Use of Intermicrophone Correlation in Estimating Signal to Noise Ratio

by

Ashish Koul

Submitted to the Department of Electrical Engineering and Computer Science in partial fulfillment of the requirements for the degrees of Bachelor of Science in Electrical Engineering and Computer Science and Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2003

© Massachusetts Institute of Technology 2003. All rights reserved.

Author

Department of Electrical Engineering and Computer Science

February 4, 2003

Certified by.

Julie E. Greenberg
Research Scientist
Thesis Supervisor

Accepted by

Arthur C. Smith
Chairman, Department Committee on Graduate Students
Use of Intermicrophone Correlation in Estimating Signal to Noise Ratio

by

Ashish Koul

Submitted to the Department of Electrical Engineering and Computer Science on February 4, 2003, in partial fulfillment of the requirements for the degrees of Bachelor of Science in Electrical Engineering and Computer Science and Master of Engineering in Electrical Engineering and Computer Science

Abstract

This thesis presents the design, analysis, and simulation of a system that uses the correlation coefficient of audio inputs gathered at two spatially separate microphones to determine the signal to noise ratio in the environment. This work is motivated by past research in microphone array hearing aids, where accurate estimates of SNR were shown to improve performance. Signal to noise ratio is defined as the ratio of energy in the direct component (audio sources originating in front of a broadside array) to energy in the interference component (sources originating from the sides of the array). The design presented is a simple hypothesis testing mechanism for determining whether the SNR exceeds a fixed level. In the analysis, behavior of the system is studied theoretically under varying conditions of reverberation in the environment, and processing parameters are determined to optimize system performance. Finally, simulations test the true performance of the system to verify the validity of the theoretical analysis.

Thesis Supervisor: Julie E. Greenberg
Title: Research Scientist
Acknowledgments

I could not have completed this research without the support and guidance of Julie Greenberg, who supervised this thesis. All my knowledge of microphone array hearing aids and speech signal processing comes from her expertise. I must credit Pat Zurek for his intellectual contributions to this project. He suggested the use of the Fisher Z-Transformation and outlined the derivation presented in Section 2.4. I also must credit my friend and colleague Baris Erkmen for going over the theoretical details of this research with me to ensure that the logic and math were in good shape. Thanks are due to my friend Joanne and my roommate Sandy for moral support and motivation. Finally, I express my gratitude to Meena, who gave me something to look forward to at the day’s end.
Chapter 1

Introduction

A conventional hearing aid is designed such that all incoming sound is amplified equally, regardless of the direction of origin. A common complaint amongst users of hearing aids is that hearing aids are not helpful when listening to speech in a crowded or noisy environment. One approach which has been proposed to resolve this problem is the use of microphone array hearing aids which selectively amplify sound based on its direction of arrival.

Greenberg and Zurek (1992) proposed a microphone-array hearing-aid system which utilized an adaptive beamforming algorithm. In general, it was found that the performance of the adaptive algorithm degrades when the signal-to-noise ratio, SNR, exceeds some limit. During typical conversation, the power in a speech signal tends to fluctuate within a range of approximately 30 dB. This leads to intermittent periods during which the SNR may exceed the limit. In order to control the behavior of the algorithm, a measure of intermicrophone correlation was used to determine if the SNR was "low" or "high" for a given time interval. In the case of a "high" decision, adaptation was inhibited for that time interval.

The simplest microphone array hearing-aid consists of a microphone mounted at each of the ears - a two-microphone broadside configuration. This is seen in Figure 1-1. The incident angle of an audio source is measured from the vertical axis in the clockwise direction. All sources and microphones in the scene are assumed to lie in the same plane. In general, sound can arrive from any incident angle in the full interval \([-\pi, \pi]\). Because the array shown in Figure 1-1 only uses two microphones, there is an implied symmetry between sources in front of the array and those behind. Thus, all analysis can be limited to the interval \([-\pi/2, \pi/2]\).
This scenario considers a potential hearing aid in which microphones are mounted on either side of a person’s head. It assumes that the focus of one’s attention is in the direction of sight, directly in front of the head. Thus, a source with an incident angle of 0 is considered to be “straight ahead”. Note that due to the front/back symmetry in this scenario, sources directly behind the array will be treated in the same manner as those directly ahead. This is an inherent limitation of any broadside microphone configuration. Sources with angle of incidence in the interval \([-\theta_0, \theta_0]\) are assumed to be desired signals, while sources with angle of incidence in either of the intervals \([-\pi/2, -\theta_0]\) or \([\theta_0, \pi/2]\) are assumed to be interfering noise. Because of the symmetry in the definition of signal versus noise, the analysis can be further limited to the interval of incident angle \([0, \pi/2]\) (See Section 2.4). This simplified situation is shown in Figure 1-1.

The basic structure of the SNR testing system utilizing the metric of intermicrophone correlation is given in Figure 1-2. The filter \(H(f)\) in the above block diagram is an ideal bandpass filter, as given in Figure 1-3. The value \(r_0\) is a constant threshold used for hypothesis testing. The bandpass filter is applied to the inputs in this system to take advantage of the characteristics of the microphone array. The range of input frequencies can be chosen such that the desired signal will yield a maximum positive correlation, and the interfering noise will yield a maximum negative correlation (frequency can be selected
such that the noise waves received at the two microphones are completely out of phase. Thus, distinguishing between signal and noise will be easier than if the full bandwidth of the speech inputs was used.

