COMBINING MULTISOURCE WIENER FILTERING WITH PARALLEL BEAMFORMERS TO REDUCE NOISE FROM INTERFERING TALKERS

Heather E. Ewalt Michael T. Johnson

Speech and Signal Processing Lab
Marquette University
Milwaukee, WI 53201, USA
Email: hgratza@yahoo.com, mike.johnson@mu.edu

Abstract: The ability to extract a primary speech signal from an environment with multiple speakers is an important issue in speech enhancement [1,2]. This paper presents a method for incorporating multiple parallel beamformers with a Wiener filter. By iteratively improving the spectral magnitude estimates of each speech source, substantial improvement in overall signal separation can be obtained. The performance of the algorithm is illustrated using a simulated multiple speaker environment with resulting SNR and sSNR plots.

1. INTRODUCTION

With the number of people with hearing damage dramatically on the rise and with the expansion of global businesses requiring the use of more sophisticated video and teleconferencing equipment, extracting a primary talker among interfering talkers is imperative to current speech signal processing research. While methods exist for a variety of beamforming techniques [2, 3] as well as for multi-source filtering in stationary noise [8], the theory has yet to be developed for integrating spatial filtering with additional enhancement methods to deal with non-stationary interference.

This paper focuses on microphone arrays with a small number of microphones and a small aperture size, as would be required for hearing aid applications where users could comfortably wear the array [1]. These smaller arrays have less ability to extract the primary signal using beamforming algorithms [2] and thus are amenable to improvement through the use of further speech enhancement algorithms such as the Wiener filter. Currently this is done through application of single source post filters [2].

In particular, the problem of enhancing a primary speech signal with two interfering speech sources and known source locations is addressed here. The use of one beamformer nulls the directions of the interfering sources to extract the primary signal, however artifacts of the interfering talkers remain even after beamforming [1, 2, 4, 7]. This paper develops a new method of utilizing three beamformers, with a coupled post-processing enhancement algorithm, to extract each speech source signal. The fixed beamformers used have far-field assumptions as given in [5] and narrowband approximations through the use of frequency bands as given in [6].

1.1 Delay and Sum Beamformer:

For an M microphone array and a narrowband signal with a frequency *f*, angle of arrival Θ, filter weights *w*, and signal array *y*, and sample point *n*, a simple delay and sum beamformer estimate, *z*, is given as:

$$\mathbf{z}_{\phi,\mathbf{f}}[n] = \mathbf{w}^{\mathbf{H}}(\phi, f)\mathbf{y}[n] \qquad (1)$$

This simple delay and sum beamformer is computationally efficient and is the baseline method used.

In order to enhance the primary source signal, three fixed beamformers are steered towards the direction of each of the three sources, assumed here to be *a priori*, to yield initial estimates of each source signal. These extracted signals are then used in a post-beamforming iterative spectral enhancement algorithm initialized by the spectral characteristics of the beamformer outputs.

Figure 1 shows a block diagram of the process. Using separate beamformers to estimate each of the signals allows the system to be able to address non-stationary noise. Prior research with multiple interfering sound sources [8] has demonstrated that results degrade as the interfering sources become more non-stationary.

1.2 Wiener Filter

The post beamforming enhancement method used here is a short-term adaptive Wiener filter. This technique has been shown to be successful in situations with stationary, non-correlated, additive noise in single channel input systems [9]. In this paper, this enhancement algorithm is extended and applied to a multi-channel system with non-stationary, additive noise from interfering speakers.

The short-term adaptive Wiener filtering method as discussed in [9] calculates the spectral characteristics of the noise estimate and the estimated clean speech signal directly and uses them to develop the Wiener filter, **H**, shown as:

