Desired-speech signal cancellation by microphone arrays in reverberant rooms

M.W. Hoffman, M.J. Link, and K.M. Buckley
Department of Electrical Engineering
University of Minnesota

Abstract

This paper addresses the problem of cancellation of a desired-signal in a head worn microphone array processor. The effect of reverberation-induced multipath on a constrained minimum variance beamformer is considered. Constrained power minimization algorithms are particularly susceptible to this desired-signal cancellation. Specifically, this paper evaluates the conditions under which the cancellation of a desired-speech signal occurs for a head worn array of microphones in a reverberant room. The effects of acoustic headshadow are incorporated in a simulated array processor. In addition, room reverberation effects are also modelled. The acoustic parameters investigated include room size, average wall absorption, and speaker-to-listener distance. Distortion-based speech intelligibility measures are employed to assess the desired-speech cancellation.

I INTRODUCTION

The work described herein is part of a larger effort investigating the potential of an array of microphones as a preprocessor for a hearing aid. Currently available single microphone hearing aids suffer from a number of problems. A common complaint of poor performance in reverberant and noisy environments suggests that a spatial filter may enhance listening for the impaired in these difficult circumstances. The advantage of binaural listening over monaural listening in complex sound fields has often been cited as motivation for this application. Past work, [Allen et al. 1977, Peterson 1989, Hoffman and Buckley 1990], has demonstrated both the advantages and the disadvantages of the multi-microphone approach. This paper attempts to establish when, and to what extent, cancellation of the desired-speech signal occurs when a power minimizing array processor is applied in a reverberant environment.

Specifically, linearly constrained minimum variance (LCMV) beamforming [Frost 1972] is considered as a potential candidate for this application. Since desired-signal cancellation can be a problem with this approach, this issue must be addressed. Other approaches to array processing which are not as susceptible to desired signal cancellation include fixed beamforming. While lacking this drawback of LCMV beamforming, fixed beamforming also lacks the most useful asset of LCMV. Namely, that LCMV is adaptive and in many diverse acoustic scenarios it can provide an enhanced desired signal which contains less noise and reverberated energy than that produced by a fixed beamformer. In considering LCMV over other data adaptive beamforming approaches it should be noted that LCMV provides a *relatively* simple, well understood, effective array processing approach. LCMV also has a number of simple, robust extensions.

Recently [Greenberg and Zurek 1991] reported that cancellation of the desired-signal can be a problem with a two channel, modified version of an LCMV based processor under certain reverberant conditions. The modification they made to the Griffiths-Jim beamformer inhibited array adaptation, hence desired-signal cancellation, in the presence of a high power desired signal relative to the background noise. With this modification the desired-signal cancellation reported was dependent on the level of the reverberated (correlated) desired-speech in the LCMV processor.

The parameters effecting desired signal cancellation which are investigated in this paper are the tap length in the LCMV processor (see figure 1) and the direct-to-reverberant energy ratio of the desired speech signal. Many individual parameters affect the direct-to-reverberant energy ratio, such as wall absorption, room size, and source-to-listener distance. The performance measures used to quantitatively describe the

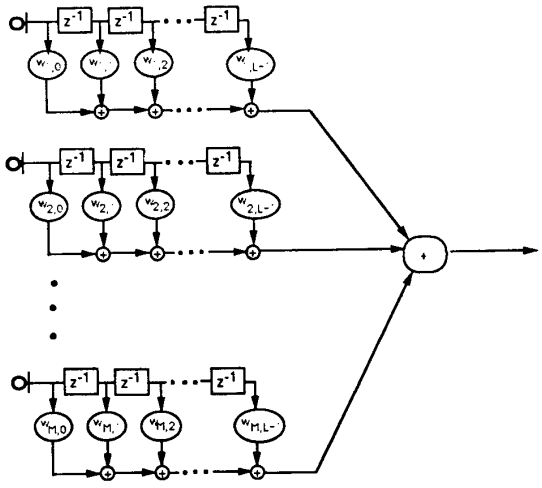

Figure (1) Broadband spatial filter structure.

