of $D(r,\omega)$ $(1+2+4+1 == 8)$ is exactly equal to that of the coefficients in the denominator $(4 * (1+1) == 8)$. Also notice that the $cos\theta$ term on the l.h.s. is balanced by the $cos\omega$ term on the r.h.s. Thus, the equality in eq (5) holds for $\theta = \omega$ and $r = 1$. But stability of the APF dictates that the magnitude of $r$ be less than 1. Table 1 shows that $D(r,\omega)$ is very close to one 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 $a = re^{j\theta}$ and $a^*$ 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 $\theta = \omega$ is given by:

$$\Phi(\omega) - 2\theta - 2tan^{-1}\left[\frac{2rsin\theta cos\theta}{1 - r + 2rsin^2\theta}\right] \quad (6)$$

If $r \approx 1$, then $r - 1 \approx 0$ and the above equation is further simplified to:

$$\Phi(\omega) \approx -2\theta - 2tan^{-1}(cot\theta)$$

$$\Phi(\omega) \approx \begin{cases} -2\theta - 2[-\frac{1}{2}\pi - cot^{-1}(cot\theta)] & \text{if } cot\theta < 0 \\ -2\theta - 2[\frac{1}{2}\pi - cot^{-1}(cot\theta)] & \text{if } cot\theta > 0 \end{cases}$$

$$\Phi(\omega) \approx \begin{cases} \pi & \text{if } cot\theta < 0 \\ -\pi & \text{if } cot\theta > 0 \end{cases} \quad (7)$$

*The phase response at $\theta = \omega$ 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 slope of $\Phi(\omega)$ at $\omega = \theta$ in terms of $r$ and $\theta$. From equation (2), we know that the derivative of $\Phi(\omega)$ w.r.t $\omega$ is given by:

$$\begin{aligned}\frac{d(\Phi(\omega)}{d\omega} &= -1 - 2\left[\frac{rcos(\omega - \theta) - r^2}{1 - 2rcos(\omega - \theta) + r^2}\right] \\ &\quad -1 - 2\left[\frac{rcos(\omega + \theta) - r^2}{1 - 2rcos(\omega + \theta) + r^2}\right] \quad (8)\end{aligned}$$

Using $\omega = \theta$ and simplifying leads to:

$$\left.\frac{d(\Phi(\omega)}{d\omega}\right|_{\omega=\theta} = -2\left[\frac{1 - 2rcos^2\theta + r^2}{(1 - r^2)sin^2\theta}\right] \quad (9)$$

The above equation is evaluated for various values of $\theta$ and $\omega$ and is tabulated in Table 2. Notice that, for a given value of $r$, the value of $d(\Phi(\omega))/d\omega$ is not very sensitive to the value of $\theta$. 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. the frequency where the phase response is the steepest ) is independent of $\theta$ and is dependent only on the value of $r$.* The *actual* value of this derivative is of little practical significance as $tan^{-1}$ is a highly compressing nonlinearity.

Table 2: *Dependence of* $\Phi'(\omega) \triangleq d(\Phi(\omega))/d\omega$ *on* $r$ *and* $\theta$.

| | r=0.80 | r=0.85 | r=0.90 | r=0.95 |
|---|---|---|---|---|
| $\theta$ | $\Phi'(\omega)$ | $\Phi'(\omega)$ | $\Phi'(\omega)$ | $\Phi'(\omega)$ |
| 0.393 | -10.41 | -13.36 | -19.67 | -39.32 |
| 0.643 | -9.51 | -12.70 | -19.24 | -39.12 |
| 0.893 | -9.26 | -12.52 | -19.12 | -39.06 |
| 1.143 | -9.16 | -12.45 | -19.07 | -39.04 |
| 1.393 | -9.12 | -12.42 | -19.06 | -39.03 |
| 1.643 | -9.11 | -12.42 | -19.05 | -39.03 |
| 1.893 | -9.14 | -12.43 | -19.06 | -39.03 |
| 2.143 | -9.20 | -12.48 | -19.10 | -39.05 |
| 2.393 | -9.37 | -12.60 | -19.17 | -39.08 |
| 2.643 | -9.86 | -12.96 | -19.41 | -39.20 |

Consider a situation where we have to design a MPO model to detect bandlimited signals centered at $\omega_c$ and of bandwidths less than or equal to $\Delta\omega$. Let us first compute the parameters of the APF and then decide the parameters of the BPF. From the above analysis we know that the pole, $a = re^{j\theta}$, of the APF should be such that $\theta = \omega_c$. The value of $r$ is dictated by the bandwidth $\Delta\omega$. Our aim is to use a value of $r$ such that the phase response, $\Phi(\omega)$, of the APF spans $-\pi/2$ to $-3\pi/2$ in $\Delta\omega$ radians centered around $\omega_c$ (i.e. the negative phase response region is spanned over the expected frequency range of input signal). This is feasible because, as equation (7) showed, the phase response at $\theta = \omega_c$ is approximately equal to $-\pi$ and the phase response is a continuous and monotonous function of $\omega$.