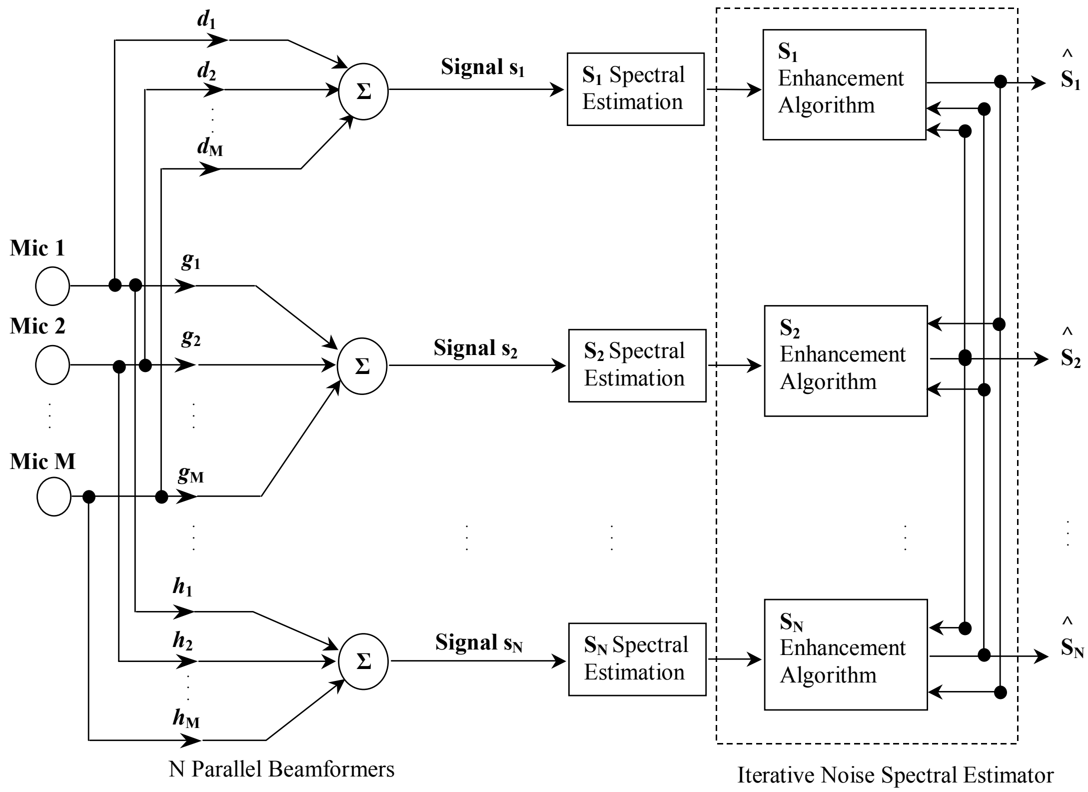


Figure 1: Block diagram of noise reduction setup.

$$\mathbf{H}(\omega) = \left[\frac{|\mathbf{S}'(\omega)|^2}{|\mathbf{S}'(\omega)|^2 + |\mathbf{N}(\omega)|^2} \right] \tag{2}$$

$$\mathbf{S}'(\omega) = \mathbf{H}(\omega)\mathbf{S}(\omega) \tag{3}$$

The estimated clean speech signal is initially the beamformer output of the source signal of interest and the noise spectral information is approximated using spectral characteristics of the remaining two sources.

The Wiener filter is then applied to the noise-corrupted signal to create a new estimate of the speech signal as shown in equation (3). The Fourier transform of the enhanced signal is given by **S'**, the Fourier transform of the noise corrupted signal is given by **S**, and the Fourier transform of the noise estimate is given by **N**. The process is repeated for a fixed number of iterations or until the signal estimates converge.

1.3 Multiple Source Wiener Filter

A multiple source Wiener filter is given by:

$$\mathbf{H}(\omega) = \left[\frac{|\mathbf{S}'(\omega)|^2}{|\mathbf{S}'(\omega)|^2 + |k_1\mathbf{N}'_1(\omega)|^2 + \ldots + |k_N\mathbf{N}'_N(\omega)|^2} \right] \tag{4}$$

The multiple source Wiener filter utilizes N noise source beamformer outputs as the initial noise estimates, $\mathbf{N}_n$. The noise corrupted signal, **S**, is the primary source beamformer output. As shown, this filter extends to N noise source signals that have no limitations on stationarity or bandwidth and are therefore applicable to a multiple speaker environment. Each noise source has an associated coupling function, k, as discussed below.

1.4 Coupling function, *k*

The coupling function determines the amount of the noise estimate's spectral energy that is filtered from the beamformed signal and is critical to equation (4). The coupling function can be calculated using the beamformer response function. It is this function that determines the spectral information of the interfering source signals that are filtered through by the beamformer and is given by:

$$W(f,\phi) = \frac{\sin\left[\frac{\pi f M d}{c}(\sin\phi_o - \sin\phi)\right]}{\sin\left[\frac{\pi f d}{c}(\sin\phi_o - \sin\phi)\right]} \tag{5}$$

This is derived from the DS beamformer and is a function of frequency and source direction. Given the beamformer function, it is possible to evaluate the coupling function across frequencies that can directly determine the spectral characteristics of the interfering sources passed through the beamformer. Incorporating this coupling function into the post filtering enhancement algorithms will remove the interfering talkers' signal artifacts with the appropriate constants across the frequency range.