The system described above is practical for applications in which computing resources are a limiting factor, such as hearing aids. Correlation coefficient can be determined with minimal computational overhead. This purpose of this study is to analyze the performance of the above system under varying conditions in order to determine optimal system parameters (bandpass filter center frequency, bandwidth, and threshold) and estimate optimal system performance.
Chapter 2

Theory

2.1 Variable Definition

In the case of a two-microphone array in free space, the two measured input variables after bandpass filtering can be defined as follows.

\[ x_1(t) = s(t) + n(t) \]  \hspace{1cm} (2.1)

\[ x_2(t) = s(t - \tau_s) + n(t - \tau_n) \]  \hspace{1cm} (2.2)

Each input consists of the summation of a signal component and a noise component, both of which have already passed through the bandpass filter. \( \tau_s \) and \( \tau_n \) represent the time delays for the desired signal and interfering noise, respectively. Note that in these expressions, the variables are regarded as continuous time rather than discrete time functions. While the actual system is to work in discrete time (see Figure 1-2), at this point in the analysis we consider the inputs \( x_1(t) \) and \( x_2(t) \) - the continuous time bandlimited reconstructions of \( x_1(n) \) and \( x_2(n) \). Assuming plane wave propagation, either \( \tau \) value can be expressed as follows,

\[ \tau = \frac{d}{c} \sin(\theta) \]  \hspace{1cm} (2.3)

where \( d \) is the distance separating the microphones, \( c \) is the speed of sound, and \( \theta \) is the incident angle of the source. Let the correlation coefficient \( \rho \) be defined as follows

\[ \rho = \frac{E[x_1(t)x_2(t)]}{\sqrt{E[x_1^2(t)]E[x_2^2(t)]}} \]  \hspace{1cm} (2.4)
In this expression, $\rho$ is the “true” correlation coefficient. In this application, we can only obtain an estimate of $\rho$ using discrete samples of the two input processes $x_1(t)$ and $x_2(t)$ taken over a finite time period. All first and second order statistics are estimated using discrete summation, and using these statistics the “sample” correlation coefficient $r$ is determined.

$$r = \frac{\sum_n x_1(n) x_2(n)}{\sqrt{\sum_n x_1^2(n) \sum_n x_2^2(n)}}$$  \hspace{1cm} (2.5)

The measured value of $r$ is used as the decision variable in this analysis.

### 2.2 Assumptions

The following initial assumptions are made to enable a tractable analysis.

1. All “signal” sources are defined to be arriving at the microphone array at an incident angle in the range $-\theta_0$ to $\theta_0$, and “noise” sources are defined to arrive at the microphone array at an incident angle either between $-\pi/2$ and $-\theta_0$ or $\theta_0$ and $\pi/2$. For either a source or noise signal, the probability distribution of the incident angle is uniform over the appropriate range. (This assumption can be extended to the full range of $-\pi$ to $\pi$ without loss of generality.)

2. All audio sources in question are accurately modeled as continuous, zero-mean, white gaussian noise processes. Signal and noise processes have variances $\sigma_n^2$ and $\sigma_w^2$, respectively. Signal-to-noise ratio is defined as $SNR = 10 \log_{10}(W)$ where $W = \frac{\sigma_n^2}{\sigma_w^2}$.

3. Reverberation exists in the environment, and is assumed to behave as a spherically-diffuse sound field. Reverberant sound is characterized by the direct-to-reverberant ratio $DRR = 10 \log_{10}(\beta)$, where $\beta$ is the ratio of energy in the direct wave to energy in the reverberant sound. We assume $\beta$ to be equal for both signal and noise sources.

4. The band pass filters used in pre-processing incoming signals are ideal band-limited filters with center frequency $f_0$ and bandwidth $B$.

In order to use the measure of sample correlation coefficient for hypothesis testing, the probability distribution of $r$ must be derived. To facilitate this derivation, we introduce the Fisher $Z$-Transformation.
2.3 The Fisher Z-Transformation

Assume that two random variables \( a \) and \( b \) are drawn from a bivariate gaussian distribution. We wish to obtain an estimate of their “true” correlation coefficient \( \rho \) using \( N \) sample pairs drawn from the joint distribution of \( a \) and \( b \). Let the “sample” correlation coefficient be denoted as \( r \). In general, the probability distribution of the estimator \( r \) is difficult to work with directly, because its shape depends on the value of \( \rho \).

The Fisher Z-Transformation \([3]\) is defined as

\[
    z = \tanh^{-1}(r) = \frac{1}{2} \ln\left(\frac{1 + r}{1 - r}\right).
\]

(2.6)

This yields the new random variable \( z \) which has an approximately gaussian distribution with mean \( \frac{1}{2} \ln\left(\frac{1 + \rho}{1 - \rho}\right) \) and variance \( \frac{1}{N - 3} \). This derived variable \( z \) has a simple distribution whose shape does not depend on the unknown value of \( \rho \).

Due to the assumption from the previous section that both the noise and signal sources are gaussian random processes, the measured microphone inputs must be jointly gaussian random processes. Furthermore, even after applying the bandpass filter, the input variables \( x_1(t) \) and \( x_2(t) \) as defined in (1) and (2) must still be jointly gaussian. Thus, the Fisher Z-Transformation may be applied to sample data gathered from the two-microphone array after bandpass filtering.

2.4 Derivation of PDF

We derive the probability density function of \( r \) in a scenario containing one signal and one noise source, following the derivation outlined in [10]. We begin by finding the PDF in the case of a single source. After discrete sampling and bandpass filtering (See Figure 1-2), \( x_1(n) \) and \( x_2(n) \) are rectangular bands of noise. Given an incident angle of \( \theta \), the true intermicrophone correlation of the filtered noise is given by [2],

\[
    \rho \theta = \frac{\cos(kd \sin \theta) \sin \left(\frac{\pi Bd}{c} \sin \theta\right)}{\left(\frac{\pi Bd}{c} \sin \theta\right)}.
\]

(2.7)
where the value $k$ is the wavenumber,

$$k = \frac{2\pi f_0}{c}. \quad (2.8)$$

Note that the above expression for $\rho_0$ is an even function of $\theta$. Thus, in the remainder of this analysis the range of incident angles can be limited to $[0, \pi/2]$ without loss of generality. Let us define $\bar{z}(\theta) = \frac{1}{2} \ln\left(\frac{1+\rho_0}{1-\rho_0}\right)$ and $\sigma_z^2 = \frac{1}{N-3}$, as described in the previous section. Thus, the conditional PDF of $z$ given a source at incident angle $\theta$ is

$$f_{z|\theta}(z|\theta) = \frac{1}{\sigma_z \sqrt{2\pi}} \exp\left(-\frac{(z-\bar{z}(\theta))^2}{2\sigma_z^2}\right) \quad (2.9)$$

The angle $\theta$ is uniformly distributed over a specific range of values, depending on the source. The expression for the joint density of $z$ and $\theta$ is

$$f_{z,\theta}(z, \theta) = \frac{1}{\theta_2 - \theta_1} f_{z|\theta}(z|\theta) \quad (2.10)$$

where $\theta_2 = \theta_0$ and $\theta_1 = 0$ for a signal source and $\theta_2 = \pi/2$ and $\theta_1 = \theta_0$ for a noise source.

