2017

Time-Frequency Masking Performance for Improved Intelligibility with Microphone Arrays

Joshua P. Morgan
University of Kentucky, JoshuaMorganUSA@gmail.com
Author ORCID Identifier:
http://orcid.org/0000-0002-5663-2264
Digital Object Identifier: https://doi.org/10.13023/ETD.2017.145

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Joshua P. Morgan, Student

Dr. Kevin D. Donohue, Major Professor

Dr. Cai-Cheng Lu, Director of Graduate Studies

Time-Frequency Masking Performance for Improved Intelligibility with Microphone Arrays

---

THESIS

A thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science
in Electrical Engineering in the College of
Engineering at the University of Kentucky

By

Joshua P. Morgan
Lexington, KY

Director: Dr. Kevin D. Donohue
Professor of Electrical and Computer Engineering
Lexington, KY

2017

ABSTRACT

Time-Frequency Masking Performance for Improved Intelligibility with Microphone Arrays

Time-Frequency (TF) masking is an audio processing technique useful for isolating an audio source from interfering sources. TF masking has been applied and studied in monaural and binaural applications, but has only recently been applied to distributed microphone arrays. This work focuses on evaluating the TF masking technique's ability to isolate human speech and improve speech intelligibility in an immersive "cocktail party" environment. In particular, an upper-bound on TF masking performance is established and compared to the traditional delay-sum and general sidelobe canceler (GSC) beamformers. Additionally, the novel technique of combining the GSC with TF masking is investigated and its performance evaluated. This work presents a resource-efficient method for studying the performance of these isolation techniques and evaluates their performance using both virtually simulated data and data recorded in a real-life acoustical environment. Further, methods are presented to analyze speech intelligibility post-processing, and automated objective intelligibility measurements are applied alongside informal subjective assessments to evaluate the performance of these processing techniques. Finally, the causes for subjective/objective intelligibility measurement disagreements are discussed, and it was shown that TF masking did enhance intelligibility beyond delay-sum beamforming and that the utilization of adaptive beamforming can be beneficial.

KEYWORDS: Distributed Microphones, Cocktail Party, Time-Frequency Masking, Beamforming, Adaptive Beamforming, Intelligibility

Joshua P. Morgan

May 2, 2017

Time-Frequency Masking Performance for Improved Intelligibility with Microphone Arrays

By

Joshua P. Morgan

Dr. Kevin D. Donohue
Director of Thesis

Dr. Cai-Cheng Lu
Director of Graduate Studies

May 2, 2017

ACKNOWLEDGMENTS

I would first like to thank my advisor, Dr. Kevin Donohue, who has been a mentor to me throughout my entire undergraduate and graduate career at the University of Kentucky. His guidance as both an academic and personal advisor has been invaluable.

I also want to thank my other thesis committee members, Dr. Michael Johnson and Dr. William Smith, for their willingness to provide feedback toward my thesis and defense completion.

Finally, I want to thank my loving parents, Matthew and Rebecca, my family, and my closest friends for their love and unending support during my college career and beyond.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Literature Review

1.1 History and Motivation for Study

The ability to isolate a single sound source from interfering noises has been a topic of study for many years. The "source of interest" (SOI) is frequently a human speaker in an environment with competing interfering sources such as ambient noise, musical sources, and other human speakers. Techniques used to isolate a sound source (particularly human speech) to improve its clarity and intelligibility are useful in a variety of applications: room surveillance, hearing aides, "smart rooms", and more. Such techniques have primarily been studied in binaural or monaural contexts which are effective and useful for hearing aides and similar applications [1, 2, 3]. More recently, however, the interest in other scenarios (surveillance, smart rooms, etc) have led to the application/adaptation of these techniques to distributed microphone arrays.

Microphone array beamforming has been an active area of research, wherein an array response is manipulated such that the array is "steered" to the source of interest, improving its isolation and intelligibility. The simple delay-sum beamformer (DSB) has been shown to be effective for this purpose and can be easily applied to the audio response of the microphone array [4]. Adaptive beamforming techniques have been developed and studied for microphone arrays and have been shown to further improve speech intelligibility of a speech recording [5].

Time-Frequency (TF) masking is an analysis technique useful for isolating an audio source of interest in the presence of other interfering sources. By considering a signal's spectral properties compared to that of interefering sources, the isolation of the target source can be improved significantly. TF masking has been studied at length in binaural and monaural applications and has been shown to improve speech intelligibility by both automated and subjective measurements [6, 7, 3, 8, 9]. TF Masking has only been briefly studied as a technique for distributed microphone arrays, but

has proven to be very useful in the separation of a target source from its interferers [10].

Though TF Masking has shown success for use with microphone arrays, an in-depth study has not yet been performed. This study aims to more fully describe the operational performance of TF masking and its dependence on a variety of parameters. We will work primarily within the context of a "cocktail party problem", wherein we attempt to isolate a single human voice from amongst interfering sources, all distributed throughout an acoustical environment. Using virtual simulation techniques, we will develop an understanding for the efficacy of the TF masking technique, and we will further compare these results to those created from real recordings in a lab environment. Finally, we will study TF Masking in conjunction with simple beamforming techniques, and also introduce the novel idea of combining TF Masking with the more advanced adaptive beamforming techniques.

In the investigation of these techniques, a variety of performance metrics have been utilized. Some studies have focused on subjective testing results [6], while others have used objective automated metrics [11, 10]. Subjective testing can provide an accurate understanding of human perception, but limits the scope and scale of an experimental study. Objective measurements do not impose this limitation, but may not accurately predict human intelligibility. For the present study, we use an objective intelligibility measurement along with informal subjective assessments throughout.

1.2 Literature Review

Beamforming techniques have been extensively studied and utilized in many applications, including monaural, binaural, and distributed microphone arrays. In the 1970s, [12] presented improvements to the traditional delay-sum beamformer by including an adaptive component to the algorithm. This was further refined some ten years later when [5] showed a simplified implementation of [12]'s work and its use with microphone arrays. In recent years, [13] showed the stability requirements of this adaptive beamformer and [11] studied potential enhancements to the adaptive beamforming technique. Studies such as [14] have verified that the adaptive beamforming techniques are beneficial. [1] has recently presented developments in binaural beamforming performance by decreasing computational complexity, particularly of use in hearing-aid applications. Other studies have investigated the use of machine learning and neural network techniques and have shown further improvement as a result [15, 16]. Finally, existing work, such as [17], have shown a benefit to combining these beamforming techniques with other isolation methods (such as TF masking).

TF masking has been successfully applied in monaural and binaural applications for many years. [3] showed use of ideal TF masking, also referred to as Ideal Binary Mask (IBM), in a monaural cocktail party environment and suggested IBM results as a performance goal for other algorithms. In [6], the study of monaural TF masking was extended to include reverberation effects and several methods of IBM creation

were studied. It was shown in [8] that TF masking is likewise effective in binaural scenarios and practical techniques were presented to estimate the ideal TF masking results. Additionally, some studies have shown success in using machine learning and neural networks to improve TF masking performance, particularly by using ideal masking behavior to train the machine learning algorithms [7, 18, 19]. These studies have primarily focused on estimating the IBM by identifying TF regions where the SOI is active based on signal features and auditory cues.

Much of the work in this space has focused on emulating a human's ability to isolate target speech from interference in monaural/binaural scenarios. With the use of distributed microphones, an advantage not afforded to humans is gained: that is, the ability to receive sound throughout an environment beyond binaural sensing. Additionally, it was shown in [20] and [21] that the microphone array arrangement (number of microphones, distribution geometry, etc) has a measurable effect on the performance of microphone array processing algorithms. [22] demonstrated integration of beamforming and spectral mask-based noise estimation with microphone arrays to isolate a target source, while [23] verified the benefits of TF masking using an IBM. [10] first proposed the technique of estimating the IBM using beamformed reference signals of each audio source, and that method is likewise used in the present study. Though studies have been done to study beamforming and masking isolation methods individually, few have compared the performance of beamforming and masking techniques in the context of distributed microphones. Further, while [10] estimated the time-frequency IBM using delay-sum beamformed reference signals, no work has yet studied IBM estimation using adaptive beamforming techniques. This work investigates mask creation using beamformed reference signals (delay-sum and adaptive beamforming) and compares performance to ideal masking and beamforming-only processing.

