AN ALTERNATE APPROACH TO ADAPTIVE BEAMFORMING USING SRP-PHAT

Avram Levi, Harvey F. Silverman.

LEMS
Division of Engineering
Box D, Brown University, Providence, RI 02912
{avramlevi,hfs}@lems.brown.edu

I. INTRODUCTION

We consider a typical noisy room environment with a microphone array and multiple simultaneous wideband sources. When more than one source is active, each microphone records an additive mixture of 1) uncorrelated background noise, 2) direct speech signals from the sources 3) correlated echoes of the sources. The result is an unintelligible mixture of speech sounds. Over the last three decades, research has been performed to utilize various microphone-array configurations to extract reliable and intelligible speech[1]. The most typical methods are beamforming[2] or blind source separation[3].

Beamforming has its roots from narrowband applications where, by adjusting the time-delays of each sensor for a particular look-direction and summing them, the signals from the desired source are added constructively while interfering signals are not. This improves the general signal-to-noise ratio. To attenuate another source in a different look direction, weights are introduced in front of each sensor to adjust the beam-pattern for that particular frequency band. In the case of stationary wideband systems, this has been generalized to pre-filtering with FIR filters. Then the beam-pattern can be adjusted so that the frequencies at which the interfering signal are active will have spatial nulls at the interfering signal’s look direction/position. For non-stationary wideband sources such as human talkers, these filters need to be constantly adapted to properly place the nulls at the correct frequencies. This problem has been widely studied since the 1970s and various optimization techniques have been proposed for determining the adaptation parameters(adaptive beamforming)[4][5][6].

II. THE SOURCE-TO-INTERFERER RATIO(SIR)

For this development, we assume the common linear-system model and point sources and a set of microphones. Consider the desired point source located at $\vec{x}_d$ and let $m_j(t)$ be the time-domain signal for the $j^{th}$ microphone, located at $\vec{x}_j$, of an N-microphone array system denoted as $\lambda$. We separate $m_j(t)$ into the sum of the direct-path signal $b_j(t)$ that originates at the desired source and another signal $i_j(t)$ that is the sum of all other signals (including other independent sources) that reach the microphone.

$$m_j(t) = b_j(t) + i_j(t).$$

(1)

$b_j(t)$ is the double convolution of the responses of the direct source to microphone path, $h_d(t, \vec{x}_d, \vec{x}_j)$ and the source transfer function, $h_s(t, \vec{\theta}_j)$ with the true source signal, $s(t)$.

$$b_j(t) = h_d(t, \vec{x}_d, \vec{x}_j) \ast h_s(t, \vec{\theta}_j) \ast s(t)$$

(2)

The source transfer function, $h_s(t, \vec{\theta}_j)$ represents the angular dependencies of a directional source with respect to microphone $j$. For this argument, we shall simplify by assuming that the microphones are in the same general direction with respect to the source (within ±60°) making the variation in the source transfer function minimal[11]. We model the response of the direct source to microphone path with the simple form of a delayed, attenuated impulse with the attenuation due to the source to microphone distance, $d_j$. Again for this argument, we shall assume that these dependencies are minimal, hence making the difference in attenuations due to $h_d(t, \vec{x}_d, \vec{x}_j)$ minimal while keeping the significance of the differences among the delays, $\tau_{js}$.

$$b_j(t) = \beta s(t - \tau_{js}) \equiv a(t - \tau_{js}),$$

(3)

In this paper we take an alternate approach to adaptive beamforming. Instead of adapting FIR filters at each microphone to modify the beam-pattern, we spectrally subtract[7][8] the frequencies dominated by interfering signals from the simple delay-and-sum beamformer output. To identify these frequencies, we introduce a simple procedure based on the SRP-PHAT[9] to determine the relation between the magnitudes of the desired and the interfering signals at each frequency at every time point, i.e. at each time-frequency point. We then assign each time-frequency point to one of the sources or as background noise and spectrally subtract all time-frequency points that are not suitably assigned to the source. This procedure requires that point-source estimates of all independent sources, as well as an estimate of the background noise spectrum in each microphone are known. In the next section, we explain in detail, how the genuine time-frequency points are determined and as well as the spectral subtraction process. We support our idea with results from real data recorded with the Huge Microphone Array(HMA)[10].
where $\beta$ is a constant, and $a$ is the attenuated form for direct-path signal $b_j(t)$. This implies,

$$m_j(t) = a(t - \tau_{js}) + i_j(t)$$  \hspace{1cm} (4)