To obtain the marginal density of $z$, the joint density is integrated over the appropriate range of $\theta$.

$$f_z(z) = \frac{1}{(\theta_2 - \theta_1)\sigma_z \sqrt{2\pi}} \int_{\theta_1}^{\theta_2} \exp\left(-\frac{(z-\bar{z}(\theta))^2}{2\sigma_z^2}\right) d\theta \quad (2.11)$$

Given the density of $z$, we can use the definition of the Fisher Z-transformation to derive the density of the sample correlation coefficient $r$. Since $z = \tanh^{-1}(r)$ is a monotonic transformation of the random variable $r$, the PDF of $r$ can be obtained through the following relation [4].

$$f_r(r) = f_z(z) \frac{dz}{dr} \quad (2.12)$$

It is known that $\frac{dz}{dr} = \frac{1}{1-r^2}$. Appropriate substitution leads to the PDF of $r$.

$$f_r(r) = \frac{1}{(1-r^2)(\theta_2 - \theta_1)\sigma_z \sqrt{2\pi}} \int_{\theta_1}^{\theta_2} \exp\left(-\frac{[\tanh^{-1}(r) - \bar{z}(\theta)]^2}{2\sigma_z^2}\right) d\theta \quad (2.13)$$

Note that this PDF only applies to the case of a single source (either noise or signal). The next step is to derive the PDF in the scenario of one noise and one signal present together.
The full expression for the sample correlation coefficient, as described in Section 2.1, is

\[ r = \frac{\sum_k [s(k) + n(k)][s(k - \tau_s) + n(k - \tau_n)]}{\sqrt{\sum_k [s(k) + n(k)]^2} \sum_k [s(k - \tau_s) + n(k - \tau_n)]^2} \]  

(2.14)

where \( s(k) \) is the direct signal and \( n(k) \) is the noise component. The corresponding expression for the direct signal component alone is

\[ r_s = \frac{\sum_k [s(k)s(k - \tau_s)]}{\sqrt{\sum_k [s(k)]^2} \sum_k [s(k - \tau_s)]^2} \]  

(2.15)

and for the noise component alone is

\[ r_n = \frac{\sum_k [n(k)n(k - \tau_n)]}{\sqrt{\sum_k [n(k)]^2} \sum_k [n(k - \tau_n)]^2} \]  

(2.16)

At this point we must make the following assumptions.

1. The \( s \times n \) cross terms in the above summation are negligible when compared with the \( s \times s \) and \( n \times n \) terms to which they add.

2. The effect of time delay on the energy can be ignored such that the following is true.

\[ \sum_k s^2(k) \approx \sum_k s^2(k - \tau_s) \]  

(2.17)

\[ \sum_k n^2(k) \approx \sum_k n^2(k - \tau_n) \]  

(2.18)

3. \( W \), as defined in Section 2.2, can be expressed as a ratio of energies.

\[ W = \frac{\sum_k s^2(k)}{\sum_k n^2(k)} \]  

(2.19)

Using the first two assumptions, we can express the intermicrophone correlation as

\[ r = \frac{\sum_k [s(k)s(k - \tau_s)] + \sum_k [n(k)n(k - \tau_n)]}{\sum_k [s^2(k)] + \sum_k [n^2(k)]} \]  

(2.20)

By then substituting the expressions for \( r_s \) and \( r_n \), we obtain

\[ r = \frac{r_s \sum_k [s^2(k)] + r_n \sum_k [n^2(k)]}{\sum_k [s^2(k)] + \sum_k [n^2(k)]} \]  

(2.21)
Finally, by dividing all terms in the expression by $\sum_k[n^2(k)]$, we derive the following.

$$
    r = \frac{Wr_s + r_n}{W + 1} = \frac{W}{W + 1}r_s + \frac{1}{W + 1}r_n
$$

(2.22)

Thus, we have an equation which expresses the full intermicrophone correlation as a linear combination of the separate correlations for signal and noise. The PDFs of both $r_n$ and $r_s$ can be obtained using equation (2.13). The probability density of their linear combination can be obtained by convolution as follows

$$
    f_{r|W}(r) = [\frac{W + 1}{W}f_{r_s}(\frac{W + 1}{W}r)] * [(W + 1)f_{r_n}((W + 1)r)]
$$

(2.23)

Therefore, conditional on a given value of SNR, we can use the above expression as an approximate PDF of the intermicrophone correlation $r$. 
Chapter 3

Reverberation

Up to this point, our probabilistic model of intermicrophone correlation has only considered the direct wave components of sound. To further generalize the model, reverberation must be incorporated into the PDF. The reverberant sound component is assumed to be statistically independent of both the direct components. Furthermore, the energy in the reverberant component is characterized by the direct-to-reverberant ratio, as described in Section 2.2.

Using the given assumptions, we can apply an argument analogous in structure to that given in the previous chapter to derive the expression for total intermicrophone correlation.

\[ r = \frac{\beta r_d + r_r}{\beta + 1} = \frac{\beta}{\beta + 1} r_d + \frac{1}{\beta + 1} r_r \]  

(3.1)

As before, the total correlation is equal to a linear combination of its components. In this expression, the signal and noise components are replaced by the direct and reverberant components \( r_d \) and \( r_r \), respectively. \( W \), the ratio of signal energy to noise energy, is replaced by \( \beta \), the constant ratio of direct component energy to reverberant energy.

From this we can define the PDF of total correlation just as before.

\[ f_r(r|\beta) = \left[ \frac{\beta + 1}{\beta} f_{r_d}(\frac{\beta + 1}{\beta} r) \right] * \left[ (\beta + 1) f_{r_r}((\beta + 1)r) \right] \]  

(3.2)

This expression is a convolution of the direct component PDF and the reverberant component PDF. The direct component PDF \( f_{r_d}(r) \) is given by equation (2.23), derived at the end of the previous chapter. In order to derive the PDF of the reverberant component \( f_{r_r}(r) \),
we follow a procedure similar to that used before.