In the study of isolation performance, a variety of evaluation methods have been proposed. Subjective testing is an effective measure of human intelligibility and has been used in studies such as [6] and [24]. [25] presented a comparison of several objective measures for speech intelligibility such as the Articulation Index (AI), Speech Intelligibility Index (SII), Speech Transmisison Index (STI), and modified implementations of these measures, and found that the SII (using modified weighting parameters for their scenario) was moderately effective at predicting human intelligibility. The SII is an objective intelligibility metric developed through extensive subjective testing [26]. [27] found that considering transients in the calculation of the STI metric was beneficial over the baseline STI metric. Further, [28] showed that the Perceptual Evaluation of Speech Quality (PESQ) measure is also useful for predicting the performance of these techniques. Finally, [29] demonstrated the intelligibility effects of TF mask sparseness and its relationship to target SNR, [30] presented a study of the impacts on TF masking and mask accuracy due to room acoustics and reverberation effects, and [31] investigated the relative importance of TF regions for speech intelligibility. The study of speech intelligibility is still an evolving science and no current method definitely predicts human intelligibility. For the purposes of this study, the SII metric will be used to compare relative performance of the target isolation techniques

alongside a limited subjective analysis to verify resulting trends.

1.3 Introduction to Beamforming

Beamforming is a filtering technique that can be used to help isolate a target audio source from an environment with interfering noise sources. In our case, an array of microphones is distributed in a near-field acoustical environment to record the audio from various sources in the environment. The beamforming technique relies on knowledge of microphone and source positions within the environment, and also requires some level of incoherence between received signals from competing sources. Additionally, a beamformer's performance is a function of the actual distribution of microphones within the space [21], but our study will only utilize a simple planar array geometry; however, the simulation and evaluation techniques used in this work can be applied to any near-field/immersive geometry. The following sections describe the beamformer's mathematical model and is agnostic of the specific array geometry being used.

1.3.1 Basics of Beamforming

Consider a three dimensional acoustical environment with M microphones and Q sound sources distributed throughout, and let $u(t;\boldsymbol{r}_q)$ be the sound source signal located at position $\boldsymbol{r}_q$. The microphone response for the mth microphone located at position $\boldsymbol{r}_m$ can be expressed as

$$x_m(t) = \sum_{q=1}^{Q} u(t;\boldsymbol{r}_q) * h(t;\boldsymbol{r}_m,\boldsymbol{r}_q) \tag{1.1}$$

where $h(\cdot)$ is the impulse response of the sound propagation from source to microphone, and $\boldsymbol{r}_m$, $\boldsymbol{r}_q$ are the x,y,z coordinates of the mth microphone and qth sound source, respectively. For a reverberant environment, this impulse response is given by

$$h(t;\boldsymbol{r}_m,\boldsymbol{r}_q) = \sum_{k=0}^{\infty} a_{qm,k}(t-\tau_{qm,k}) \tag{1.2}$$

where $a_{qm,k}$ is the attenuation response from the kth propagation path of the source signal, $\tau_{qm,k}$ is the corresponding time delay from source to microphone, and $k=0$ represents the *direct* path from source to microphone. Transforming the received

signal of Equation 1.1 into the frequency domain yields:

$$X_m(f) = \sum_{q=1}^{Q} \sum_{k=0}^{\infty} U(f; \boldsymbol{r}_q) \cdot A_{qm,k}(f) e^{-j2\pi f \cdot \tau_{qm,k}} \tag{1.3}$$

Given the frequency domain representation of the microphone array response, X, the generic beamformer output can be described in the frequency domain as

$$Y(f; \boldsymbol{r}_p) = \sum_{m=1}^{M} w_m(f) \cdot X_m(f) \tag{1.4}$$

where $\boldsymbol{r}_p$ is the beamformer *focal point*, $w_m(f)$ is a set of complex weights applied to each individual microphone response, and the time-domain output signal $y(t; \boldsymbol{r}_p)$ is given as

$$y(t; \boldsymbol{r}_p) = \mathcal{F}^{-1} \{Y(f; \boldsymbol{r}_p)\} \tag{1.5}$$

1.3.2 Delay-Sum Beamformer

Given the previous description of a beamformer, the Delay-Sum Beamformer (DSB) is created if the complex weights $w_m(f)$ are selected to be purely phase terms with unity amplitude such as

$$w_m(f) = e^{-j2\pi\alpha} \tag{1.6}$$

Then, the beamformer output becomes

$$Y(f; \boldsymbol{r}_p) = \sum_{m=1}^{M} X_m(f) e^{-j2\pi\alpha} \tag{1.7}$$

The frequency domain phase terms, $w_m(f)$, are equivalent to time delays in the time domain. For the DSB, these delays, τ_{pm}, are chosen to be the time required for sound to propagate distance d_{pm} from the beamformer focal point (x_p, y_p, z_p) to the mth microphone at position (x_m, y_m, z_m) through the *direct* path, which can be expressed as

$$\tau_{pm} = \frac{d_{pm}}{c} = \frac{\sqrt{(x_p - x_m)^2 + (y_p - y_m)^2 - (z_p - z_m)^2}}{c} \tag{1.8}$$

where c is the propagation speed of sound. This yields the DSB response in the time domain as

$$y(t; \boldsymbol{r}_p) = \sum_{m=1}^{M} x_m(t - \tau_{pm}) \tag{1.9}$$

where $x_m(t)$ is the mth microphone response at time t. Finally, because we will work exclusively with digital computer systems, the microphone response and beamformer output signals must be discretely sampled with some sampling frequency f_s. The discrete representation of the beamformer output is given by

$$y(\boldsymbol{r}_p)[n] = \sum_{m=1}^{M} x_m[n - \tau_{pm} \cdot f_s] \tag{1.10}$$

Although the delay-sum beamformer can be designed with the complex weights chosen as described, the weights are traditionally selected having a magnitude of $1/M$ such that the sum of all weights equals unity [5, 32] This gives:

$$y(\boldsymbol{r}_p)[n] = \frac{1}{M} \sum_{m=1}^{M} x_m[n - \tau_{pm} \cdot f_s] \tag{1.11}$$

Finally, these complex weights may, instead, be chosen with their magnitude as a function of the distance between source and microphone d_{pm}, which allows for adjusting the influence of microphones based on their distance from the source. This can be expressed as:

$$y(\boldsymbol{r}_p)[n] = \sum_{m=1}^{M} |w_m| \cdot x_m[n - \tau_{pm} \cdot f_s] \tag{1.12}$$

For this study, however, the simple magnitude of $1/M$ is chosen for the complex weights as shown in eq.(1.11). A visualization of the alignment, or "steering", of the beamformer is shown in Figure 1.1. It can be seen in the figure how the delay-sum beamformer aligns the individual microphone responses such that the target source becomes coherent post-alignment.

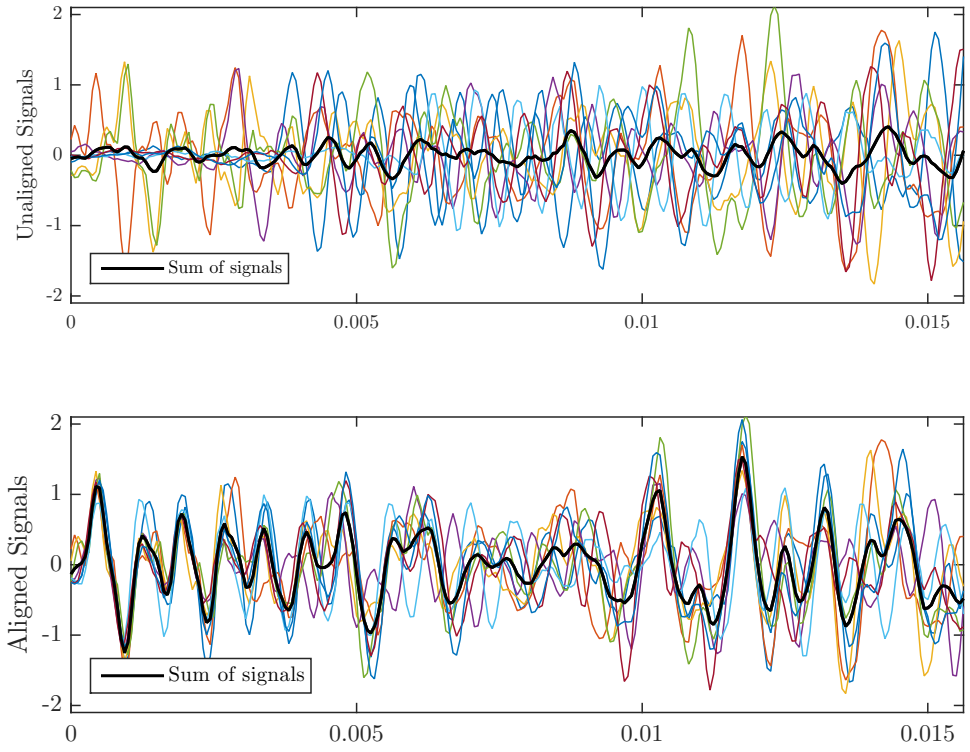

Figure 1.1: Visualization of beamformer's alignment of microphone signals: colored lines indicate individual microphone responses, solid black line depicts sum of microphone responses

