A Triple-Microphone Real-Time Speech Enhancement Algorithm Based on Approximate Array Analytical Solutions

Ryan Ritch, Meng Yu, and Jack Xin

Department of Mathematics, University of California, Irvine, CA 92697, USA
rritch@uci.edu, myu3@uci.edu, jxin@math.uci.edu

Abstract

A novel triple microphone array speech enhancement algorithm is developed based on closed form solutions to an approximate algebraic system of equations on noise directivity and speech spectral energy. The system is derived from the ensemble averaged spectral energy equations of a first order differential microphone array. For sufficiently small microphone spacing, closed form analytical solutions are derived using suitable approximations for the normal speech frequency range. The resulting algorithm is simple to implement and efficient for real time noise reduction in a reverberant environment. The algorithm’s limitation, the case of spatially almost overlapping sources, is analyzed mathematically, leading to an effective alternative solution utilizing array rotation. The algorithm is evaluated and compared with well-established beamforming algorithms. Results indicate robust gains in objective measures of speech relative to the baseline algorithms in the presence of multiple/mobile noise sources.

Index Terms: triple-microphone array, spectrum estimation, noise directivity, closed form solutions, real-time algorithm.

1. Introduction

Speech enhancement aims to eliminate noise and unexpected interferences that degrade speech quality and intelligibility in realistic listening situations. It is an indispensable technique in telecommunication and assistive listening devices such as hands-free mobile phones and hearing aids. Most of hearing aid users’ complaints stem from reduced quality and intelligibility of voices in the presence of noise. Though substantial research has been done in this area, only a limited number of methods are effective in both real time and real world conditions. Difficulties include varying noise types (incoherent, coherent, diffuse), a priori unknown number of noise sources, mobility of source locations, room reverberations, and non-stationarity. Classical speech enhancement methods rely on data from a single microphone. Spectral estimation methods, such as spectral subtraction, Wiener filtering, and the subspace listening devices such as hands-free mobile phones and hearing aids. Most of hearing aid users’ complaints stem from reduced quality and intelligibility of voices in the presence of noise. Though substantial research has not been done in this area, only a limited number of methods are effective in both real time and real world conditions. Difficulties include varying noise types (incoherent, coherent, diffuse), a priori unknown number of noise sources, mobility of source locations, room reverberations, and non-stationarity. Classical speech enhancement methods rely on data from a single microphone. Spectral estimation methods, such as spectral subtraction, Wiener filtering, and the subspace method are most widely used [1]. However in recent years, microphone array techniques have been developed and are recognized as more powerful and promising solutions [2]. Beamforming is a well-known class of algorithms in array processing. Fixed beamformers focus the array on a target sound source by forming linear combinations of delayed and weighted versions of the input signals in each microphone using the direction of the target sound. Several examples of beamforming algorithms are the delay-and-sum and superdirective beamformers [2]. Adaptive beamformers also attenuate unwanted sounds based on array focusing; however unlike fixed beamformers, they use signal information to optimize the weights, facilitating superior signal separation. Examples of adaptive beamforming algorithms include adaptive first order differential microphone arrays (FDMA) and variants [3, 4] and generalized sidelobe cancellers (GSC) [5, 6]. The performance of either FDMA or GSC degrades rapidly if the number of noise sources increases, the noise locations vary significantly in time, or a noise direction approaches that of the target sound. Additionally, the adaptive methods are iterative in the estimation of weights and hence require sufficient computation time to achieve convergence. A simpler method based on coherence function and small spacing dual-microphone ([7] and its references) has been shown recently to suppress multiple spatially separated noises. Array processing may also be carried out in convolutive generative models in a blind source separation framework ([8] and its references). However, such approaches involve adaptive filter estimation and are often unable to perform well under real time constraints.