In the previous chapter, the value of the true intermicrophone correlation $\rho$ for the direct signal was dependent on the incident angle $\theta$ of the source (see equation 2.7). We assume reverberation in this environment to behave as a spherically diffuse sound field, thus the true correlation of the reverberant component $\rho_r$ will have no dependence on incident angle. Let us first assume that all audio sources in the environment produce only a single pure tone. In this case, the reverberant component would consist of only this tone. The true intermicrophone correlation coefficient of a pure tone in a diffuse sound field is given by

$$\rho_r = \frac{\sin(kd)}{kd},$$  \hspace{1cm} (3.3)

where $k$ is the wavenumber, and $d$ is the microphone spacing [5]. This applies to the total reverberant component, including both signal and noise reverberation. In our scenario, rather than pure tones all sound sources are bandlimited gaussian white noise processes. The previous expression can be modified to account for this yielding the following simple closed-form approximation [6]

$$\rho_r = \text{sinc}(\frac{\pi Bd}{c}) \frac{\sin(kd)}{kd},$$ \hspace{1cm} (3.4)

where $B$ is the bandwidth, $c$ is the speed of sound, and $\text{sinc}(x) = \frac{\sin(x)}{x}$. From this point forward, this expression will be used as the true intermicrophone correlation for reverberant sound $\rho_r$.

Now we apply the definition of the Fisher Z-Transformation as given in Section 2.3.

$$z = \tanh^{-1}(r_r) = \frac{1}{2} \ln\left(\frac{1 + r_r}{1 - r_r}\right).$$ \hspace{1cm} (3.5)

Here $z$ is defined as a function of $r_r$, the reverberant sample correlation coefficient. As before, the random variable $z$ will have an approximately gaussian distribution, with $\overline{z} = \frac{1}{2} \ln\left(\frac{1 + \rho_r}{1 - \rho_r}\right)$ and $\sigma_z^2 = \frac{1}{N-3}$.

$$f_z(z) = \frac{1}{\sigma_z \sqrt{2\pi}} \exp\left(-\frac{[z - \overline{z}]^2}{2\sigma_z^2}\right)$$ \hspace{1cm} (3.6)

Finally, by applying the transformation of equation (2.12) to the above expression we arrive
at the PDF of intermicrophone correlation for the reverberant component.

\[ f_{r}(r) = \frac{1}{(1 - \tau^2)\sigma_{z}\sqrt{2\pi}} \exp\left(-\frac{[\tanh^{-1}(r) - \bar{z}]^2}{2\sigma_{z}^2}\right) \]  

(3.7)

This can be combined with the PDF for the direct wave component using the convolution relation in equation (3.2) to obtain the PDF for total intermicrophone correlation.
Chapter 4

Hypothesis Testing

The goal of the system under consideration is to distinguish between two potential situations: "low" SNR and "high" SNR, denoted $H_0$ and $H_1$, respectively. The previous analysis was done under the assumption that all sources were white gaussian noise processes. However, the true system is intended to work with speech to detect intervals of high and low SNR which occur due to the natural fluctuations in speech. During typical conversation, it has been empirically determined that speech signal power tends to fluctuate within the range $(X+12)$ dB and $(X-18)$ dB, where $X$ is the average signal power. Tracking of power fluctuations can be done reliably using a temporal window of roughly 10 msecs, due to the nature of phonemes in speech. For the purposes of analysis and testing, we will define $H_0$ to be $10 \log(W) < 0$ dB and $H_1$ to be $10 \log(W) > 0$ dB. The choice of 0 dB as the cutoff point is related to the adaptive algorithm mentioned in Chapter 1 [1]. Typically, degradation of the algorithm's performance becomes noticeable at positive values of SNR.

In the previous chapter $W$ was treated as a known constant, but for the purpose of formulating a hypothesis test it is regarded as a random variable. Thus, it becomes necessary to know an approximate probability distribution for $W$. In this study, we will make the assumption that the SNR is uniformly distributed between -20 dB and 20 dB. Therefore, the variable $U = 10 \log(W)$ is uniformly distributed between -20 and 20. Under this assumption, the two hypotheses $H_0$ and $H_1$ both have equal prior probability, so the hypothesis test reduces to a Maximum Likelihood situation.

To derive the conditional PDF of $r$ under either hypothesis, the conditional expression
obtained in the previous chapter must be integrated over the appropriate range.

\[ f_{r|H_0}(r|H_0) = \int_{-20}^{0} f_{r|W}(r|W)dU \]  \hspace{1cm} (4.1)

\[ f_{r|H_1}(r|H_1) = \int_{0}^{20} f_{r|W}(r|W)dU \]  \hspace{1cm} (4.2)

To evaluate these expressions, the substitution \( W = 10^{\frac{U}{10}} \) must be made. The issue of hypothesis testing will be considered in more detail in the following chapter.
Chapter 5

Analysis

5.1 Frequency and Bandwidth Considerations

In this section we present an analysis of the proposed model. The functional parameters to be considered here are center frequency $f_0$, filter bandwidth $B$, and correlation threshold $r_0$. The issues involved in selecting each of these parameters will be addressed in turn. In order to perform this analysis, we produced approximate calculations of the intermicrophone correlation PDFs described in chapter 3 (equations 3.2 and 3.7) - all approximations were determined by applying Simpson's rule [7] to perform numerical integration and numerical convolution.

Let us first in general describe how the parameters affect the structure of the intermicrophone correlation PDF. In Figure 5-1, an array of PDFs is given. The term which we consider as the center frequency variable is the expression $kd$, the product of wavenumber and intermicrophone distance. The value of $d$ is fixed at 0.14 meters in this analysis, so $kd$ is simply a fixed linear function of the center frequency $f_0$. In addition, the filter bandwidth $B$ will be defined as some fraction of the center frequency $f_0$ throughout this analysis, thus we will use the term fractional bandwidth to denote the expression $\frac{B}{f_0}$.