1.4 Griffiths-Jim General Sidelobe Canceler

Although the simple delay-sum beamformer can perform effectively in many cases, it does not consider the dynamic nature of an audio scene with multiple fluctuating sources. The Frost algorithm [12] was one of the first methods proposed to dynamically adapt a beamformer's operation in response to the incoming signal. A simplified implementation of the Frost algorithm was later proposed by Griffiths and Jim [5] as the General Sidelobe Canceler (GSC) and is presented here for use in our study. An overview of the GSC algorithm is shown in Figure 1.2 and is described below.

To begin, we modify our notation such that $x_m[n]$ represents the *aligned* response of the mth microphone; that is, signals $x_1[n], \ldots, x_M[n]$ have already been aligned (the "delay" portion of the delay-sum beamformer) for a chosen focal point, though these signals have not yet been weighted. We refer to the collective vector of microphone response signals as $\boldsymbol{X}[n]$ which is an $O \times M$ matrix with each of the M matrix columns corresponding to the response from one of M microphones and contains O total samples.

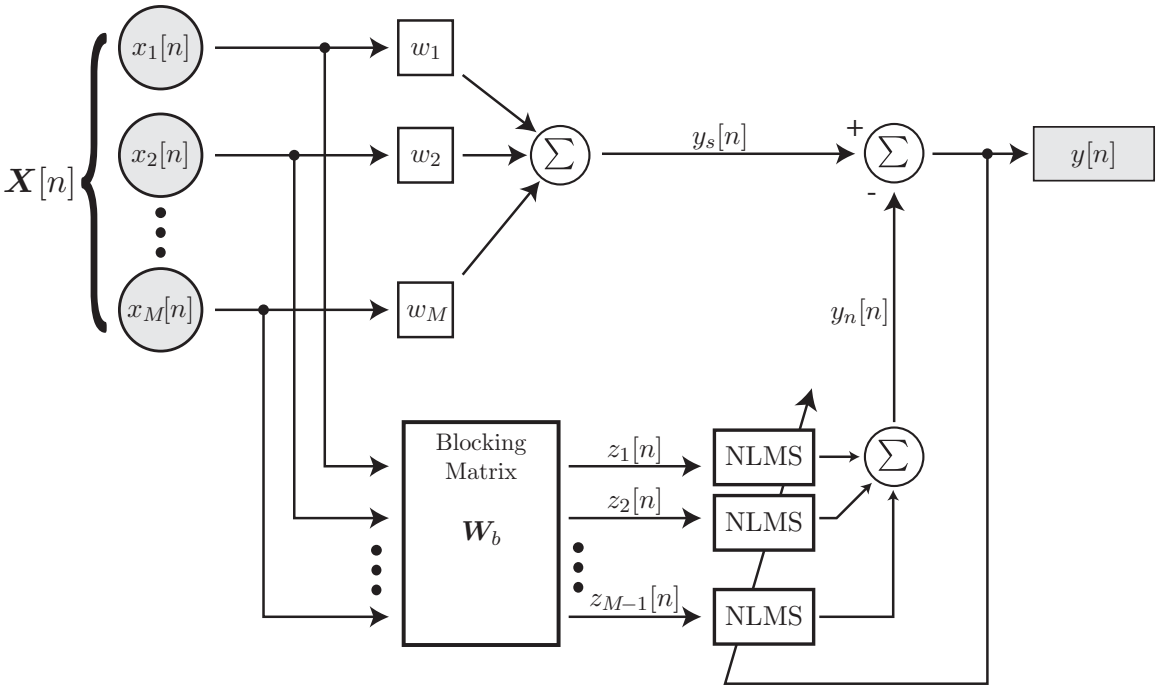

Figure 1.2: Griffiths-Jim General Sidelobe Canceler (GSC) adaptive beamforming algorithm

The top branch applies an individual weighting factor $(w_1, \ldots, w_M)$ to each of the M microphone signals and is called a Fixed Beamformer because its behavior is constant with time. The weighting factors may be chosen freely but are usually selected as $1/M$, which makes this beamfomer equivalent to the traditional delay-sum beamformer [11]. The Fixed Beamformer yields the signal $y_s[n]$ which contains both the target signal as well as interfering noise:

$$y_s[n] = \sum_{m=1}^{M} w_m \cdot x_m[n] = \frac{1}{M} \sum_{m=1}^{M} x_m[n] \tag{1.13}$$

The bottom branch implements the adaptive nature of the GSC by first passing $X[n]$ through a Blocking Matrix, an algorithm designed to eliminate the target signal from the incoming data to form a reference of the interfering noise. Griffiths-Jim selects the blocking matrix to be the simple pair-wise difference of the M signal tracks, yielding the $O \times M - 1$ vector $\boldsymbol{Z}[n]$, which is computed as the matrix product of the blocking matrix and input data vector:

$$\boldsymbol{Z}[n] = \boldsymbol{X}[n]\boldsymbol{W}_b \tag{1.14}$$

where $\boldsymbol{W}_b$ is given as the $M \times M-1$ matrix:

$$\boldsymbol{W}_b = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & -1 \end{pmatrix} \tag{1.15}$$

Using the total output signal $y[n]$ as a reference, the noise reference vector $\boldsymbol{Z}[n]$ passes through adaptive filters using the Normalized Least Mean Square (NLMS) algorithm and is summed to create the total noise reference signal $y_n[n]$. The Fixed Beamformer output $y_s[n]$ contains both the target signal and interfering noise, while the noise reference $y_n[n]$ ideally contains *only* interfering noise. By subtracting the noise reference from the Fixed Beamformer output, the NLMS filters attempt to minimize the total output power of $y[n]$. In an ideal case, this minimizes the interfering noise signature in the output without any effects on the pure target signal. In practice, however, some amount of the target signal will leak through the blocking matrix, causing a decrease of target power in the final output, decreasing the performance of the GSC beamformer from ideal. We can describe the total GSC beamformer output as:

$$y[n] = y_s[n] - \sum_{k=1}^{M-1} \boldsymbol{w}_k^T[n]\boldsymbol{z}_k[n] \tag{1.16}$$

where $\boldsymbol{z}_k[n]$ is the kth Blocking Matrix output track of length O, and $\boldsymbol{w}_k[n]$ is the kth column of the NLMS filter tap weight matrix $\boldsymbol{W}$ of length O. The adaptive filters are updated with the NLMS algorithm:

$$\boldsymbol{w}_k[n+1] = \beta \boldsymbol{w}_k[n] + \mu y[n] \frac{\boldsymbol{z}_k[n]}{||\boldsymbol{z}_k[n]||^2} \tag{1.17}$$

where β is the forgetting factor ($0 < \beta < 1$), $||\cdot||^2$ is the squared Euclidean norm, and μ is the step size parameter ($\mu > 0$). The parameter μ determines how much the filter tap changes with each iteration, with large values resulting in rapid convergence toward a steady-state signal with large misadjustment, and small values resulting in slower convergence but with small misadjustment. The forgetting factor β adjusts the influence of the previously calculated tap weights on future weights [11]. The selection of the β and μ parameters affect the stability of the NLMS filters as described in [13]. A full discussion of GSC stability is beyond the scope of this work, but the β and μ parameters are chosen, as 0.9 and 0.1, respectively, throughout this study. These values were selected to maintain stability, while providing sufficient dynamic filter response to the output signal.

1.5 Estimating TF Mask with Distributed Microphones

Consider an environment with multiple stationary sound sources (eg. multiple human speakers) distributed throughout. The beamforming technique can be used to improve the SNR of the source/speaker of interest (SOI) and, thus, improve intelligibility. Time-Frequency (TF) masking is a technique that can be used to further improve the intelligibility of the SOI in the presence of interfering sources [10].