The goal of this paper is to develop an effective method capable of reducing multiple mobile noise interferences in real time with an array of three omnidirectional microphones. The algorithm’s block diagram, shown in Fig. 1, combines single-channel power spectrum estimation with the FDMA techniques to yield a closed form approximate solution. The addition of a third microphone allows the formation of another pair of inputs for approximating noise directivity therefore facilitating both signal and noise spectral energy estimation. We show that the noise directivity estimation holds within regular voice frequency range. As is the case with beamformers, as the noise direction approaches that of the target source, the system of algebraic equations for signal and noise energy estimation becomes singular. For example, in [9], the improvement of their beamformer was less than 1 dB when the angle between the noise and target source was smaller then 45°. We analyze this limitation mathematically and develop an alternative array steering method to improve the tolerance on the included angle of target source and noise for the proposed triple-microphone method. This paper is organized as follows. In section 2, we show that with no knowledge of noise arrival direction, the standard two-microphone array is under-determined. This leads us to a triple-microphone array where a closed system of equations is available for estimating noise directivity and the target speech spectrum. Next, we discuss the case of a small included angle between the target source and the noise. Experimental and evaluation results are in section 3. Concluding remarks are in section 4.
2. Array Equations, Solutions and Analysis

For most listening devices, the speech in front of the user is the desired signal, as shown in Fig. 2 with $\theta = 0^\circ$. We shall begin with the well-known two-microphone array as a motivation for pursuing a triple microphone array to reach an approximate statistically closed system of equations for speech and noise spectral energy. We also explain the difference between end-fire and broadside array configuration in the regime of similar incident angles of speech and noise sources where the array system becomes singular. A helpful remedy is then proposed.

2.1. Dual-Microphone Array

Two omni-directional microphones, $Mic_1$ and $Mic_2$, are spaced a distance $d$ apart. The desired speech comes from a direction $\theta$ and the noise signal from $\psi$, both being recorded by $Mic_1$ and $Mic_2$. This gives

$$Mic_1 = s(t) + n(t)$$

$$Mic_2 = s \left( t - \frac{d}{c} \cos \theta \right) + n \left( t - \frac{d}{c} \cos \psi \right).$$

The recorded signals are delayed by a factor of $d/c$, that is, $Mic_j(t) \rightarrow Mic_j(t - \frac{d}{c})$ for $j = 1, 2$. Based on the FDMA method [2, 3], channel cross differencing with delay produces

$$Ch_1(t) \triangleq Mic_1(t) - Mic_2 \left( t - \frac{d}{c} \right) = s(t) + n(t)$$

$$- s \left( t - \frac{d}{c} (1 + \cos \theta) \right) - n \left( t - \frac{d}{c} (1 + \cos \psi) \right).$$

The equations (3)-(4) are Fourier transformed to

$$Ch_1(\omega e^{j\omega}) = \left( 1 - e^{-j\omega \frac{d}{c} (1 + \cos \theta)} \right) S(\omega e^{j\omega}) + \left( 1 - e^{-j\omega \frac{d}{c} (1 + \cos \psi)} \right) N(\omega e^{j\omega})$$

$$Ch_2(\omega e^{j\omega}) = \left( e^{-j\omega \frac{d}{c} \cos \theta} - e^{-j\omega \frac{d}{c} \cos \psi} \right) S(\omega e^{j\omega}) + \left( e^{-j\omega \frac{d}{c} \cos \theta} - e^{-j\omega \frac{d}{c} \cos \psi} \right) N(\omega e^{j\omega}).$$

By standard statistical independence assumption of $S$ and $N$, the power spectra of channel 1 and channel 2 are

$$E(|Ch_1(\omega e^{j\omega})|^2) = 2 \left[ 1 - \cos \left( \frac{\omega d}{c} (1 + \cos \theta) \right) \right] E(|S(\omega e^{j\omega})|^2)$$

$$+ 2 \left[ 1 - \cos \left( \frac{\omega d}{c} (1 + \cos \psi) \right) \right] E(|N(\omega e^{j\omega})|^2).$$

(7)

$$E(|Ch_2(\omega e^{j\omega})|^2) = 2 \left[ 1 - \cos \left( \frac{\omega d}{c} (1 + \cos \theta) \right) \right] E(|S(\omega e^{j\omega})|^2)$$

$$+ 2 \left[ 1 - \cos \left( \frac{\omega d}{c} (1 - \cos \psi) \right) \right] E(|N(\omega e^{j\omega})|^2).$$