This work was supported by NSF under contract MIP 9057071.

cancellation are based on the Articulation Index [Kryter 1962] and other intelligibility indices. The remainder of the paper is subdivided as follows. In section II LCMV processing is reviewed in the context of the microphone array problem. Also in II, the acoustic model is described. The performance measures mentioned above will be described in more detail in section III. Section IV presents the results of the computer simulations performed. Section V summarizes the main results.

II Array Observation Model

(A) LCMV array processor

Figure 1 illustrates the structure of an M sensor, L tap, spatial filter. Define

$$\mathbf{x}(n) = [\chi(n)^T, \chi(n-1)^T, \cdots, \chi(n-L+1)^T]^T$$

$$\text{and} \qquad \chi(n) = [x_1(n), x_2(n), \cdots, x_M(n)]^T$$

where ^{T} denotes transpose and $\chi(n)$, is the vector of output signals from each of the M sensors. $\mathbf{x}(n)$ is called the "stacked snapshot" vector, since it consists of the vector of the received signals at each microphone at a point in time, $\chi(n)$, stacked into a single vector. This vector, $\mathbf{x}(n)$, represents all of the received signal present in the beamformer at time instant n. For a stationary input into the array, the data covariance matrix is given by:

$$\mathbf{R} = E[\mathbf{x}(n)\mathbf{x}(n)^T] \tag{1}$$

The dimensions of this matrix are $ML \times ML$. The response of the array to a narrowband source from direction θ at frequency f is denoted by $\mathbf{a}(\theta,f)$. $\mathbf{a}(\theta,f)$, the array response vector, is an $ML \times 1$ vector consisting of the magnitude gains and phases for each of the samples of the "stacked snapshot" vector relative to the array phase reference at a frequency f and from a direction θ. For a stationary signal, $s(n)$, with power spectral density, $S(f)$, impinging on the array from direction θ, the data covariance can be obtained from the Wiener-Hopf relation as:

$$\mathbf{R} = \int_0^B S(f)[\mathbf{a}_c(\theta,f)\mathbf{a}_c(\theta,f)^T + \mathbf{a}_s(\theta,f)\mathbf{a}_s(\theta,f)^T]df \quad . \tag{2}$$

Here B is the bandwidth of interest and the subscripts c and s denote the real (cosine) and imaginary (sine) values used in processing real data.

Consider the standard problem of a linearly constrained minimum variance beamformer. Given an observed array data covariance matrix, $\mathbf{R}$, and a set of linear constraints on the array weights, $\mathbf{w}$, the problem statement is simply:

$$\min_{\mathbf{w}} \quad \mathbf{w}^T\mathbf{R}\mathbf{w} \tag{3}$$

$$\textit{subject.to} \quad \mathbf{C}^T\mathbf{w} = \mathbf{g}$$

where $\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, \cdots \mathbf{c}_J]$ is $ML \times J$ and contains as columns the set of J linear constraint vectors; the elements of $\mathbf{g}$ are the specified responses to each of the J constraints; $\mathbf{w}$ is $ML \times 1$; and $\mathbf{R}$ is $ML \times ML$, positive definite and symmetric. The well-known solution to (3) is given by:

$$\mathbf{w}_{opt} = \mathbf{R}^{-1}\mathbf{C}(\mathbf{C}^T\mathbf{R}^{-1}\mathbf{C})^{-1}\mathbf{g} \quad . \tag{4}$$

When the input data is spectrally white noise, the quiescent weight vector, $\mathbf{w}_q$, is given by:

$$\mathbf{w}_q = \mathbf{C}(\mathbf{C}^T\mathbf{C})^{-1}\mathbf{g} \quad . \tag{5}$$

The simplest form of linear constraints for power minimization beamforming would be a set of look direction constraints. These constraints fix the beamformer response in the direction of the desired signal. Subject to this fixed response, output power from the beamformer is minimized. [Cox et al. 1987] proposed a robustness constraint for practical array processing which limits the norm (or length) of the weight vector to prevent desired-signal cancellation due to array imperfections and misteering. This approach can be implemented by the structure shown in figure 2 and known as the Generalized Sidelobe Canceller (GSC). The weight vector $\mathbf{w}_q$ from figure 2 is fixed, while the weight vector $\mathbf{v}$ is subject to a norm constraint. See [Hoffman et al. 1990] for details of how this algorithm can be applied to the microphone array problem to inhibit signal cancellation due to misteering of the array. For the present study, no quadratic constraint has been incorporated.