In Figure 5-1 we are assuming an anechoic environment and are using a fixed fractional bandwidth of 0.22. In the left column, PDFs are shown for the circumstance in which the signal energy is predominant ($W = 1000$), while the right column shows the opposite extreme, where the noise energy is predominant ($W = 0.001$). Each row corresponds to a single value of $kd$. From this figure, we can see that in the case of high SNR, the intermicrophone correlation is concentrated around 1, while for low SNR it is concentrated
toward the opposite extreme.

The system proposed in this study uses intermicrophone correlation as a hypothesis testing variable to distinguish between the situations of high SNR and low SNR. Thus, in order to obtain good performance from the system, the parameters should be chosen such that the PDFs representative of high SNR and low SNR (the left and right columns respectively) are concentrated as far apart as possible and have minimal overlap. From this point forward, the PDFs corresponding to high SNR and low SNR will be referred to as the signal and noise PDFs, respectively.

In this array of graphs, we can see in general the effects of varying the center frequency parameter $kd$. From the left column, it is apparent that for any choice of frequency the signal PDF is rather heavily weighted towards 1. This makes sense, since the source origin in this case is in front of the microphone array, and thus the signal will always yield a strong positive correlation for any choice of $kd$. From the right column, we see that the structure of the noise PDF is strongly affected by the value of $kd$. In particular, for $kd = \pi$ the noise PDF is sharply concentrated at -1. The reason for this is that when $kd = \pi$, noise originating directly from the side of the microphone array (where the incident angle is $\pi/2$) will yield a perfect negative correlation - the intermicrophone distance is equal to exactly half the wavelength. Deviating the center frequency parameter from this causes the noise PDF to shift its mass away from -1.

At first glance, it would seem that the choice of $kd = \pi$ would be optimal for hypothesis testing, but on closer inspection we see that the noise PDF for $kd = \pi$ spans a rather large range. The tail end of the PDF reaches values very close to 1. Since PDF overlap adversely affects the performance of hypothesis testing, this long tail will cause an increase in probability of error. However, if we look at the noise PDF corresponding to the choice of $kd = 4\pi/3$, even though the mass of the PDF is not so sharply concentrated at -1, the PDF tail is greatly reduced. An optimal selection of center frequency must balance the needs for strong negative correlation in the noise PDF as well as overlap minimization. The choice of this parameter will be discussed further shortly.

In the array of PDFs given in Figure 5-2, the center frequency $kd$ is held constant at $\pi$, and the rows correspond to increasing fractional bandwidth. Again we are assuming an anechoic environment. This figure shows the general effect of increasing the filter bandwidth. The signal PDF (left column) is largely unaffected by bandwidth, because the signal
Array of PDFs for $kd=\left[\frac{2\pi}{3}, \pi, \frac{4\pi}{3}\right]$

Figure 5-1: In this set of intermicrophone correlation PDFs, the fractional bandwidth is fixed at 0.22. The environment is assumed anechoic.

... originates from directly in front of the microphone array. The signal will always yield strong positive correlation for any range of frequency. However, for the noise PDF the effects are substantial. As the bandwidth increases, the mass of the PDF contracts away from -1. This contraction effect occurs for any value of $kd$, and increasing bandwidth will in general have a detrimental effect on theoretical system performance. Thus, from this point forward in this chapter we will not consider the parameter of filter bandwidth; all analysis will be done assuming a fixed fractional bandwidth of 0.22. The choice of filter bandwidth will be considered later in the simulation chapter. The reason for this is that the true system is intended to work with speech sources, while the theory is based on the assumption of bandlimited white noise sources. In the case of true speech, energy content is not distributed uniformly across the bandwidth. The fractional bandwidth must be selected appropriately, so as to
capture enough energy from the speech to make a good judgement of the SNR while not being adversely affected by energy fluctuations in the nonstationary speech signal.

Array of PDFs for \( kd = \pi \) and Varying Bandwidth

Figure 5-2: In this set of intermicrophone correlation PDFs, the center frequency is fixed at \( kd = \pi \). The environment is assumed anechoic.

Now, we return to the issue of frequency selection. As described before, the center frequency must be selected to yield an optimal combination of strong negative correlation in the noise PDF and overlap minimization. To reconcile this issue, we refer to Figure 5-3. In this figure are given two graphs, one which gives the measured moment of the signal PDF about +1 versus \( kd/\pi \), and another which gives the moment of the noise PDF about -1 versus \( kd/\pi \). These graphs assume a fixed fractional bandwidth of 0.22 and an anechoic environment. The metric of PDF moment about the extremal points takes into account both the necessary issues - it is affected substantially by the long tails present in the noise PDFs. Thus, we propose that the frequency which minimizes these moments will have
the appropriate balance of characteristics. In the case of signal PDF, we can see that the moment increases with frequency. A lower frequency always yields a more concentrated signal PDF. However the scale of this graph is insignificant when compared with that of the noise PDF. The noise PDF moment is minimized roughly at the point $kd = 4\pi/3$, so by our metric this is the optimal choice for center frequency. From this point forward in the analysis, we will consider three choices in frequency selection - $\pi$, $4\pi/3$, and $3\pi/2$. We include $kd = 3\pi/2$ because at this point the moment of the noise PDF has risen up to roughly the same level as for $kd = \pi$.

![Graph](image)

Figure 5-3: These graphs assume an anechoic environment and fixed fractional bandwidth of 0.22.

### 5.2 Reverberation

Up to this point in the analysis, the issue of reverberation has been ignored. The choice of center frequency was made based on the assumption of an anechoic environment, but we
must now examine how this choice is affected by a reverberant environment. The greatest obstacle presented by reverberation is the fact that it is a completely unknown variable. We must choose our system parameters such that we have robust performance given any level of reverberation.

Let us first consider the general effects of reverberation. In Figure 5-4, an array of PDFs is given. These PDFs all assume fixed center frequency of $kd = \pi$ and fractional bandwidth of 0.22. The top row corresponds to an anechoic environment, while the next two rows correspond to increasing levels of reverberation. (The value of $\beta$ equals the ratio of direct signal energy and reverberant energy, as described in Section 2.2) From these graphs, it is apparent that increasing reverberation causes a contraction away from the extreme in both the signal and noise PDFs. Thus, as the value of $\beta$ decreases, the system performance will degrade.