In a normal environment, $i_j(t)$ is the sum of 1) general background noise, 2) correlated noise, i.e., reflections from the source and, 3) uncorrelated noise from other sources. Our goal is to extract a reasonable estimate of $a(t)$ from the noisy data by first attenuating the background noise and correlated noise through simple delay-and-sum beamforming and second removing all the remaining uncorrelated noise from other sources. First, we steer each microphone to the source location by appropriately time-shifting the data, defining the steered microphone signal as $m_j^s(t)$.

$$m_j^s(t) = m_j(t + \tau_{js}) = a(t) + i_j(t + \tau_{js})$$  \hspace{1cm} (5)

In the frequency domain,

$$M_j^s(\omega) = A(\omega) + I(\omega)e^{-j\omega\tau_{js}}$$  \hspace{1cm} (6)

Again, for this development, we assume the magnitude of the interfering signal, $|I_j(\omega)|$, is approximately the same for every microphone, i.e., $|I_j(\omega)| \approx |I(\omega)|$. This is a reasonable assumption for background noise and when variations in the source-to-microphone distances are minimal for each source in a room (i.e., the microphones are grouped together) that has fairly uniform reflectivity off of all surfaces. Thus, the signal-to-interferer ratio at every microphone is the same function of frequency $\omega$.

$$SIR(\omega) \equiv SIR_j(\omega) \equiv \frac{|A(\omega)|^2}{|I(\omega)|^2}$$  \hspace{1cm} (7)

In general, at any given time, the $SIR(\omega)$ term is dominated by one of the following: The signal-to-background noise ratio, $SNR(\omega)$, the signal-to-reverberance ratio, $SRR(\omega)$ and the signal-to-cross talk ratio, $SCR(\omega)$.

A simple delay-and-sum beamformer steered to the desired source produces,

$$M_k^s(\omega) = \frac{1}{N} \sum_{j=1}^{N} M_j^s(\omega) = A(\omega) + \frac{1}{N} \sum_{j=1}^{N} I_j(\omega)e^{-j\omega\tau_{js}}$$  \hspace{1cm} (8)

Note that the phase term for the interfering signal is different for each microphone, resulting in the attenuation of the interfering signal in $M_k^s(\omega)$. In typical room environments, the level of background noise and reverberation are on the same order as $A(\omega)$, $SIR(\omega) \approx 1$, and the attenuation capabilities of the delay-and-sum beamformer coupled with a spectral subtraction procedure are considered sufficient to yield intelligible speech. On the other hand, the level of the interfering signal coming from cross-talk may be much greater than $A(\omega)$, $SIR(\omega) \ll 1$, and in this case, the attenuation capability of the delay-and-sum beamformer will be insufficient to attenuate the interfering signal, severely impacting the intelligibility of the desired signal. In the next section, we discuss a measure to identify the frequencies at which this phenomenon happens.

### III. SIR($\omega$) AND GCC-PHAT

The GCC-PHAT[12], between microphones $j$ and $k$ of an array, for each signal steered to the talker is defined in the frequency domain as

$$\Psi_{jk}(\omega) \equiv \frac{M_j^s(\omega) M_k^s(\omega) \Phi_{jk}(\omega)}{|M_j^s(\omega)||M_k^s(\omega)|}.$$  \hspace{1cm} (9)

Substituting Equation 6, this becomes,

$$\Psi_{jk}(\omega) = \frac{|A(\omega)|^2 e^{j\Phi_{jk}(\omega)} + |I(\omega)|^2 e^{j\Phi_{jk}(\omega)} + C(\omega)}{|M_j^s(\omega)||M_k^s(\omega)|}.$$  \hspace{1cm} (10)

where $\Phi_{jk}^h(\omega)$ and $\Phi_{jk}^l(\omega)$ denote the phase angle components due to the direct wave of the desired source and the interference respectively. As both microphones are steered to the talker, $\Phi_{jk}^h(\omega) \approx 0$. Since the phase values due to the steering delays and the internal phase value of $I_j(\omega)$ are somewhat arbitrary, $\Phi_{jk}^l(\omega)$ is not necessarily zero. Rather, a random model for $\Phi_{jk}^l(\omega)$ is a random variable between $-\pi$ and $\pi$. Due to limitations in scope, we do not theoretically analyze the distribution of this random variable. However simulated results for two point sources (desired source and an interferer) are shown in Figure 1 where it is shown that the distribution of phase values for high SIR time-frequency points have a sharp peak around 0 for an SIR value of 20 dB, whereas it is more or less uniformly distributed for lower SIR values.