Given an environment with Q sound sources at distinct locations, we create Q beamformed signals, each having a unique source as its beamformer focal point. For each beamformed signal, a short time window (order of 20-50ms) is selected and its frequency spectrum calculated. The Time-Frequency representation, Y, of a beamformed signal can be estimated with the discrete function

$$Y\left[k, i, \boldsymbol{r}_p\right] = \sum_{q=1}^{Q} G_{pq}[k] \cdot X\left[k, i, \boldsymbol{r}_q\right] \tag{1.18}$$

where i is the index of the selected time window, k is the discrete frequency index (frequency bin), $X[\cdot]$ is the time-frequency representation of the audio source at position $\boldsymbol{r}_q$, and $G_{pq}[k]$ is the discrete beamformer transfer function for the source at $\boldsymbol{r}_q$ with the beamformer focal point at $\boldsymbol{r}_p$.

Although the beamformer gain at its focal point is higher than for a source away from the focal point, the power spectrum in each TF window can be dominated by interfering speakers during periods of quiet speech from the SOI or loud speech from interfering speakers. A spectral power ratio is used to determine TF windows where the SOI is the dominant source and TF windows where interferers dominate:

$$S_{pq}[k, i] = \frac{\left|Y\left[k, i, \boldsymbol{r}_p\right]\right|^2}{\left|Y\left[k, i, \boldsymbol{r}_q\right]\right|^2} \tag{1.19}$$

for the SOI at beamformer focal point $\boldsymbol{r}_p$ and single interfering source at $\boldsymbol{r}_q$. A binary mask is then chosen as

$$T_{pq} = \begin{cases} 1, & \text{if } S_{pq}[k, i] \geq 1 \\ 0, & \text{if } S_{pq}[k, i] < 1 \end{cases} \tag{1.20}$$

When multiple interfering sources are present, the mask is chosen as the multiplication (or binary "AND" operation) of each mask corresponding to individual interferers:

$$T_p[k, i] = \prod_{q=1, q \neq p}^{Q} T_{pq}[k, i] \tag{1.21}$$

With these definitions, the output signal spectrum for a specific TF window is given as

$$Y'[k,i,\boldsymbol{r}_p] = T_p[k,i] \cdot Y[k,i,\boldsymbol{r}_p] \tag{1.22}$$

Finally, the time domain signal can be reconstructed through an inverse FFT process. By masking Time-Frequency areas where the SOI is overpowered by interfering sources, the intelligibility of the SOI can be improved.

As a demonstration of the TF masking operation, consider the case of a single target at $\boldsymbol{r}_1$ and single interferer at $\boldsymbol{r}_2$; then, $|Y[k,i,\boldsymbol{r}_1]|^2$ can be shown to equal

$$|G_{11}[k]|^2 \cdot |X[k,i,\boldsymbol{r}_1]|^2 + |G_{12}[k]|^2 \cdot |X[k,i,\boldsymbol{r}_2]|^2 + 2\left|\Re\left\{\left(G_{11}[k]G_{12}^*[k]\right)\left(X[k,i,\boldsymbol{r}_1]X^*[k,i,\boldsymbol{r}_2]\right)\right\}\right| \tag{1.23}$$

and, similarly, $|Y[k,i,\boldsymbol{r}_2]|^2$ equals

$$|G_{21}[k]|^2 \cdot |X[k,i,\boldsymbol{r}_1]|^2 + |G_{22}[k]|^2 \cdot |X[k,i,\boldsymbol{r}_2]|^2 + 2\left|\Re\left\{\left(G_{22}[k]G_{21}^*[k]\right)\left(X[k,i,\boldsymbol{r}_2]X^*[k,i,\boldsymbol{r}_1]\right)\right\}\right| \tag{1.24}$$

where $\Re\{\cdot\}$ denotes the real part. If it is assumed that the target source is uncorrelated with the interfering source, the cross-spectra terms, $2|\Re\{\cdot\}|$, go to 0 under the expected value operation. However, for a finite microphone implementation, these terms form a zero-mean random walk scaled by small coefficients, which effectively describe the noise floor generated by interfering speakers for the power spectrum estimates [10]. If the beamformer and number of microphones are sufficient to make the cross-spectra terms small with respect to the first part of (1.23) and (1.24), then (1.19) is approximately equal to:

$$S_{12}[k,i] = \frac{|G_{11}[k]|^2 \cdot |X[k,i,\boldsymbol{r}_1]|^2 + |G_{12}[k]|^2 \cdot |X[k,i,\boldsymbol{r}_2]|^2}{|G_{21}[k]|^2 \cdot |X[k,i,\boldsymbol{r}_1]|^2 + |G_{22}[k]|^2 \cdot |X[k,i,\boldsymbol{r}_2]|^2} \tag{1.25}$$

Note that G_{11} and G_{22} are high gain coefficients because they represent the beamformer gain at focal points $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$, respectively. For a typical beamfomer, these are always greater than the off-focal beamformer gains G_{12} and G_{21}.

Consider the case where the SOI (at position $\boldsymbol{r}_1$) becomes loud or dominates the interferer for certain TF intervals. Then, (1.25) becomes:

$$\lim_{|X[k,i,\boldsymbol{r}_1]|^2 \to \infty} S_{12}[k,i] = \frac{G_{11}[k]}{G_{21}[k]} > 1 \tag{1.26}$$

indicating that these frequency regions will not be masked and will pass through the masker. Conversely, consider the case where the SOI becomes quiet or is dominated by the interferer. In this case, (1.25) becomes:

$$\lim_{|X[k,i,\boldsymbol{r}_1]|^2 \to 0} S_{12}[k,i] = \frac{G_{12}[k]}{G_{22}[k]} < 1 \tag{1.27}$$

indicating a TF interval that will be masked and not passed to the masker output.

For the masker to be effective, the SOI must remain sufficiently unmasked in the TF space, while the interference is blocked by the mask enough to enhance speech intelligibility. The mask creation process is subject to errors which will degrade performance. For example, when the beamformer quality is poor, such that the ratios in (1.26) and (1.27) are close to unity, the cross-spectra terms can influence the ratio and cause errors in the binary mask. Further, when there are correlations between the target source and interferers, the cross-spectra terms again can create errant mask values. Finally, there is the extreme case when interference is so dense that, even though mask creation may be done perfectly, too many target-containing TF regions are masked out such that an insufficient amount of target is present in the final output.

1.6 Speech Intelligibility

It is useful to quantify the *intelligibility* of a speech signal; that is, how well a human can perceive spoken words even when in the presence of interfering noise. While a simple SNR measurement is useful for many tasks, it is not an ideal metric for speech intelligibility. This is primarily because human perception is not only a function of SNR, but also one of the structure of interfering noise. It may be easier for a human to understand a low SNR speech signal when white noise is the primary interferer, than a higher SNR signal in which the interfering noise is that of other spoken words. For this reason, a more advanced technique is desired to quantify speech intelligibility of a signal.

The Speech Intelligibility Index ("SII") is an improvement over simple SNR for this quantification [26]. The index is calculated by computing the SNR of multiple frequency bands between the SOI and interfering sources. These frequency-band SNR values are scaled nonlinearly and are weighted according to human perception based on subjective testing. A plot of the frequency band weighting function can be seen in Figure 1.3. The SII values range from 0 (completely unintelligible) to 1 (perfectly intelligible). The SII measurement is used in this study as a means to quantify predicted human intelligibility of a signal with interfering noise.