Although the beamformer function is a sinc function in theory, practically, it is judicious to define the coupling function through using the envelope in order to make it more robust to discrepancies in the source locations. Given this, the coupling function is shown in Figure 2 and results in the following equation:

$$k = \begin{cases} \dfrac{\sin\left[\dfrac{Mf\pi d}{c}(\sin\phi_o - \sin\phi)\right]}{\dfrac{f\pi d}{c}(\sin\phi_o - \sin\phi)} & \text{low } f \\[2em] \dfrac{1}{\sin\left[\dfrac{f\pi d}{c}(\sin\phi_o - \sin\phi)\right]} & \text{high } f \end{cases} \tag{6}$$

2. EXPERIMENTAL SETUP

To test the algorithm, a three-source experiment was simulated. With the appropriate time shifted signals, the primary speaker was placed at a ϕ of zero while the second speaker's ϕ was $\pi/3$, and the third speaker's ϕ was $-\pi/3$. Ten variations in the SNR were created by magnitude changes of the second and third speakers' signals.

The data used to simulate the multiple speaker environments was obtained from DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus database. The speakers and sentence waveforms were chosen at random and include both male and female speakers. These waveforms were then combined to simulate a multiple speaker signal.

The spacing of the microphones at 2.5 centimeters allows the analysis of speech signals in the frequency ranges up to 6800 hertz. The input speech signal is divided into 512 point, 32 millisecond, triangular windowed frames to approximate stationarity of the speech signal.

Each of the frames was run through a filter bank to produce ten bins of equal frequency bandwidths across the given range. The filters used are twelfth order FIR band pass filters. Given the filtered and framed data for each of the microphone signals and given the *a priori* knowledge of each signal source direction, each source is beamformed.

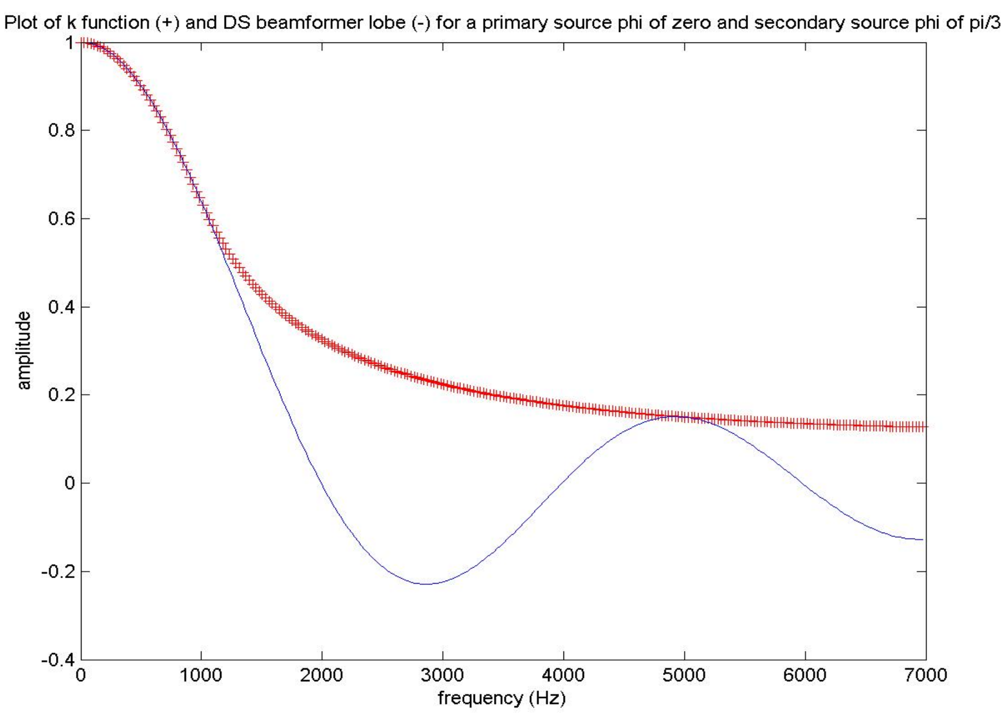
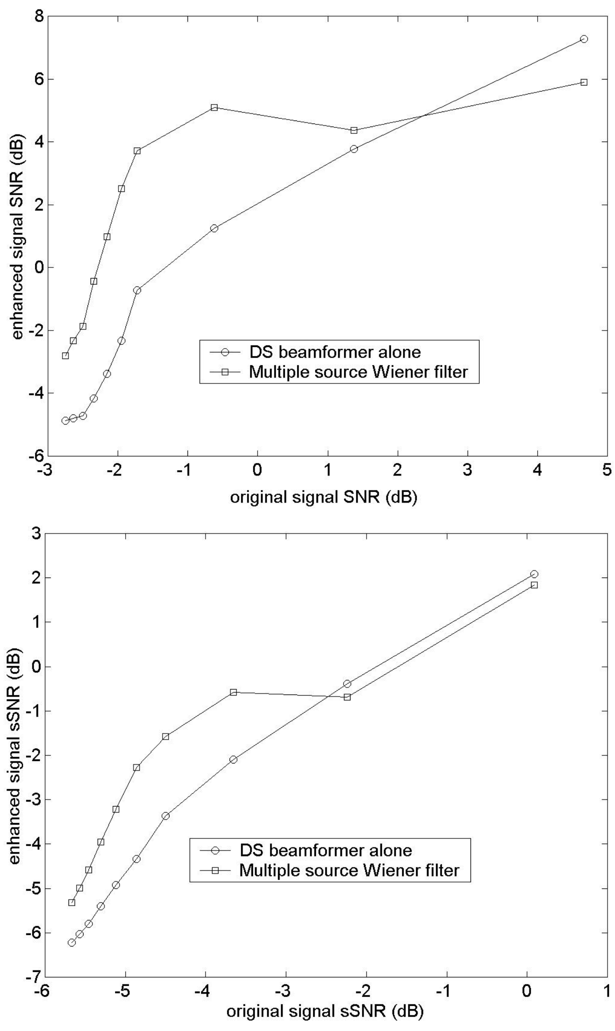