Equation (9) can be thought of as a linear approximation to the relation between $r$ and the phase response at $\omega_c$. But this relation cannot be extended for $\omega$ values far from $\omega_c$ as the phase response function (i.e. $tan^{-1}$) is highly nonlinear. Thus, the value of $r$, such that the phase response spans $-\pi/2$ to $-3\pi/2$ in $\Delta\omega$ radians centered around $\omega_c$, needs to be found using empirical experiments. Assume that the optimal value of $r$ was found to be $r = r_c$ and the frequencies at which the phase response is $-\pi/2$ and $-3\pi/2$ to be $\omega_1$ and $\omega_2$ respectively. We need to now decide the parameters of the FIR BPF. The passband of this BPF should clearly include the frequency range $\omega_1$ to $\omega_2$. In addition, it should also include some frequency regions where the relative phase difference between the parallel paths is such that the output is in-phase. This is to ensure that the output of this pair of the MPO will be zero(or positive) when the input is white noise or a wideband signal spanning the entire passband of the filter.

Thus, the output of the MPO model will be negative when a bandlimited signal is present and the output will be positive in the presence of broadband noise.

Fig. 3 shows the magnitude response of the BPF and the phase response of the APF used to detect a signal centered around 1000Hz and with a bandwidth of about 150 Hz (i.e. spread of about 75Hz on each side). The magnitude of the pole of the APF is $r = 0.975$, the frequencies at which the phase response of the APF is $-\pi/2$ and $-3\pi/2$ are 940Hz and 1080Hz respectively. The BW of the BPF is about 600Hz (300Hz on each side of 1000 Hz). Fig. 4 shows the distribution of the output of the MPO model shown in Fig. 3 for about 5000 frames each of white noise and a bandlimited signal centered at 1000Hz and of bandwidth 150 Hz corrupted with white noise at $\infty$, 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 $\infty$ dB SNR and the distribution of the bandlimited signal

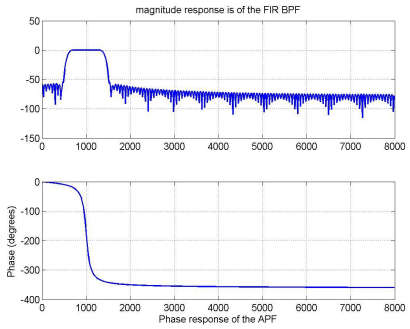


Figure 3: *Top: The magnitude response of the BPF; Bottom: The phase response of the APF*

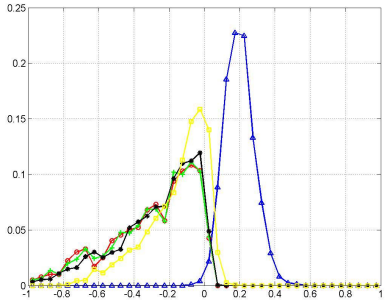

Figure 4: *Distribution of the output of MPO model when the input is white noise (blue curve:*$\triangle$ *); bandlimited signal at* $\infty$ *dB SNR (red curve:o); at 20 dB SNR (green curve: +); at 10 dB SNR (black curve: *); at 0 dB SNR (yellow curve:*$\square$ *)*

corrupted by white noise remains quite similar over the wide range of SNRs used in this study ($\infty$ to 0 dB). 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).

### 3.1. Application to speech enhancement

When the input signal is a speech signal, the frequency regions of the spectral peaks are not known beforehand. Moreover, the frequency composition will change over time. Hence the overall MPO model consists of a set of MPO structures (like the ones shown in Fig. 2), each tuned to a different frequency region. The input speech signal is split into overlapping frames and it is assumed that the frequency composition does not vary much in a given frame. In the present work, a separate MPO filter pair is used every 50 Hz and the magnitude of the pole of the APF is adjusted so that the signals with bandwidth of $\Delta\omega = 150Hz$ or less will lead to negative outputs.

The output of all the MPO filter pairs whose APF has phase response between $-\pi/2$ and $-3\pi/2$ at frequencies corresponding to that of the narrow band signal will be negative. The output of all the other MPO filter pairs will be positive ( or zero). The threshold for this discrimination is trained using the ML-based LRT technique mentioned above. In the simplest form of speech enhancement using the MPO, input speech signal is split into different frequency channels using a bank of overlapping BPFs. From each channel only those regions where the