(B) Acoustic model

Two primary components comprise the acoustic model used for the computer simulations of the hearing aid array. The first component is the room reverberation model which is based upon the method of images [Allen and Berkley,1979]. The second is the model of the effects of acoustic headshadow on the observations of the array of microphones. The simulations generate the composite pulse response from the speaker to each microphone mounted on the head of the listener.

[Allen and Berkley,1979] outline a basic procedure for using the method of images to model a reverberant room transfer function. The computed pulse response, $h_{tot}(n)$, from this method can be written as follows:

$$h_{tot}(n) = \sum_{i=0}^{N} \frac{\beta_i * \delta(n - [n_i]_{rnd})}{4\pi|\mathbf{R}_i|} \tag{6}$$

where, $\delta(n) = 1$ if $n = 0$ and $\delta(n) = 0$ otherwise; β_i is the aggregate frequency independent wall reflection attenuation for the ith reflection; $\mathbf{R}_i$ is the distance from the ith image source to the listener; and $[\,]_{rnd}$ indicates integer rounding. While (6) is sufficient for estimating the pulse response of a room from a speaker to a single microphone, more accurate accounting of the relative phase between microphones in an array is needed. [Peterson] incorporates this via a time domain interpolation for each reflection in (6). In the acoustic model employed in the following simulations proper accounting for phase is incorporated with the headshadow effects.

In [Hoffman and Buckley 1990], a model of acoustic headshadow was described. Data taken from psychoacoustical literature for the interaural pressure level and time delay differences was used to form array response vectors incorporating acoustic headshadow effects. These frequency domain data are used here to incorporate the headshadow effects into the room reverberation model. Figure 3 shows a block diagram of the general process. The frequency dependent time delays and pressure levels are stored in a two dimensional table with frequency and azimuthal angle of arrival as the dependent variables. For a specific azimuth, the values of pressure magnitude and time delay are interpolated linearly at each frequency. In the domains of pressure magnitude and time delay the headshadow data are quite smooth. An elevation correction is applied to these data for sources arriving from outside the azimuthal plane. For a specific angle of arrival, the time delay data and pressure magnitude data are coupled with the noninteger portion of the reverberation delay. The amplitude attenuation factors in (6) are incorporated. An IFFT of the data provides a relatively short time domain response which incorporates the frequency dependent effects of the acoustic headshadow for each reflection. The effects of the fractional time delay are included within the IFFT. While this represents a circular convolution in the time domain, the effects are negligible compared to other non-ideal aspects of the model. The overall room/head pulse response is the sum of the appropriately delayed individual contributions from each image source.

III Performance Measures

(A) Distortion and MSE

Distortion can be defined as the mean-squared difference between a desired signal and a received distorted desired signal. For the general case of a received distorted signal plus corrupting noise, this measure is simply the mean squared error (MSE) between the received signal and the original signal transmitted. In the case where no noise is present, the term distortion is sometimes used to characterize the effects that the communications channel has on the desired signal only. In evaluating the effects of reverberation on a desired speech signal without any interfering signals, this study employs this type of measure. Given a composite transfer function from source to listener, $H_C(f)$, the distortion that a channel introduces to a signal with power spectral density $S(f)$ is given by:

$$D = \int_0^B |H_C(f) - H_d(f)|^2 S(f)\,df \qquad (7)$$

where $H_d(f)$ is the desired channel response consisting of unity magnitude and a constant group delay over B.

For an array with a limited spatial aperture, better beamforming capability exists at higher frequencies. In addition, high level, low frequency sounds have a tendency to mask higher frequency sounds which can contain other information important to speech intelligibility. For these and other reasons, microphone arrays exhibit better performance when using pre-whitened speech versus cases without whitening. For the purposes of measuring distortion, this results in a frequency weighted MSE cost function. Distortion after pre-whitening can be defined as:

$$D_W = \int_0^B |H_C(f) - H_d(f)|^2 S_W(f)\,df \qquad (8)$$

here $S_W(f)$ is the whitened input spectrum. Rewriting (8) gives:

$$D_W = \int_0^B |H_C(f) - H_d(f)|^2 |H_W(f)|^2 S(f)\,df$$

$H_W(f)$ defines the preferential weighting of the distortion

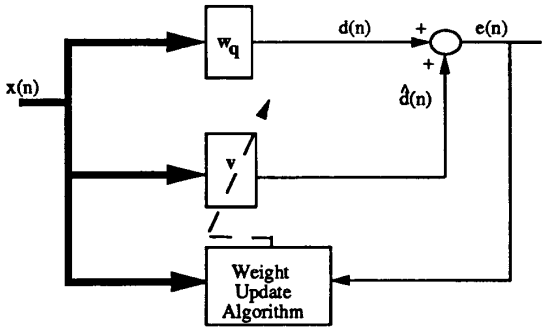

Figure (2) Generalized Sidelobe Canceller.

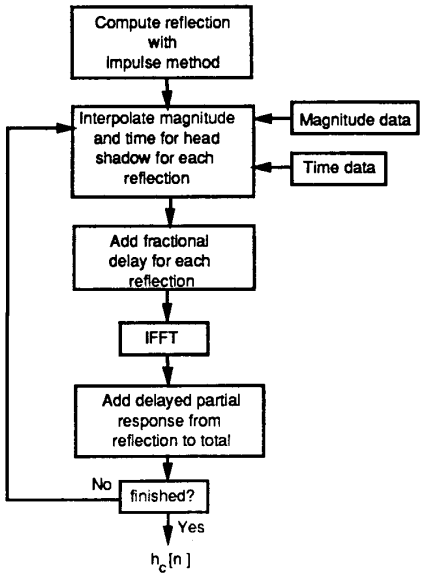

Figure (3) Room reverberation and headshadow model block diagram.

measure. In the present context it is the inverse of the long-time speech spectrum.

Another distortion measure can be defined which treats the first 50 msec or so of reverberated energy as helpful to speech intelligibility. [Houtgast et al. 1980] show a close relationship between this good/bad reverberation model and the Speech Transmission Index (STI). All signal energy arriving within 50 msec of the direct path defines the desired signal, while all other reverberation paths contribute to distortion. The effectiveness of STI in predicting intelligibility in reverberant situations and it's relationship to this definition of distortion motivate the following definition:

$$D_{50} = \int_0^B |H_C(f) - H_{50}(f)|^2 S_W(f)\,df \qquad (9)$$

$H_{50}(f)$ represents the reverberation transfer function for the first 50 msec of the pulse response. The input signal is again assumed to be pre-whitened.

(B) Articulation Index measures

The Articulation Index (AI) [Kryter 1962] was developed to provide an objective measure of speech intelligibility for a given physical situation. Many empirically based adjustments are made to provide accurate predictions of speech intelligibility under certain situations (e.g., level dependent masking). The use of a simplified version of AI fails to incorporate several important features which are, in fact, integral parts of the Articulation Index. The above caveat having been made, AI is used in the present investigation to indicate the effect the desired signal cancellation will have on the intelligibility of the received/cancelled speech.

[Peterson 1989] developed a measure G_{AI}, called intelligibility weighted gain, to quantify the effects that interfering speech has on the intelligibility of desired speech processed with a microphone array. The incorporation of some features of the Articulation Index in G_{AI} allows the prediction of intelligibility increases when a microphone array replaces a single omni-directional microphone. The application of an AI-based performance measure in the present study is similarly motivated. A simplified version of AI can be defined as follows:

$$\mathbf{AI} = \sum_{i=1}^{15} w_{AI,i}\, 10 * log_{10}([SNR_i]_{1/3})_{[0,30]} \qquad (10)$$

where

$$[SNR_i]_{1/3} = \frac{\int_{1/3oct_i} |S(f)|\,df}{\int_{1/3oct_i} |N(f)|\,df}$$

is the signal to noise ratio of the third octave averaged speech and noise (or distortion) spectra, $S(f)$ and $N(f)$, respectively. The $w_{AI,i}$ in (10) are the empirically derived [Kryter 1962] third octave weights applied to 15 third octave bands of speech. Note that the weighted sum of signal to noise ratios is applied to the dB value of the smoothed SNR clipped to fall between 0 and 30 dB. AI is a value between 0 and 1. AI = 0 corresponds to no intelligibility, AI = 1, to very high intelligibility.