The last term in Equation 10, $C(\omega)$, is shorthand for the cross-terms,

$$C(\omega) \equiv A(\omega)I_k^*(\omega)e^{j\omega\tau_{rk}} + A^*(\omega)I_j(\omega)e^{-j\omega\tau_{jr}}.$$  \hspace{1cm} (11)

To obtain the dependence of the GCC-PHAT on $SIR(\omega)$, we factor out $A(\omega)$ from the numerator and rewrite Equation 10 as

$$\Psi_{jk}(\omega) \equiv \frac{|A(\omega)|^2}{|M_j^s(\omega)||M_k^s(\omega)|} \left\{ e^{j\Phi_{jk}^h(\omega)} + e^{j\Phi_{jk}^l(\omega) SIR(\omega)} + C(\omega) \right\},$$  \hspace{1cm} (12)

where it can be shown that

$$|C(\omega)| \leq \frac{2}{\sqrt{SIR(\omega)}}.$$  \hspace{1cm} (13)

The magnitude of the GCC-PHAT is always 1 by definition, therefore the weight of the three separate terms in Equation 12 will be manifested in the phase. If the interference is relatively
Fig. 2. Experimental setup and room overview from above. The two talkers, denoted by the large dots are facing the 16 microphone array. The microphones are denoted by the x's.

large \((SIR(\omega) \ll 1)\) the middle term, the phase of which is a random variable, dominates.

\[ \Psi_{jk}^s(\omega) \approx e^{j\phi_{jk}^s(\omega)} \] (14)

At the other extreme, if the source signal is relatively large \((SIR(\omega) \gg 1)\), the first term, the phase of which is \(\approx 0\) dominates,

\[ \Psi_{jk}^s(\omega) \approx e^{j\phi_{jk}^s(\omega)} = e^0 = 1 \] (15)

III-A. Extension to SRP-PHAT

In 2000 Dibiase [9] showed that the steered response power (SRP) of a beamformer may be computed from the sum of the generalized cross correlation (GCC) of all possible pairs of microphones within the beamformer over an interval. This property also holds for the GCC-PHAT. We define the SRP-PHAT value of an N-microphone array, \(\lambda\), steered to the source at a frequency \(\omega\),

\[ \Upsilon_{\lambda}^s(\omega) = \frac{1}{N^2} \sum_{j=1}^{N} \sum_{k=1}^{N} \Psi_{jk}^s(\omega)^2. \] (16)

Using Equations 14 and 15, we deduce that for large SIR(\(\omega\))

\[ \Upsilon_{\lambda}^s(\omega) \approx \frac{1}{N^2} \sum_{j=1}^{N} \sum_{k=1}^{N} e^{j\phi_{jk}^s(\omega)} = 1 \] (17)

where as for small SIR(\(\omega\)),

\[ \Upsilon_{\lambda}^s(\omega) \approx \frac{1}{N^2} \sum_{j=1}^{N} \sum_{k=1}^{N} e^{j\phi_{jk}^s(\omega)} \ll 1. \] (18)

This is validated through simulated data where the counts of SRP-PHAT values are presented for different SIR values in Figure 1.

Thus, the SRP-PHAT value has a very high correlation with the SIR value. This means it can be used effectively as a discriminator for a particular point in time and frequency to improve the beamformer output by reducing interference for a particular source.