To further explore the effects of reverberation, consider the PDFs given in Figure 5-5. These are PDFs of purely reverberant sound for differing values of $kd$. (Recall that in our theoretical model, reverberation is assumed to behave as a spherically diffuse sound field) Again, all graphs assume a fixed fractional bandwidth. As the value of $\beta$ decreases, the reverberant component PDF will have a more pronounced effect on the PDF of total intermicrophone correlation. This effect is dictated by the convolution relation in equation (3.2). This explains the effects of decreasing $\beta$ seen in Figure 5-5. As the value of $\beta$ approaches zero, the total correlation PDF will equal that of the reverberant component alone, however in this study we will only consider values of $\beta$ greater than or equal to one. We assume that there is a substantial component of direct wave sound in the environment - otherwise the proposed system would not be effective in testing SNR.

We now must consider the issue of threshold selection. A single threshold parameter must be chosen for hypothesis testing, and it must be valid in varying levels of reverberation. As reverberation increases, the structure of the signal and noise PDFs will change, and thus the optimal threshold will change. When choosing system parameters, we must take into account that the single threshold must perform well in any level of reverberation. As one can see in Figure 5-5, the reverberant PDFs for different center frequencies have largely the same shape. The contraction effects on the total correlation PDF will be roughly similar for all frequencies in question. Thus, the effects of reverberation on optimal system performance would be roughly similar for the different frequencies if we were allowed to
Array of PDFs for $kd=\pi$

Figure 5-4: In this set of intermicrophone correlation PDFs, the center frequency is fixed at $kd = \pi$ and the fractional bandwidth is 0.22.

choose our threshold given a fixed value of $\beta$ - but this is not the case. The only difference between the reverberant PDFs is the mean value of the distributions. This mean value is dictated by equation (3.4). For $kd = \pi$, the mean value of the reverberant PDF is roughly zero, while for the other two values of $kd$ it is roughly -0.2. Thus, in the case of $kd = 4\pi/3$ or $kd = 3\pi/2$, the effect of reverberation on the total correlation PDF is not only a blurring and contraction, but also a shift. Due to the shifting effect, intuitively it seems reasonable to use the center frequency $kd = \pi$ exclusively. One would think that as reverberation increases, the choice of optimal threshold would change considerably for the two other choices of center frequency. Ideally, this would be true, but in our model this intuition fails. Consider Figures 5-6, 5-7 and 5-8. Each of these figures considers a different of center frequency. In each figure, signal and noise PDFs are plotted together for
Reverberant PDFs for $kd=[\pi, (4\pi)/3, (3\pi)/2]$

Figure 5-5: In this set of reverberant component PDFs, the fractional bandwidth is fixed at 0.22.

Varying levels of reverberation (the signal PDF corresponds to $W = 100$ and the noise PDF to $W = 0.01$). The choice of optimal threshold is the point at which the signal and noise PDFs are equal. It is apparent that for any choice of center frequency, the optimal threshold changes considerably with increasing reverberation. The signal and noise PDFs generated by our theoretical model are sufficiently assymetric and oddly shaped that the ideal case does not apply. It is difficult to make an intuitive judgement regarding parameter selections. Thus, to make our final choices for center frequency and threshold, in the next section we use a series of brute-force numerical computations to find the optimal parameters.
PDFs for \( kd=\pi \), Varying Beta

Figure 5-6: In this set of graphs, the center frequency is fixed and the fractional bandwidth is fixed at 0.22.

5.3 Receiver Operating Characteristics

Using the mathematical machinery developed in our probabilistic model of intermicrophone correlation, conditional PDFs \( f_{r|H_0}(r|H_0) \) and \( f_{r|H_1}(r|H_1) \) as described by equations (4.1) and (4.2) were generated. These PDFs are representative of the two hypotheses in question: low SNR and high SNR. All computations were performed by applying Simpson’s rule [7] for numerical integration. Furthermore, appropriate integrations were performed to generate receiver operating characteristic curves for three values of center frequency in varying levels of reverberation. The probability of detection, \( P_D \), is defined as the probability of deciding \( H_1 \) (high SNR) given that \( H_1 \) is true. This is determined by integrating \( f_{r|H_1}(r|H_1) \) over the range \( r > r_0 \). The probability of false alarm, \( P_F \), is defined as the probability of deciding
PDFs for $kd = (4\pi)/3$, Varying Beta

Figure 5-7: In this set of graphs, the center frequency is fixed and the fractional bandwidth is fixed at 0.22.

$H_1$ given that $H_0$ is true. In the same manner, this is determined by integrating $f_{r|H_0}(r|H_0)$ over the range $r > r_0$. Again, analysis was performed assuming a fixed fractional bandwidth of 0.22. Figures 5-9, 5-10, and 5-11 show these ROC curves. From this set of figures it is apparent that the higher choices of center frequency, $kd = 4\pi/3$ and $kd = 3\pi/2$ have better performance than $kd = \pi$ for all levels of reverberation.

Now we move on to the issue of threshold selection. Figure 5-12 shows a set of three graphs, one for each choice of center frequency. These graphs plot the probability of error versus threshold for varying levels of reverberation. Probability of error is defined as $P_E = P(H_0)P_F + P(H_1)(1 - P_D)$. In this circumstance, since the two hypothesis have equal prior probability, this reduces to $P_E = \frac{P_F + 1 - P_D}{2}$. For $kd = \pi$, one can see that the optimal choice for threshold is at roughly 0.1 for all levels of reverberation. However, for the other values
PDFs for $kd = (3\pi)/2$, Varying Beta

![Graphs showing PDFs for different values of Beta](Figure 5-8)

Figure 5-8: In this set of graphs, the center frequency is fixed and the fractional bandwidth is fixed at 0.22.

of $kd$, it is apparent that the optimal threshold actually shifts with increasing reverberation. This is a result of the shifting of the total correlation PDF due to reverberation, as described in the prior section. In the case of $kd = 4\pi/3$, by choosing 0 as the threshold we will obtain a hypothesis test which performs better than $kd = \pi$ and also has fairly constant performance for all levels of reverberation. In the case of $kd = 3\pi/2$, there is no good choice of threshold - the impact of reverberation on the optimal threshold point is too extreme. Thus, from this point forward we will no longer consider the choice of $kd = 3\pi/2$ for center frequency.