(8)

In case of an end-fire array, we let $\theta \rightarrow 0$ and obtain

$$E(|Ch_1(\omega e^{j\omega})|^2) = 2 \left[ 1 - \cos \left( \frac{2\omega d}{c} \right) \right] E(|S(\omega e^{j\omega})|^2)$$

$$+ 2 \left[ 1 - \cos \left( \frac{\omega d}{c} (1 + \cos \psi) \right) \right] E(|N(\omega e^{j\omega})|^2).$$

(9)

$$E(|Ch_2(\omega e^{j\omega})|^2) = 2 \left[ 1 - \cos \left( \frac{\omega d}{c} (1 - \cos \psi) \right) \right] E(|N(\omega e^{j\omega})|^2).$$

(10)

The percentages of noise power spectra in channels 1 and 2 are functions of the noise azimuth $\psi$. Combining (9) and (10), we have

$$E(|Ch_1(\omega e^{j\omega})|^2) = 4 \sin^2 \frac{\omega d}{c} E(|S(\omega e^{j\omega})|^2)$$

$$+ \frac{1 - \cos \left( \frac{2\omega d}{c} \right)}{1 - \cos \left( \frac{\omega d}{c} (1 + \cos \psi) \right)} E(|N(\omega e^{j\omega})|^2).$$

(11)

We define

$$g_\psi(\psi) \triangleq \frac{1 - \cos \left( \frac{2\omega d}{c} \right)}{1 - \cos \left( \frac{\omega d}{c} (1 + \cos \psi) \right)}.$$

(12)

This directivity coefficient is needed to estimate the speech power spectrum. Without the knowledge of noise direction, especially if the noise source is mobile, this two microphone equation (11) is under-determined, necessitating the input of more data. The natural progression to a triple-microphone array is considered next.

2.2. Triple-Microphone Array

Consider the addition of a third microphone to form a linear array, as shown in Fig. 2. Cross channel differencing with delay similar to (3) and (4), we produce

$$Ch_1(t) \triangleq Mic_1(t) - Mic_3 \left( t - \frac{d + d'}{c} \right) = s(t) + n(t)$$

$$- s \left( t - \frac{d + d'}{c} (1 + \cos \theta) \right) - n \left( t - \frac{d + d'}{c} (1 + \cos \psi) \right).$$

(13)

$$Ch_3(t) \triangleq Mic_3(t) - Mic_1 \left( t - \frac{d + d'}{c} \right) = s \left( t - \frac{d + d'}{c} \cos \theta \right) - n \left( t - \frac{d + d'}{c} \right).$$

(14)
By calculating the Fourier transform and channel power spectra of (13) and (14), we find the analogue of (11) to be
\[
E(|C_H(\omega)|^2) = \sin^2 \left( \frac{\omega(d + d')}{c} \right) E(|S(\omega)|^2)
\]
\[+ \cos \left( \frac{\omega(d + d')}{c} (1 + \cos \varphi) \right) E(|C_H(\omega)|^2) \cdot \sin \left( \frac{\omega(d + d')}{c} \right) E(|C_H(\omega)|^2). \tag{15}
\]
In view of (11) and (15), we obtain the following equation with unknown \(\cos \varphi\) (hence directivity):
\[
E(|C_H(\omega)|^2) \sin^2 \left( \frac{\omega(d + d')}{c} \right) - E(|C_H(\omega)|^2) \sin^2 \left( \frac{\omega d'}{c} \right) = \sin^2 \left( \frac{\omega(d + d')}{c} \right) \frac{1 - \cos \left( \frac{\omega(d + d')}{c} (1 + \cos \varphi) \right)}{1 - \cos \left( \frac{\omega(d + d')}{c} (1 - \cos \varphi) \right)} E(|C_H(\omega)|^2)^2 - \sin^2 \left( \frac{\omega d'}{c} \right) \frac{1 - \cos \left( \frac{\omega d'}{c} (1 + \cos \varphi) \right)}{1 - \cos \left( \frac{\omega d'}{c} (1 - \cos \varphi) \right)} E(|C_H(\omega)|^2)^2.
\tag{16}
\]
Once \(\cos \varphi\) is solved in (16), target power spectrum can be calculated from (11). However, (16) is not exactly solvable in \(\varphi\).