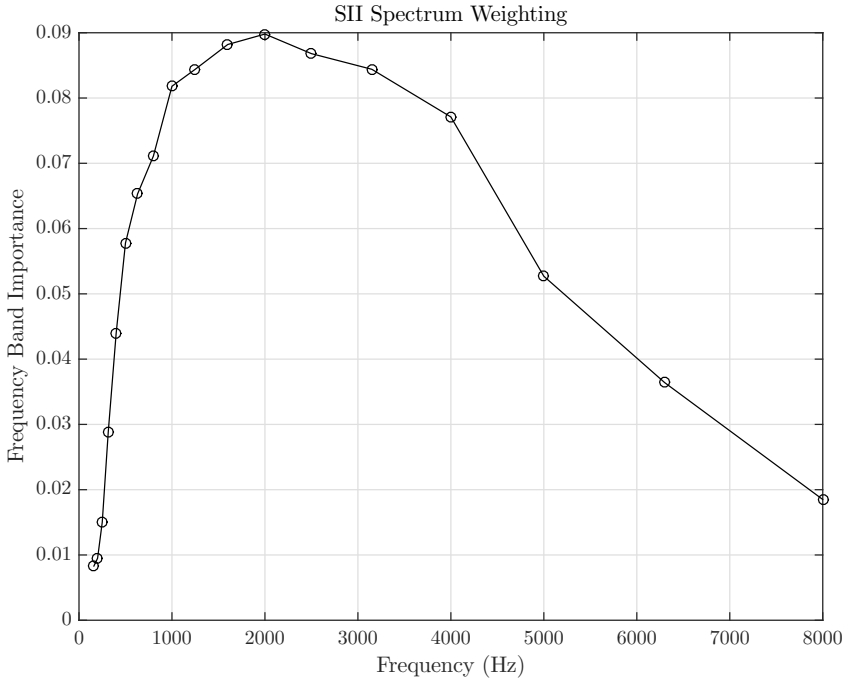

Figure 1.3: SII frequency band importance

1.7 Conclusion

The objective of this thesis is to more thoroughly study the TF masking technique when used with distributed microphones. In particular, by demonstrating performance of an ideal TF masker, we will evaluate and compare the performance of practical TF masking techniques. Further we will compare TF masking to traditional and adaptive beamformers as well as the new technique of combining adaptive beamforming with TF masking. We hypothesize that TF masking will show intelligibility enhancement over beamforming-only processing and, further, that adaptive beamforming combined with TF masking will be advantageous. Finally, we will utilize objective intelligibility metrics along with informal subjective assessments to evaluate and compare the speech intelligibility enhancements provided by these techniques.

1.8 Organization of Thesis

The organization of the remainder of this thesis is as follows. Chapter 2 introduces the methods used to produce simulated and recorded data for use in this study, along with discussing the techniques used to evaluate the performance of the target signal isolation techniques. Chapter 3 provides a detailed explanation of the implementation of the isolation techniques and the methods required to facilitate their evaluation. Chapter 4 presents an experiment that establishes an upper-bound on TF masking

performance, and Chapter 5 describes an experiment to determine the benefits of TF masking over those of traditional beamforming. Chapter 6 presents an experiment to determine the benefits of applying adaptive beamforming methods alongside the TF masking technique. Finally, Chapter 7 summarizes the research and results of this study and provides suggestions for future work on this topic.

Chapter 2

Data Collection, Simulation, and Evaluation Techniques

2.1 Introduction

The purpose of this study is to investigate the performance of the Time-Frequency Masking analysis technique and how it compares to that of other sound source isolation methods. Test data from a microphone array in a cocktail party scene is created for analysis by means of both real microphone recordings and simulations. The real microphone recordings are used to demonstrate the practical performance of these analysis techniques, while the simulations are created to provide a wide variety of testing parameters and allow for a more ideal setup in which to test these analyses.

In addition to these data creation methods, Monte Carlo techniques are utilized to expand the range of data on which to test isolation performance. A description of this process is provided in this chapter.

2.2 Real-World Data Collection Techniques

2.2.1 Microphone and Sound Source Placement

For the real data collected in this study, a microphone array was set up in a typical office-type environment. A total of 8 microphones were regularly distributed on the ceiling of a support structure, and their positions were selected to represent a typical "smart-room" environment. A photo of the microphone array can be seen in Figure 2.1 while a 2D representation of their positions on the ceiling is shown in Figure 2.2.

Additionally, 6 locations were chosen as source positions for data collection. A loudspeaker was placed in each of these positions and an audio recording was played while

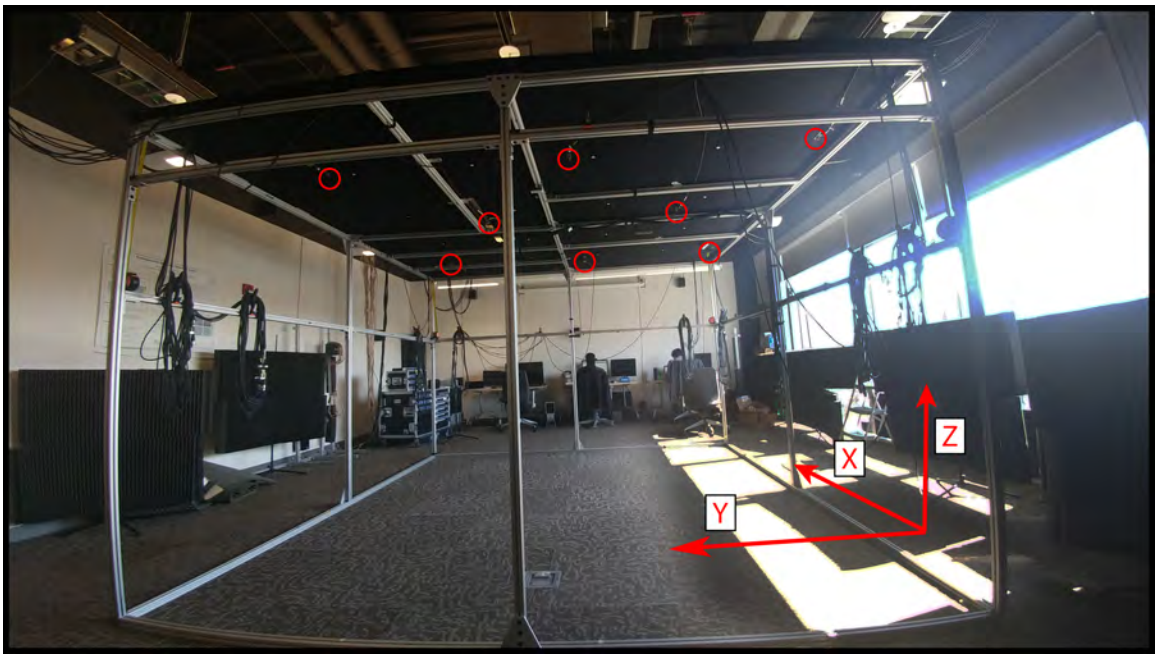

Figure 2.1: Photo of research lab with microphone array on ceiling of support structure

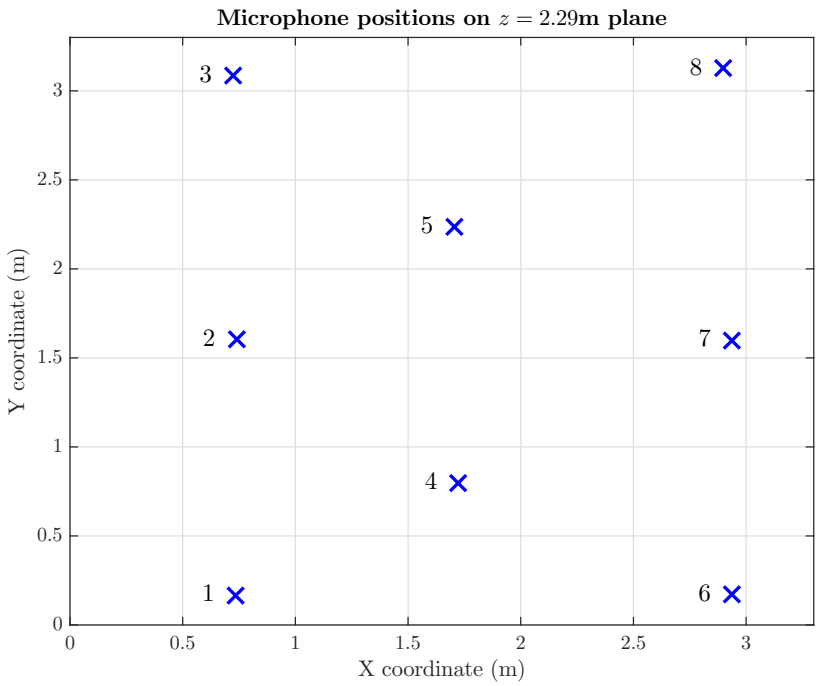

Figure 2.2: Distribution of microphone array on ceiling

the microphone array response was recorded. The volume of the source was subjectively set to approximately equal that of a human speaking at conversational volume. A 2D representation of these 6 source locations is shown in Figure 2.3.