Two distortion measures, D_W and D_{50}, will be used to generate two AI-based measures. AI_{MS} is the AI-measure computed when D_W distortion, i.e., the weighted MSE distortion, is considered the noise in (10). Specifically,

$$[SNR_{MS,i}]_{1/3} = \frac{\int_{1/3oct_i} S_W(f)\,df}{\int_{1/3oct_i} |H_C(f) - H_d(f)|^2 S_W(f)\,df} \qquad (11)$$

AI_{50} is the AI-measure computed when D_{50} distortion is considered noise and $|H_{50}(f)|^2 S_W(f)$ is considered the signal:

$$[SNR_{50,i}]_{1/3} = \frac{\int_{1/3oct_i} |H_{50}(f)|^2 S_W(f)\,df}{\int_{1/3oct_i} |H_C(f) - H_{50}(f)|^2 S_W(f)\,df} \qquad (12)$$

AI_{MS} and AI_{50} are both computed as per AI in (10). Note that since AI is a function of SNR, pre-whitening has no direct effect on it. However, pre-whitening does effect the LCMV processor, hence the processor output spectra, hence AI.

IV Results

Computer simulations were run on a five microphone, head worn array with the microphones evenly spaced from the left ear to the right ear on an arc around the head. Two reverberant rooms were modelled: both the living room and conference room described in [Peterson 1989]. Average wall absorptions of 0.3, 0.6, and 0.9, along with speaker-to-listener distances ranging from 0.9m to 4.6m, provided simulated reverberant environments with Direct-to-Reverberant energy ratios ranging from -14 to +12 dB. The modelled acoustic headshadow described above was also incorporated. Exact covariances of stationary, pre-whitened signals of interest produced optimum weights for an LCMV beamformer. The number of taps in the processor were varied from 8 to 60, with a sample rate of 10 kHz. The number of linear look direction constraints varied in proportion with the number of taps. Twelve constraints were employed in cases of 8 taps, while 66 were used in cases of 60 taps. While no quadratic constraint was employed, a white noise at a level of ~50 dB below the maximum data covariance eigenvalue was added to ensure numerical stability.

Figure 4 presents computed values of AI_{50} and AI_{MS} for a 30 tap processor over a range of 7 different direct-to_reverberant energy ratios. AI_{50} and AI_{MS} were each computed for two weight vectors, the optimum processor and the quiescent (see Eqn (5)). On the high end of the Dir/Rev

axis, the AI_{50} measures predict excellent intelligibility (AI=1), in a situation where it is expected, a fixed weight vector, $\mathbf{w}_q$, in a room with a short reverberation time of <100 msec. The AI_{MS} measure, however, predicts rather poor intelligibility (AI=0.4). For the optimum LCMV weight vector, the performance predicted is very similar to that for the quiescent in both cases. For all combinations of tap length and Dir/Rev, the AI_{MS} provided what seem to be overly pessimistic predictions of intelligibility. With neither measure, however, is there any evidence of a significantly cancelled signal of interest as evidenced by decrease in performance with optimum over quiescent processor. Again this finding was consistent over all trials.

In addition, for the range of tap lengths considered, there was no relationship evidenced between an increase in tap length and a subsequent increase in signal cancellation as depicted in figure 5. While these results seem at odds with [Greenberg and Zurek 1991], it is not inconsistent considering that they adaptively processed speech data, used long tap lengths (100 and 400), and applied a different performance measure (G_{AI}). While longer tap lengths increase the nulling capability of the array, that nulling is achieved at the cost of greater computational complexity and a greater susceptibility to cancellation of the desired signal.

As a final comparison, the AI_{50}'s from the 30 tap processor (shown in figure 4) are replotted in figure 6 against the predicted AI_{50} for the single microphone positioned at the right ear. AI_{50} for the optimum and quiescent are higher throughout the transition range of Dir/Rev values. The consistency of AI_{50} in predicting good intelligibility with a single microphone in the high Dir/Rev cases suggests a useful test case in which to process speech.

V Summary

While high amounts of weighted mean squared error suggest significant amounts of reverberation-related distortion, simulation results did not suggest any significant signal of interest cancellation for an optimum LCMV processor using up to 60 taps. An interesting AI-based performance measure was applied to the simulated room reverberation and headshadow models. Application of AI_{50} to processed speech signals of interest will provide a better idea of its utility.