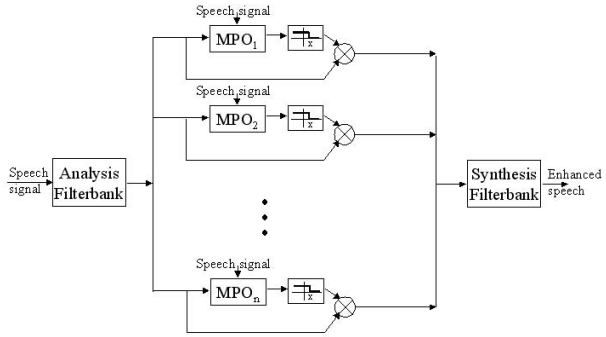


Figure 5: *Schematic of speech enhancement using MPO. The threshold $x$ is trained using ML-LRT technique and all the regions with output above this threshold are suppressed.*

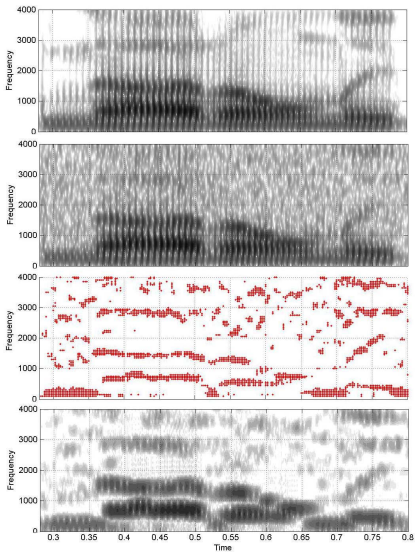


Figure 6: *Spectrogram of speech utterance "not only" in clean (Top); in 10 dB SNR ( Second); MPO enhanced utterance (Fourth); The third panel shows the output of the MPO model. Dark regions are the ones where the MPO algorithm detected signal.*

MPO is below the threshold (indicating presence of a signal) are used for reconstruction. The schematic of the enhancement scheme is shown in Fig. 5.

The frequency locations and the bandwidths of spectral peaks of speech signals are known to change over time. Thus, the MPO model has to change the properties of the MPO filter-pairs in order to efficiently detect the formants when their bandwidths vary or when two formants come close to form one effective wide spectral peak. The MPO model presented here includes an adaptation stage which tracks the strength and movement of spectral peaks in previous frames and uses this information to modify the properties of the MPO filter-pairs. Fig. 6 shows the output of the MPO model on a speech utterance corrupted by white noise at 10 dB SNR. Notice that the MPO algorithm is able to remove the noise from in between two spectral peaks while maintaining the peaks. The adaptation stage is partially successful in tracking the formants even when they come close (around 1000Hz at 0.6 sec in Fig. 6). Although, a few frames after 0.6 sec are missed when the two formants have moved close but haven't yet completely merged to form one wide spectral peak.

## 4. Conclusions and future work

The algorithm presented here, for enhancing speech signals corrupted by additive noise, is effective in enhancing the spectral peaks while suppressing the noise. There are a few spectro-temporal regions where the output of the MPO model is spuriously on (e.g. between 0.55 and 0.60 sec, around 2000-3000Hz in Fig. 6). This leads to the well-known musical noise effect. Informal listening tests reveal that the quality of the enhanced speech is greatly increased but the musical noise effect increases as the SNR reduces. Work is in progress to study this phenomenon and propose algorithms to eliminate or minimize the musical noise effect. Work is in progress to modify the adaptation stage so that the spectral peaks are tracked more accurately as they move close.

## 5. Acknowledgments

This work was supported by NSF BCS0236707.

## 6. References

[1] Lippmann, R., "Speech recognition by machines and humans", Speech Communication, vol. 22, 1-15, 1997

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

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

[4] M. Gabrea, "Robust adaptive Kalman filtering-based speech enhancement algorithm":301-304, ICASSP 2004

[5] Y. Ephraim, H. L. Van Trees, "A signal subspace approach for speech enhancement", IEEE Trans. on Speech and Audio Proc., 3, pp. 251-266, 1995

[6] C-T Lu, H-C Wang, "Speech enhancement using robust weighting factors for critical-band-wavelet-packet transform", ICASSP-04, pp. 721-724, 2004

[7] Y. Hu, P. C. Loizou, "Speech enhancement based on wavelet thresholding the multitaper spectrum", IEEE Trans. on Speech and Audio Proc., 12(1), pp. 59-67, 2004.

[8] Y. M. Cheng, D. I'Shanughnessy, "Speech Enhancement based conceptually on auditory evidence", IEEE Trans. on Signal Proc., 39(9), pp. 1943-1954, 1991

[9] J. H. Hansen and S. Nandkumar, "Robust estimation of speech in noisy backgrounds based on aspects of the auditory process", J. Acoust. Soc. Am. 97(6) : 3833-49, 1995

[10] S. Shamma, "The acoustic feature sos peech sounds in a model of auditory processing: vowels and voiceless fricatives" Journal of Phonetics, Vol. 16, pp. 77-91, 1988.

[11] N. Mesgarani, S. A. Shamma, "Speech enhancement based on filtering the spectrotemporal modulations", ICASSP-05, 2005

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