Figure 2: Plots of the coupling function and the beamformer function

3. RESULTS

SNR and segmental SNR are used here as the enhancement algorithm performance metrics. Although neither metric is a direct measure of perceived speech signal quality, it has been established that the sSNR is more correlated to perception than SNR [9]. The results are shown in Figure 3. These plots use the delay and sum beamformer SNR as a baseline to compare to the enhancement algorithm increases in SNR.



As shown, there is significant improvement above the beamformer alone, of an average of 2.8dB SNR and 1.1dB sSNR. The amount of improvement by the enhancement algorithm decreases when the interfering source power is minimal, as the SNR and segmental SNR increases. It is important to note that this multiple source enhancement algorithm performs best for an intermediate range of SNR and sSNR's. This is because the algorithm is dependent upon the interfering talkers' signal estimation, and with increased the interfering talkers' signal power, the parallel beamformers are able to obtain better noise signal estimates. However, as the interferer power becomes much larger, as is true for extremely small SNR's, the algorithm does not perform well due to the inability of the beamformer to obtain an adequate estimate of the primary signal.

4. CONCLUSION

The use of multiple, parallel beamformers integrated with a multiple source Wiener filter shows substantial improvement in the SNR and segmental SNR for a low SNR speech signal. The speech enhancement method presented in this paper contends with nonstationary, broadband noise that occurs in a multiple speaker environment.

Additional work being conducted in this area includes integration of other traditional enhancement methods, such as spectral subtraction and Ephraim Malah filtering, into the multiple source domain. Other improvements being investigated include utilizing a higher resolution beamformer, such as the minimum variance distortionless response (MVDR) beamformer, with the algorithms.

5. REFERENCES

[1] Widrow, Bernard; "A Microphone Array for Hearing Aids;" *IEEE Circuits & Systems Magazine*, Vol. 1, No. 2, 2001.

[2] Brandstein, M. and D. Ward; Microphone Arrays: Signal Processing Techniques and Applications; Springer, New York: 2001.

[3] Dundgeon, Dan E. and Don H. Johnson; Array Signal Processing: Concepts and Techniques; PRT Prentice Hall, Englewood Cliffs, NJ: 1993.

[4] Omologo, M., et al; "Microphone Array Based Speech Recognition With Different Talker-Array Positions;" *Proceedings of IEEE ICASSP*, pp. 227-230, 1997.

[5] Ryan, James G. and R. Goubran; "Near-Field Beamforming for Microphone Arrays;" *Proceedings of the IEEE ICASSP*, pp. 363-366, 1997.

[6] McCowan, I.A. and S. Sridharan; "Microphone Array Sub-Band Speech Recognition;" *Proceedings of IEEE ICASSP*, 2001.

[7] Frost III, O.L.; "An Algorithm for Linearly Constrained Adaptive Array Processing;" *Proceedings of the IEEE*, vol. 60, pp. 916-935, 1972.

[8] Saruwatari, H., et al; "Speech Enhancement Using Nonlinear Microphone Array With Noise Adaptive Complementary Beamforming;" *Proceedings of IEEE IEEE ICASSP*, pp.1049-1052, 2000.

[9] Deller, J.R., J. Hansen, and J. Proakis; Discrete-Time Processing of Speech Signals; IEEE Press, New York: 2000.