To further illustrate the aforementioned shifting effect due to reverberation, we present Figure 5-13. In this array of graphs, we see plots of probability of detection and probability of false alarm for varying levels of reverberation. In the case of $kd = \pi$, at the optimal threshold point of 0.1 which was just selected, the detection and false alarm probabilities
Figure 5-9: For this set of ROC curves, the fractional bandwidth is fixed at 0.22.

are relatively stable across different levels of reverberation. However, for \( kd = 4\pi/3 \) at the optimal threshold point 0 which was just selected, the detection and false alarm probabilities are influenced by increasing reverberation. The probability of error remains fairly constant because losses in probability of detection are compensated for by losses in probability of false alarm as the value of \( \beta \) decreases. This is entirely due to the shifting effect, but in the case of \( 4\pi/3 \) the effect is mild enough that a suitable threshold can be found.

Finally, having now settled upon appropriate parameter choices we shall continue on to the simulation phase of this study. In the chapter which follows, our system will be critically evaluated using the selected parameters to determine if the true performance agrees with our theoretical model.
Figure 5-10: For this set of ROC curves, the fractional bandwidth is fixed at 0.22.
Figure 5-11: For this set of ROC curves, the fractional bandwidth is fixed at 0.22.
Figure 5-12: For this set of curves, the fractional bandwidth is fixed at 0.22.
Figure 5-13: For this set of curves, the fractional bandwidth is fixed at 0.22.
Chapter 6

Simulation

6.1 System Specification

In this chapter we present the experimental results of simulation using system parameters selected in the previous chapter. The general flow of information through the system is as given by the block diagram in Figure 1-2. For the purpose of this simulation, exact specifications of system operation are as follows.

1. Signals received at both microphones are digitally sampled at a rate of 10 kHz.

2. Both sampled signals are passed through the same bandpass filter, with appropriate center frequency and bandwidth. In this simulation we use an 81 point optimal FIR filter designed using the Parks-McClellan algorithm.

3. Filtered signals are then broken into segments of 100 samples. (This corresponds to 10 milliseconds at a sampling rate of 10 kHz, a window length appropriate for tracking power fluctuations in speech) For each corresponding pair of segments, the sample correlation coefficient is computed (see equation 2.5).

4. Finally, the sample correlation coefficient is compared to the threshold. If it exceeds the threshold, the system outputs "true" for the given segment (high SNR), otherwise it outputs "false" (low SNR).
6.2 Noise Simulation

The first set of simulations was performed using white gaussian noise sources. Both sources consisted of 28000 samples of zero-mean white gaussian noise. In the case of the noise source, the variance was fixed at 1. The signal source was broken into 2000-sample intervals of constant variance - this variance was incremented such that the SNR would increase in 3 dB steps between -19.5 dB and 19.5 dB.

The simulations are based on a model of a room with dimensions 5.2 x 3.4 x 2.8 meters. The distance separating the microphones is 14 cm, and the microphones are located at coordinates (2.7495, 1.3505, 1.600) and (2.6505, 1.4495, 1.600). The array center is the midpoint of these coordinates - (2.7, 1.4, 1.6). All sources in the room are located on a circle around the array center in the horizontal plane at height 1.7 meters. The forward direction ($\theta = 0$) is defined to be directly broadside of the array in the direction of positive coordinates, and increasing the incident angle refers to clockwise progression of source angle when viewed from above. The radius of source locations and coefficient of absorption for the walls is selected depending on the desired level of reverberation. For an anechoic environment, the radius is fixed at 1.0 meters and the absorption coefficient is 1.0. For $\beta = 2$, the radius is fixed at 1.07 meters and the coefficient is 0.6. Finally, for $\beta = 1$, the radius is fixed at 1.62 meters and the coefficient is again 0.6. These three level of reverberation were all tested separately in this set of simulations.

The signal and noise sources were convolved with source-to-microphone impulse responses before being passed into the SNR detection system. These impulse responses were generated numerically using the image method described by Allen and Berkley (1979) with modifications by Peterson (1986). The signal source angle was varied between 0 and 12 degrees and the noise source angle between 18 and 90 degrees, both in 4 degree increments. For all 76 combinations of signal and noise source angles, the system as described in Section 6.1 was simulated. A binary value of true or false was produced for every 100 samples (10 milliseconds) of input. Each of these binary outputs was compared to the true SNR for the given segment. The input is structured such that the SNR $< 0$ dB for the first 14000 samples, and the SNR $> 0$ dB for the last 14000 samples. Thus, the first half of the signal was used to determine the probability of false alarm $P_F$, and the second half used to determine the probability of detection $P_D$. The values of $P_D$ and $P_F$ were averaged over all
combinations of signal and noise source angle.

Simulations were run using two sets of system parameters. In the first set, the value of \( kd \) is set equal to \( \pi \) and the threshold is set at 0.1. With \( d \) fixed at 14 cm, this implies that the filter center frequency is 1238 Hz. The fractional bandwidth was varied between 0.1 and 1.5 times the center frequency (124 Hz to 1856 Hz). In the second set, \( kd \) is set equal to \( 4\pi/3 \) and the threshold is set to 0. In this case the filter center frequency is 1650 Hz, and again the fractional bandwidth is varied between 0.1 and 1.5 times the center frequency (165 Hz to 2475 Hz). Results of these simulations are shown in Figure 6-1.

**Noise Simulation Performance Results versus Fractional Bandwidth**

---

**Figure 6-1:** This set of graphs shows noise simulation results for the two values of \( kd \), and varying fractional bandwidth. In these plots, the probability of detection and probability of false alarm are denoted by o and x, respectively, for each fractional bandwidth. The probability of error is denoted by +. Finally, for fractional bandwidth equal to 0.22 the theoretical values for probabilities of detection, false alarm and error are denoted by the dot, square, and triangle, respectively.
The results presented in these plots are the $P_F$, $P_D$, and $P_E$ values for varying values of fractional bandwidth. Overlaid on the simulation results are the theoretically values predicted for fractional bandwidth equal to 0.22. (Recall that all analysis was done using this value.) In the case of an anechoic environment, the theoretical predictions are very accurate. As reverberation increases however, the theoretical predictions tend to deviate from the true results. This is explained by the model of reverberation used in the analysis. The assumption was made that reverberation could be accurately described as a spherically-diffuse sound field - this assumption enabled a tractable analysis. However, the true simulation also takes into account reflections in the environment.