REFERENCES

[1977] J.B Allen, D.A. Berkley,and J. Blauert, "Multi-microphone signal processing technique to remove room reverberation from speech signals," *Journ. of the Acoust. Soc. of Am.*, vol. 62, No. 4, pp. 912-915, Oct. 1977.

[1989] P.M. Peterson, "Adaptive Array Processing for Multiple Microphone Hearing Aids," *MIT Ph.D. Dissertation, RLE Technical Report No. 541*, Feb., 1989.

[1990] M.W. Hoffman and K.M. Buckley, "Constrained optimum filtering for multi-microphone digital hearing aids," *Proc. Twenty-Fourth Asilomar Conf. on Signals, Systems and Computers*, Nov. 1990.

[1991] J.E. Greenberg and P.M. Zurek, "Adaptive Beamformer Performance in Reverberation," *Proc. 1991 IEEE ASSP Workshop on Appl. of Sig. Proc. to Audio and Acoustics*, Oct., 1991.

[1987] H.Cox, R.M. Zeskind, and M.M. Owen, "Robust adaptive beamforming," *IEEE Trans. on Acoust., Speech and Signal Processing*, vol. ASSP-35, pp. 1365-76, Oct., 1987.

[1979] J.B Allen and D.A. Berkley, "Image method for efficiently simulating small room acoustics," *Journ. of the Acoust. Soc. of Am.*, vol. 65, No. 4, pp. 943-950, Apr. 1979.

[1962] K.D. Kryter, "Methods for the calculation and use of the articulation index," *Journ. of the Acoust. Soc. of Am.*, vol. 34, No. 11, pp. 1689-1697, Nov. 1962.

[1980] T. Houtgast, H.J.M Steeneken and R. Plomp, "Predicting Speech Intelligibility in Rooms from the Modulation Transfer Function. I. General Room Acoustics," *Acustica*, vol. 46, 1980, pp. 60-72.

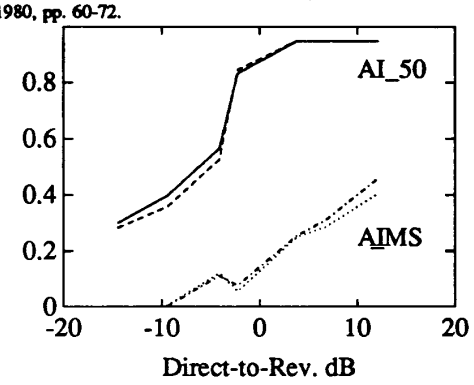

Figure (4) Two AI measures for $\mathbf{w}_q$ and w_{opt} weight vectors. 30 tap per element, 5 element processor.

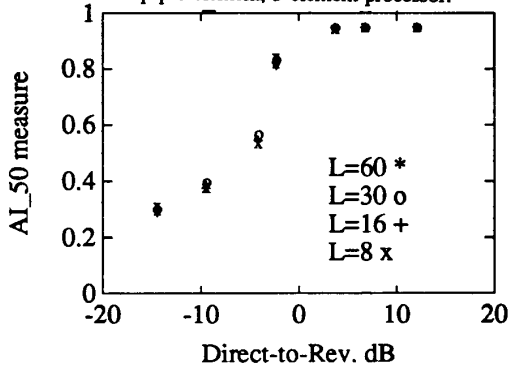

Figure (5) AI_{50} for 8, 16, 30 and 60 taps per element.

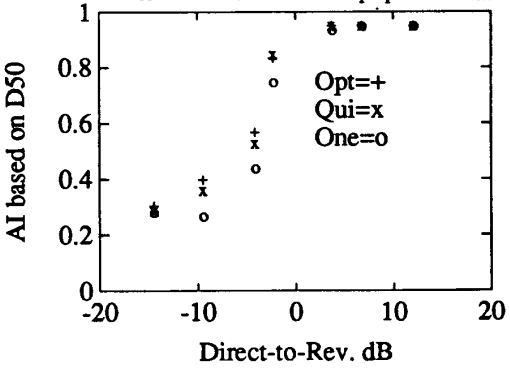

Figure (6) AI_{50} for 30 tap optimum (+) and quiescent (x) processors and single microphone (o).