IV. IMPROVED BEAMFORMING ALGORITHM

In this section, we will describe a simple algorithm that exploits the correlation between \(SIR(\omega)\) and \(\Upsilon_{\lambda}^s(\omega)\) to remove cross-talk from one or more interferers from a delay-and-sum beamformer output. The main idea is that the array is aimed at each source as a function of SIR in an interval. This property where the desired signal is prevalent, almost all time-frequency points are correctly attributed to the desired source. Finally, a time-frequency point is assigned to source \(\hat{s}\) if

\[ \frac{\Upsilon_{\lambda}^s(\omega)}{\Upsilon_{\lambda}^\hat{s}(\omega)} > \rho \quad \forall \omega \neq \hat{s} \] (19)

\[ \frac{|M_{\lambda}(\omega)|^2}{|M_{\lambda}(\omega)|^2} > \rho. \] (20)

Here, \(\rho\) and \(\rho\) are predetermined ratio values and \(|M_{\lambda}(\omega)|^2\) is defined as the average magnitude of background noise at frequency \(\omega\). The time-frequency points that are assigned to talker \(\hat{s}\) are kept as they are. The magnitudes of the time-frequency points that are not assigned to talker \(\hat{s}\) are reduced to a predetermined small value, \(\mu\) to assure that those time-frequency points will not be audible and the output is reconstructed. In order to assure that each time-frequency point is assigned to a single source, \(\rho\) should be greater than 1. Its exact value will determine how likely a time-frequency point in which more than one source is active will be assigned to a given source. The \(\rho\) value is set to eliminate the musical noise generated by simple spectral subtraction. Its value should be proportional to the variance of the background noise spectrum. This whole idea is essentially very similar to the spectral subtraction algorithm by Beroult et. al [7] with the significant contribution being the use of SRP-PHAT to determine the correlated noise sources in addition to the uncorrelated background noise.

V. EXPERIMENTS

To test our method, we set up two speakers at the locations shown in Figure 2. We then played 8 consecutive speech files per speaker from the TIMTIT database using the 16 microphones at a 20 kHz sampling rate. The first few seconds of the recordings were only background noise. We divided the recorded speech into frames of 1024 samples (51.2ms) advancing every 100 samples (5ms). This yielded more than 4900 speech frames. For each frame, we calculated the energy and the SRP-PHAT value at each frequency point. Then for each talker we chose the frequency points using the inequalities at Equations 19 and 20. The \(\rho\) value was set to correspond to a ratio of 1.5dB, the \(\rho\) value was set to 3 and the \(\mu\) value was set to 0.01. Comparative, qualitative spectrographic results are presented in Figure 3 where the removal of almost all cross-talk is observable. Note the large difference in the region pointed by the arrows.

To quantitatively evaluate the proposed method’s performance we report the percentage of time-frequency points attributed to the desired source as a function of SIR. To obtain the SIR value for time-frequency points, we took advantage of the repeatability of using recorded data through speakers and took two additional recordings, 1) with only the desired source active 2) with only the interferer active. We applied the simple delay-and-sum beamformer aiming at the desired source’s location for both of the recordings. The energy ratio of the outputs at each time-frequency point yielded an approximation of the SIR in the original beamformed signal with both the desired and the interfering source active. We then looked at the time-frequency points in frequency bands 250 Hz wide for a number of frequencies and plotted the percentage of frames attributed to the desired source as a function of SIR in Figure 4. The distributions are similar for different frequencies. For high SIR values where the desired signal is prevalent, almost all time-frequency points are correctly attributed to the desired source.
We propose a novel technique for isolating sources using a microphone array where the source locations are known. The method determines whether a time-frequency point in a delay-and-sum beamformer output is authentic to the aimed source. The algorithm replaces all time-frequency points that are deemed inappropriate with a predetermined fraction of the noise floor at that frequency. In the end, we were able to successfully remove most (up to 90%) of the time-frequency points with high interference while keeping almost all of the low interference time-frequency points. This resulted in a significant increase in the intelligibility and quality of speech of each source independently. Because of limited space, the comparison of this method to other related published methods in terms of performance and computational complexity will be discussed in a separate journal paper[13]. For a simple demonstration of the method please visit http://www.lems.brown.edu/~avramlevi

VI. CONCLUSION

Fig. 3. An example for two seconds of speech with two sources: a) Original desired source b) Original interfering source, c) Typical single microphone, d) Delay-and-sum beamformer aimed desired source, e) Proposed method aimed at desired source. The arrows point to a region where the effects of the method is easily observed

For SIR values less than 0dB, around 90% of the time-frequency points are removed through spectral subtraction.

VII. REFERENCES