We observe that \(E(|S(\omega)|^2)\) in (11) is a function of \(g_d\).

If \(d + d' \ll 1\) or \(\frac{\omega(d + d')}{c} \ll 1\), the following approximations of \(g_d\) can be made:
\[
g_d \approx \frac{1 - \cos \left( \frac{\omega d'}{c} (1 + \cos \varphi) \right)}{1 - \cos \left( \frac{\omega d'}{c} (1 - \cos \varphi) \right)} \approx \frac{1 + \cos \varphi}{2},
\]
\[
\approx \frac{1 + \cos \varphi}{2} = \frac{1 + \cos \varphi}{2},
\]
\[
\approx \frac{1 + \cos \varphi}{2} = \frac{1 + \cos \varphi}{2},
\]
\[
\approx \frac{1 + \cos \varphi}{2} = \frac{1 + \cos \varphi}{2}.
\tag{17}
\]

Therefore, the directivity coefficients have the following approximate analytical expression:
\[
g_d \approx g_{d + d'} \approx \frac{\omega d'}{2},
\]
\[
\approx \frac{E(|C_H(\omega)|^2)}{E(|C_H(\omega)|^2)} \sin^2 \left( \frac{\omega d'}{c} \right) E(|C_H(\omega)|^2)^2 - \sin^2 \left( \frac{\omega d'}{c} \right) E(|C_H(\omega)|^2)^2.
\tag{18}
\]

To justify the above approximation, we computed \(\frac{g_d g_{d'}}{g_d + g_{d'}}\) in case \(d = d'\) and plotted it as a function of frequency \((\omega = 2\pi f)\) in Fig. 3 for five noise directions. Within the voice frequency range, the maximum deviation from 1 is 9%, making this approximation well-justified within a fairly wide range of parameters.

\[\text{Figure 3: Illustration of the deviations of } g_{d+d'} \text{ from 1 the value used in the small spacing approximation, } d = 0 \text{ mm, frequency in the speech range.}\]

\[\text{Figure 4: Illustration of the ill-conditioning problem of the linear system (7)-(8) as the included angle of target and noise sources varies. Microphone spacing } d = 5 \text{ mm and frequency } f = 2000 \text{ Hz. The red region boundaries denote the angle where the condition number exceeds 50.}\]

\[\text{2.3. Spatial Overlapping Problem}\]

The determinant of the system (9) and (10) is:
\[
\det \approx 4 \left( 1 - \cos \left( \frac{2\omega d'}{c} \right) \right) \left( 1 - \cos \left( \frac{\omega d'}{c} (1 + \cos \varphi) \right) \right) \left( 1 - \cos \left( \frac{\omega d'}{c} (1 - \cos \varphi) \right) \right).
\tag{19}
\]

As \(\varphi \to 0\), the determinant goes to 0, and so equation (11) is not valid for small \(\varphi\). Likewise in (17), \(\cos^4 \left( \frac{\omega d'}{2} \right)\) is undefined at \(\varphi = 0^\circ\). Thus the proposed solution breaks down if the target source and noise form a small included angle.

To examine a possible solution to this limitation, fix the included angle \(\varphi - \theta\) of the target and noise in the range \((0^\circ, 40^\circ)\) and rotate the array clockwise about its mid-point. For simplicity, the azimuth of the front microphone \(\theta\) is always taken as \(0^\circ\). As a result, the rotation is equivalent to rotating the target source and noise together about the array mid-point. The target/noise bisector is \(\frac{\omega d'}{\omega} \approx \frac{\omega d'}{c}\) and is annotated along the edge of the circle in Fig. 4. The condition number of equations (7) & (8) is computed every \(10^\circ\). A large condition number indicates that the system is close to singular, implying the degradation of target spectrum estimation. In Fig. 4, good enhancement performance is expected if the speech/noise bisector angle points away from the red region where the condition number is greater than 50. The bisector angle for the end-fire microphone array is \(\varphi\), which enters the red region once \(\varphi \) is smaller than \(45^\circ\). A large red area occurring with small included angles degrades end-fire array enhancement. However as the angle bisector of the target source and noise becomes perpendicular to the microphone array (broadside), the condition number reaches its minimum as demonstrated in Fig. 4. For example, in the top left subplot \(\varphi - \theta = 10^\circ\), as the microphone array is rotated clockwise by \(85^\circ\) from the end-fire position, the ill-conditioning problem is maximally cured, allowing the triple-microphone enhancement to remain effective for nearly overlapping sound sources. Experimental results will be shown next.