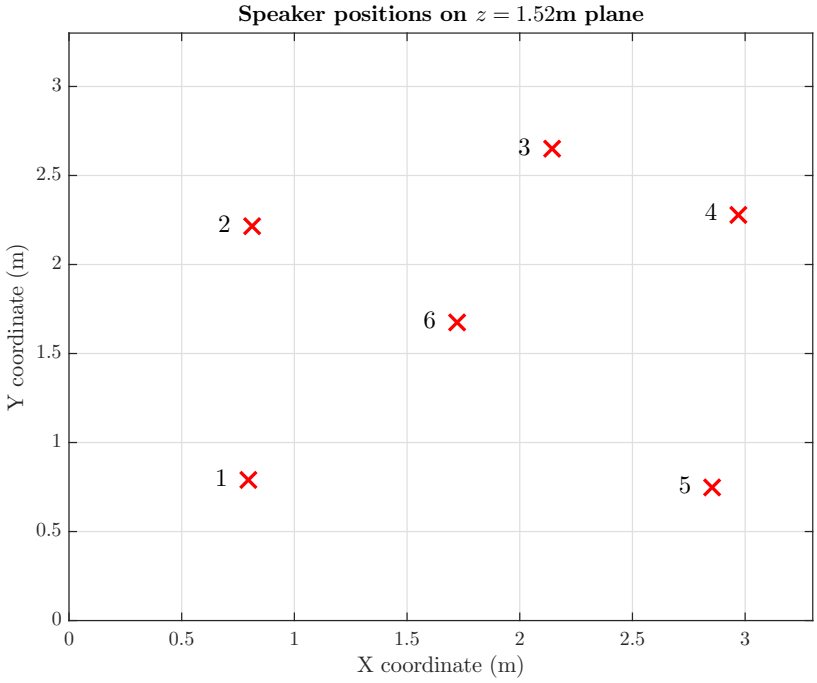

Figure 2.3: Sound source placements within the recording environment

The microphone signals were amplified with RME Octamic-D preamplifiers, sampled with RME HDSP9652 sound cards at 44.1 kHz, and downsampled to 16kHz for all processing and analysis.

2.2.2 Sound Sources and Recording

A collection of 8 different audio recordings are used as the sound sources in this study. Each recording contains a human voice speaking approximately 20 seconds of sentences. Both men and women voices are included in the set of recordings, and each recording contains a unique set of spoken words. For each of the 6 sound source locations in the recording environment, every source recording is played individually through the loudspeaker and the microphone array response is recorded. This yields a total of 48 total microphone array response recordings (6 sound source locations $\times$ 8 sound source recordings).

Because of the linear nature of the microphone response, these 48 recordings can be combined to create a full cocktail party scene (where multiple sources are active at once) as desired. For example, the microphone response to sound source # 1 at location #1 can be summed with the response to sound source #2 at location

#2, yielding an overall response equivalent to that if a single microphone response was recorded with both sound sources active simultaneously. Using this technique, a cocktail party scene can be created with anywhere from 1 to 6 active speakers in the scene. In addition, it provides a convenient way to parametrically diminish the speaker of interest for studying performance at high and low SNRs.

2.3 Simulation Techniques

To understand the effects of the analysis methods over a wide range of scenarios, virtual simulations are performed. A virtual environment can be defined to emulate that of a real-life scenario (source and microphone positioning, room reverberation effects, etc). A collection of MATLAB scripts (available online at the University of Kentucky Vis Center website [33]) is used to create this virtual environment. Placing sound sources at defined locations in the virtual environment allows for a simulation of one or more speakers simultaneously speaking in a room. Additionally, microphone arrays of varying geometries can be defined and simulated, and effects such as frequency-dependent attenuation and room reverberation can be accurately simulated, creating a realistic representation of a real-life scenario. Because there are no position measurement errors (as there could be with real recordings), simulations can provide an understanding of analysis performance in "ideal" scenarios, while also providing freedom from the requirements of a stationary and quiet lab space for recordings. Finally, simulations allow for great flexibility in quickly changing an environment, reducing time required to test various analysis techniques.

For this study, the simulated audio environment (microphone positions, source positions, and propagation speed of sound) were chosen to equal or closely match those parameters from the real recordings. This allows for the most meaningful comparison between results created from simulated and real microphone recordings. Similar to the real recordings, the simulations are performed with only a single source active in the scene. Using the same 8 source recordings (real voice recordings in a quiet room) as with the real recordings, 48 total simulated microphone responses are created and stored for future analysis.

2.4 SII Calculation

As described in Section 1.6, the Speech Intelligibility Index (SII) is used in this study to quantify the intelligibility of a given speech audio signal. The SII metric models nonlinear relationships to human intelligibility determined through extensive subjective testing and is based on weighted SNR values across an audible frequency range [26]. The SII calculation requires separation of the target source and interfering sources, and the microphone response data must be designed to accommodate the creation of signals which maintain target/interferer separation.

Because the simulation and real recordings were performed where only one source was active at a time, it is possible to maintain this signal and noise separation required for SII calculation. Consider an example cocktail party scene with 3 active sources (source #1 at location #1, source #2 at location #2, and source #3 at location #3). If we choose source #1 to be our Source Of Interest (SOI), meaning we want to isolate source #1 from the two interfering sources, a microphone response containing only the SOI, $\boldsymbol{X}_S$, can be taken as simply the microphone response simulated or recorded with source #1 active. To create the microphone response containing only the interfering sources (the "noise"), $\boldsymbol{X}_N$, the simulated or recorded response with source #2 active is summed with the response with source #3 active. Together, $\boldsymbol{X}_S$ and $\boldsymbol{X}_N$ fully represent the microphone response if it was simulated or recorded with all three sources active simultaneously, $\boldsymbol{X}_{S+N}$, because $\boldsymbol{X}_S + \boldsymbol{X}_N = \boldsymbol{X}_{S+N}$. This separability can be maintained throughout the isolation process (as described in Chapter 3), which facilitates the calculation of the SII metric.

2.5 Monte Carlo Techniques

Monte Carlo techniques are applied throughout the experiments in this study to survey a wide range of setup parameters. Though many parameters can be altered in each simulation or recording, this study primarily investigates algorithm performance for enhancing intelligibility in high levels of noise. The relative strengths of each active source (target and interferers) is varied, providing a range of relative SNRs between active sources. This allows for performance measurement of TF masking and other isolation techniques across a variety of conditions ranging from "worst-case" to "best-case" data.

As described in the previous sections, a cocktail party scene can be created at will by the simple summation of simulated or recorded microphone responses. To change the relative SNRs between source, we can simply pre-multiply the respective microphone responses before summation, which is used to weight each source's presence in the final response. Weighting each contributing source by a unity gain results in each active source having approximately the same power in the final microphone response, while deviating from unity factors directly adjusts the SNR between the corresponding sources. In this way, a small amount of simulated or recorded microphone array responses (total of 48) can be used to create an extremely large number of unique cocktail party scenes for analysis.

2.6 Overview

An overview of the complete process from data collection to performance evaluation is shown in Figure 2.4. First, a single SOI and k interfering audio speech recordings are selected for evaluation. Microphone and sound source locations are created in

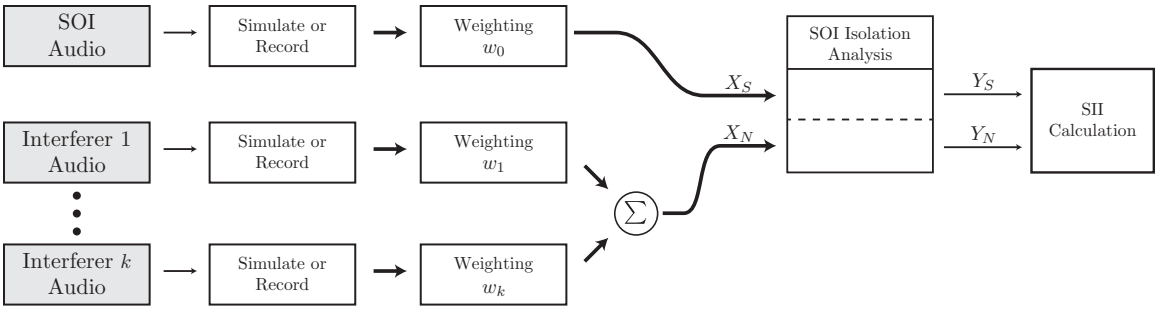

Figure 2.4: Overview of simulation, recording, and analysis process

a real audio environment, or virtually within a simulated environment. The microphone array response to the real/simulated environments are recorded/simulated for a single active source. Each of the array response vectors is weighted by its respective weighting factor ($w_0 \dots w_k$) to adjust the relative SNRs. The interfering sources are summed together into one overall noise response, X_N. This noise response is used along with the SOI microphone response, X_S, by the various SOI isolation techniques to create the final separate signal and noise outputs, Y_S and Y_N. Finally, these signal and noise outputs are used to calculate the SII performance metric.