As the bandwidth increases, there is an initial boost in system performance, followed by the expected decrease. This is evident from the measure of $P_E$ in each of the six plots - there is a definite minimum in the midrange of fractional bandwidth. This contradicts the expectations based on the analysis regarding the detrimental effects of increasing bandwidth. Previously, we had observed that increasing bandwidth yields contraction in the noise PDF away from the extremal values. While it is desirable to push the PDF as far to the extremes as possible, the contraction due to increasing bandwidth also tends to reduce overlap between the signal and noise PDFs. By increasing bandwidth, the measure of intermicrophone correlation becomes less sensitive to fluctuations in individual frequency components, and both the signal and noise PDFs will benefit from reductions in variance. As was discussed in the analysis chapter, overlap minimization is also important in improving system performance. As it turns out from simulation results, the initial benefits of increasing fractional bandwidth outweigh the detriments - we see both an increase in $P_D$ and a decrease in $P_F$. This is especially apparent in the case of reverberation. (In the anechoic case, the initial improvement is negligible) For $\beta = 1$ or $\beta = 2$, the additional PDF contraction caused by the reverberant energy component comes into play. As fractional bandwidth increases beyond a certain level, the effect of noise PDF contraction away from the extreme tends to dominate, and $P_F$ increases dramatically. For all levels of reverberation, we can see that there is a definite optimal fractional bandwidth of about 1.0 for $kd = \pi$ and about 0.67 for $kd = 4\pi/3$. Overall, the noise simulations yielded promising results. The performance of the system when $kd = 4\pi/3$ is in general better than that when $kd = \pi$, as expected. Having considered the noise case, we now move on to simulations using true speech.
6.3 Speech Simulation

Simulation using true speech sources followed the exact protocol as that of the noise sources. The signal source consisted of a concatenation of several phonetically-balanced sentences (IEEE, 1969) read by a single male speaker. The noise source consisted of 12-talker SPIN babble (Kalikow, et al., 1977). The babble was set to be the same length as the speech and normalized to have the same total power. As before, all sources were sampled at a rate of 10 kHz. For each 100 sample (10 millisecond) segment, the output of the SNR testing system was compared to the true SNR for that segment. The true SNR was calculated for each segment by taking the ratio of total power in the speech segment and total power in the babble segment. In this manner, $P_D$ and $P_F$ were determined as before and averaged for all combinations of signal and noise source angle. Results of these simulations are given in Figure 6-2.

The results presented in these plots are the $P_F$, $P_D$, and $P_E$ values for varying levels of fractional bandwidth. The general trends in performance in the case of speech sources are the same as those observed in the case of noise. The obvious difference is the drop in system performance when using speech rather than noise. Both the values of $P_D$ and $P_F$ are strongly affected by the use of speech sources. The reason for this lies in the fact that the SNR in the case of speech is not uniformly distributed in the range -20 dB to 20 dB, as was assumed for the noise. Rather, fluctuations in the SNR tend to be concentrated at less extreme values. In this reduced range of potential SNR values, the performance of the testing system will degrade. The majority of mistakes made by the system will occur in regions of transition. This is evident from Figure 6-3, which shows the true SNRs for a 3 second speech segment along with the locations of misses and false alarms.

The second major difference in the case of speech is the sensitivity of performance to increasing bandwidth. The levels of $P_D$ and $P_F$ are affected far more in the case of speech sources when compared with noise sources. This is due to the fact that the energy in the speech signal is not uniformly distributed across the bandwidth. The issue of bandwidth as it relates to speech was considered in the analysis chapter. The fractional bandwidth must be selected appropriately, so as to capture enough energy from the speech to make a good judgement of the SNR while not being adversely affected by energy fluctuations in the nonstationary speech signal.
Speech Simulation Performance Results versus Fractional Bandwidth

Figure 6-2: This set of graphs shows speech simulation results for the two values of $kd$, and varying fractional bandwidth. In these plots, the probability of detection and probability of false alarm are denoted by o and x, respectively, for each fractional bandwidth. The probability of error is denoted by +.

From the results obtained through simulation, we again see in general that the choice of $kd = 4\pi/3$ for center frequency yields better performance. Also, the optimal choice of fractional bandwidth in the case of either value of $kd$ is 0.67. This yields the lowest probability of error $P_E$ across all levels of reverberation.
Figure 6-3: In the upper plot, a trace of SNR for a 3 second speech segment is given. In the lower plot, a trace of intermicrophone correlation is given for the same speech segment. On the upper plot, misses are represented by x, and false alarms are represented by +. The majority of misses and false alarms are concentrated around regions of transition near 0 dB. The system parameters used for testing in this plot are $kd = \pi$ and fractional bandwidth 0.22.
Chapter 7

Discussion

The purpose of this study was to investigate the use of the intermicrophone correlation coefficient as a metric for testing the SNR in an arbitrary acoustic environment. A hypothesis testing system for distinguishing between high and low levels of SNR was proposed and analyzed theoretically. The system was then critically evaluated through simulation using both noise and speech sources. Though the theory made many assumptions to allow for a tractable analysis, it was found that simulated results corresponded well with predictions, despite violating many of those assumptions.

Two sets of parameters were proposed by the theory - one with center frequency $kd = \pi$ and threshold 0.1 and the other with center frequency $kd = 4\pi/3$ and threshold 0. The results of simulation revealed better performance across all levels of reverberation for the higher choice of center frequency. It was also determined in the case of speech simulation that a choice of 0.67 for fractional bandwidth yields the best results. While the potential for error does exist, in the case of speech it was shown that mistakes tend to occur almost exclusively in regions of transition. In regions of extreme SNR, either high or low, the system rarely fails.

The performance of the system in general was good in all environments, and provides a method for determining SNR with relatively low computational complexity.
Bibliography