\[\text{3. Experiment and Evaluation}\]

To evaluate the performance of the proposed triple-microphone based method, we used four male and four female utterances from the TIMIT speech corpus. Clean speeches were corrupted by interfering speeches from other speakers, where the room
impulse responses were synthesized by image-source method [10] with \(T_{\text{ms}} = 0.2s\), microphone spacing as 1cm, and the distance from sources to microphone array at 1m. The STFT was computed using a 64ms Hanning window with a 60ms overlap. In order to facilitate real time processing, the expectation in (7) and (8) is approximated by single time frame. On an Intel i7-2600K with 16GB of RAM running Mac OS 10.7.3 and MATLAB 7.10.0, processing a 9.98 second recording required approximately 2 seconds, indicating the potential for performing the separation in real time. The comparison between the proposed method and other beamformers is shown in Fig. 5. The proposed method achieves higher signal-to-noise ratio (SNR) than the other adaptive beamformers. In terms of noise reduction, it performs significantly better than adaptive GSC at large included angles of target and noise, while better than adaptive FDMA at small included angles. Following the discussion in section 2.3 and Fig. 4, in Fig. 6, we tested the proposed method with array steering on two small included angles (60\(^\circ\) and 40\(^\circ\)). The maximum gain on target speech is achieved when the array forms almost a right angle with the bisector of the incoming sound directions (broadside), consistent with Fig. 4. Take the included angle \(\varphi - \theta = 40^\circ\) as an example, the SNR is improved from 1.76 dB (end-fire) to 6.64 dB (after rotating the array by 100\(^\circ\)). This experiment also shows that the performance is robust in directivity variations. 

We then tested the proposed real-time algorithm on speech signals in an ordinary office room. The various numbers (1 to 4) of speakers are speaking simultaneously 1 m away from an end-fire positioned triple microphone array (Xbox Kinect). Experimental results are listed in the Table 1. Finally, the proposed method is tested on a mobile interfering speaker walking from \(\varphi = 60^\circ\) to 120\(^\circ\) at uniform speed, with SNR improvement shown in Fig. 7.

Table 1: Performance on the recorded data with end-fire array

<table>
<thead>
<tr>
<th>Noise sources</th>
<th>SNR(_{\text{after}}) [db]</th>
<th>SNR(_{\text{after}}) [db]</th>
</tr>
</thead>
<tbody>
<tr>
<td>30(^\circ)</td>
<td>0</td>
<td>2.4</td>
</tr>
<tr>
<td>60(^\circ)</td>
<td>4.7</td>
<td></td>
</tr>
<tr>
<td>150(^\circ)</td>
<td>0</td>
<td>17.3</td>
</tr>
<tr>
<td>60(^\circ), 90(^\circ)</td>
<td>-1.0</td>
<td>3.3</td>
</tr>
<tr>
<td>90(^\circ), 150(^\circ)</td>
<td>-1.0</td>
<td>14.0</td>
</tr>
<tr>
<td>90(^\circ), 120(^\circ), 150(^\circ)</td>
<td>-2.0</td>
<td>10.1</td>
</tr>
<tr>
<td>90(^\circ), 120(^\circ), 150(^\circ), 180(^\circ)</td>
<td>-3.0</td>
<td>8.6</td>
</tr>
</tbody>
</table>

4. Conclusion

We presented a real-time speech enhancement method based on triple microphone array, its approximate analytical solutions and proper steering. It adaptively estimated noise directivity and suppressed directional noises efficiently. The experimental results indicated the robust performance and efficiency for speech enhancement.

5. References