2.7 Subjective Evaluation

Beyond the automated intelligibility measurement of the SII, we recognize the need to validate the speech intelligibility by a human listener. A formal subjective experiment is beyond the scope of this work, but we have defined a scale by which to subjectively rate a speech's intelligibility, which is shown below in Table 2.1.

Table 2.1: Scale for subjectively rating speech intelligibility

Subjective Rating	**Description**
No	No words are discernible. May not even hear target's presence.
Barely	Target's presence is detected, but words are only sparsely perceived.
Moderately	Multiple words are intelligible, but complete sentences or phrases are not discerned.
Mostly	Most words are accurately perceived, and sentences are sufficiently complete to discern meaning.
Yes	All words are accurately perceived with careful listening. Interfering noise may be present, but do not prevent complete intelligibility of target.

This informal subjective scale will be used for discussion, to compare the performance of the processing techniques, and to assess limitations and biases of the SII metric.

Chapter 3

Analysis Techniques

3.1 Introduction

With microphone array response data created as described in Chapter 2, techniques to isolate the SOI from interfering noise can be applied. The following sections describe the implementation of these isolation techniques, especially as it relates to maintaining the separability of signal and noise data for SII calculation.

3.2 Delay-Sum Beamformer Implementation

The delay-sum beamformer is not dependent on the nature of the signal itself, as it only applies a constant time-delay (equal to propagation time from source to microphone) to each sample in the signal. Because of this, the application and analysis of the DSB effects are relatively simple.

Consider an environment with a single target source (SOI) at position $\boldsymbol{r}_p$ and K interfering sources at positions $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_K$. As described earlier, two microphone responses are created (by simulation or recording) that either contain only the SOI or only the interfering speakers: X_S and X_N, respectively. Beamforming is applied to both of these signal vectors, selecting the beamformer focal point as $\boldsymbol{r}_p$. This results in a beamformed signal Y_S with only the target present and another beamformed signal Y_N (still beamformed *at* the target position $\boldsymbol{r}_p$) containing only the interferers, which are the two signals required for SII calculation. Finally, both of the resulting beamformed signals can be summed to give the overall beamformed result Y_{S+N}. An overview of this process can be seen in Figure 3.1.

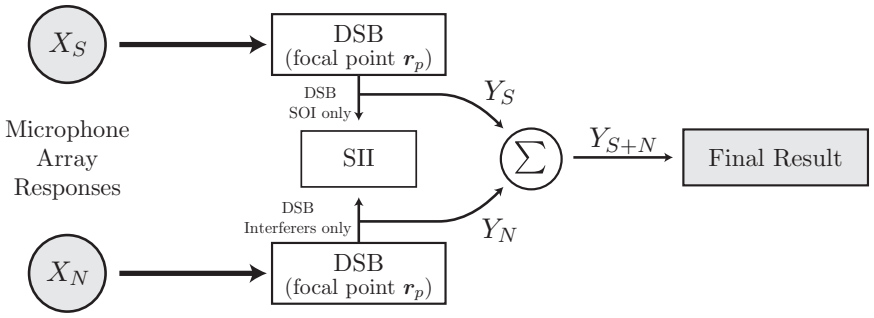

Figure 3.1: Overview of DSB analysis technique for single SOI and K interfering sources. Separate target and interfering signals are available for SII calculation.

3.3 Griffiths-Jim GSC Implementation

As described in Section 1.4, the GSC beamformer performs a simple delay-sum beamforming operation in conjunction with an *adaptive* blocking matrix branch. Because of this adaptive process, the target and noise signals can not simply be superimposed in the same manner as the DSB simulations, since the NLMS filters are based on past and current values of both the signal and noise. Instead, a technique is used to create maximal separability of target and noise signals, while maintaining the adaptive advantages of the GSC beamformer. This process is described below and an overview is shown in Figure 3.2.

We first continue with the previous notation where $\boldsymbol{X}_S$, and $\boldsymbol{X}_N$ represent data vectors containing the microphone response due to only SOI and only interferers, respectively. We also add $\boldsymbol{X}_{S+N}$ being the combination of these such that $\boldsymbol{X}_{S+N} = \boldsymbol{X}_S + \boldsymbol{X}_N$. However, we temporarily modify our notation such that these vectors have already been aligned to the target source; that is, the "delay" portion of a delay-sum beamformer has already been performed to steer the array response. These signals are then weighted and summed which completes the delay-sum beamforming process for these vectors. Each of these vectors is also converted to a blocking matrix: a linear process consisting of pair-wise differences of vector columns. The adaptive filter receives feedback from the total output Y_{S+N} and attempts to minimize the noise signature in the total output. To ensure the accuracy of the signal and noise outputs (Y_S and Y_N) such that $Y_S + Y_N = Y_{S+N}$, the signal and noise blocking matrices are filtered by the *same* filter as that created from the total output Y_{S+N}. This ensures that the nonlinear filtering process is applied equally to signal and noise vectors, which allows for the separability required for SII calculation.

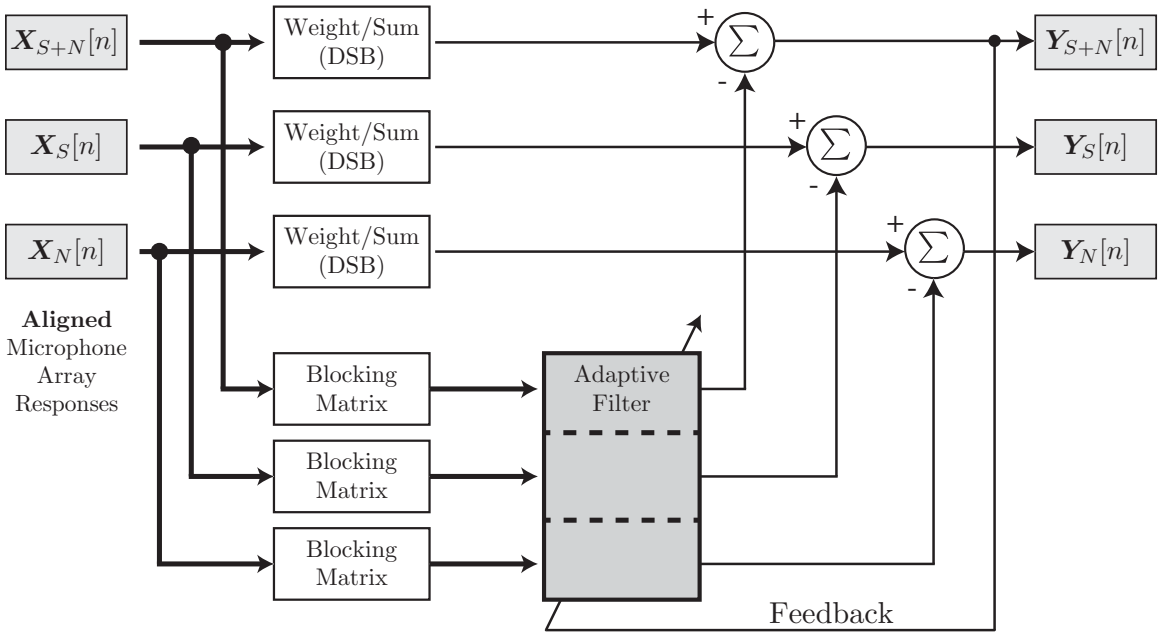

Figure 3.2: Overview of GSC analysis technique with maintained target signal and noise separation for SII calculation.

3.4 Time-Frequency Masking Implementation

In contrast to the DSB technique, the TF masker does not perform a constant operation across an entire audio signal. Instead, the TF binary mask dynamically updates corresponding to the relative spectral power ratio between target and interferers in a given TF window. For this reason, TF masking cannot be simply applied to target and interferer signals separately (as with the DSB) but, rather, requires the input of both the target and all individual interfering source signals.

Consider the example environment with a single target source (SOI) at positon $\boldsymbol{r}_p$ and K interfering sources at positions $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_K$, and with unaltered, un-steered, microphone response vectors $\boldsymbol{X}_S$ and $\boldsymbol{X}_N$. Beamforming is applied to the complete microphone response $\boldsymbol{X}_{S+N}$ at all source locations $(\boldsymbol{r}_p, \boldsymbol{r}_1, \ldots, \boldsymbol{r}_K)$, yielding $K+1$ beamformed signals. TF masking is done on these signals as described in Section 1.5, which gives the overall result of the TF masking process. However, to create separate target and noise signals for SII calculation, the binary TF masks are saved at the time of their calculation. These TF masks are then applied to two beamformed signals (focal point $\boldsymbol{r}_p$), one containing only the target source, and the other containing only the interfering sources, which are then used for SII calculation. This process allows for realistic TF binary mask creation (by calculating binary masks in the context of every source being active), while still maintaining separate target and noise signals for intelligibility estimation. An overview of this process can be seen in Figure 3.3.

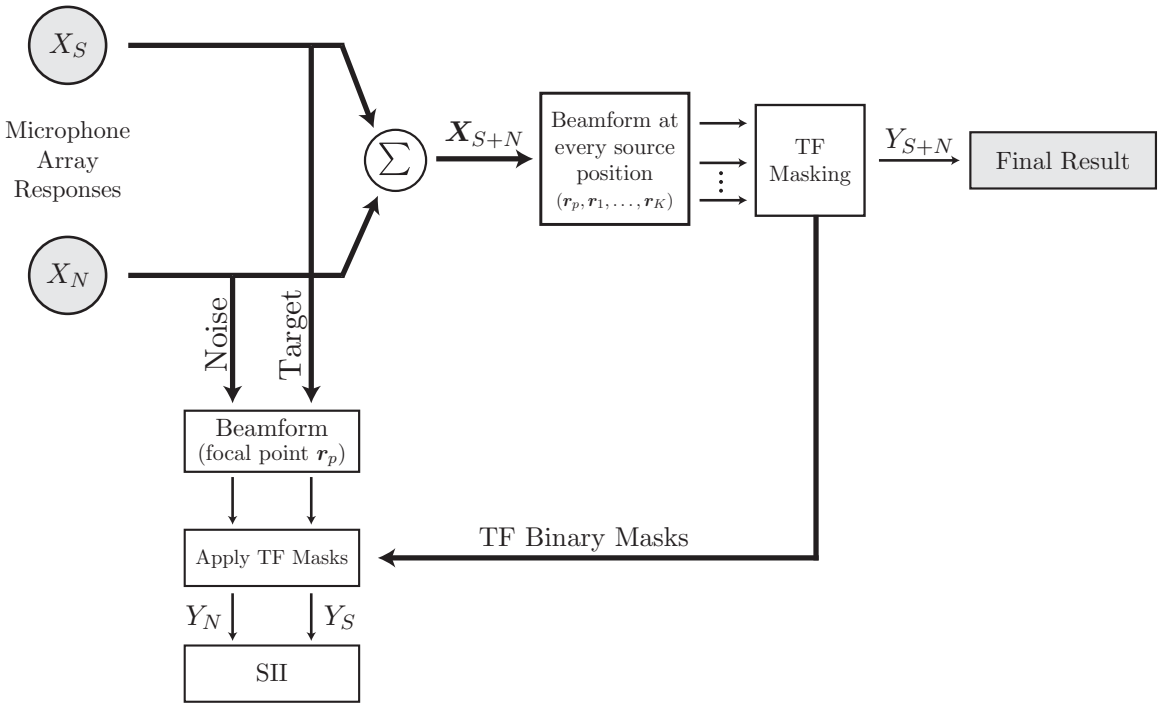

Figure 3.3: Overview of TF masking analysis technique for single SOI and K interfering sources. Separate target and interfering signals are available for SII calculation.

Chapter 4

Ideal TF-Masking Performance

4.1 Introduction

To begin our study of TF masking performance with microphone arrays, we first investigate the ideal, or best-case, performance. Recall from the discussions in Section 1.5 and Section 3.4 that the TF masking algorithm operates by creating binary TF masks through comparing the beamformed target TF signature to that of the interferers. In a practical application, the microphone array response will contain the target signal and the interfering noise together. To create reference signals for the TF comparisons, the array response is beamformed on each source (target *and* interferers) to create the best possible reference of each individual source for the mask creation process. Because audio from each source will be present in the beamformed response of other sources, the TF masking technique is unable to create perfect masks to block out interfering noise. Recall from the discussion of Equation 1.25 in Section 1.5 that when the target and noise energy occupy the same TF intervals, the masker performance depends on the beamformer performance. In other words, because some amount of the target signal is present in the beamformed response of a given interfering source, the associated binary mask may allow the interfering noise signal to pass through when it should not.

Although the target signal and interfering noise are not separable in a practical microphone array setup, our technique of maintaining signal/noise separability allows for the TF masking performance to operate in an ideal sense. That is, if we provide the mask creation algorithm signals containing only a single source (target or interferers), then the binary mask can be created in the best-case scenario, which is used to understand the ideal or upper-bound performance of the TF masking technique. Although the implementation of the TF masking is very similar to that described in Section 3.4, a slight modification is required to facilitate the creation of ideal binary masks. An overview of the modified implementation is shown in Figure 4.1.

It can be seen from the figure that the only modification to the TF masking imple-

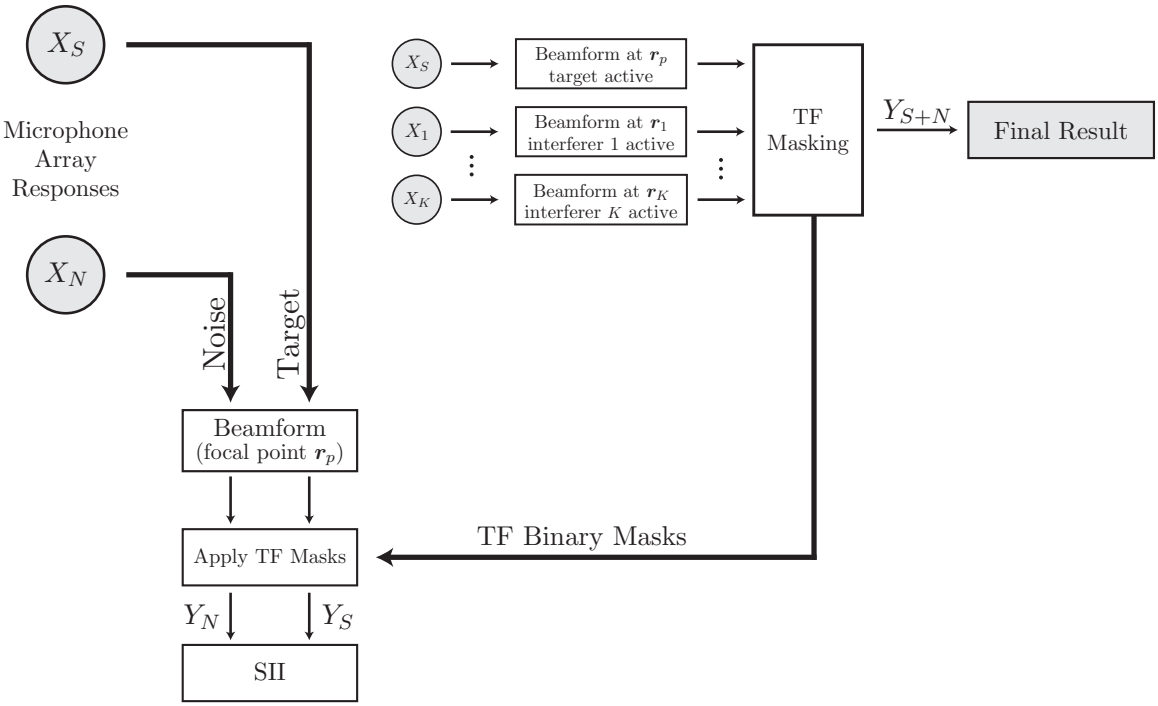

Figure 4.1: Overview of modified TF masking analysis technique for ideal performance evaluation

mentation is that the signals used for mask creation are pure signals containing only a single active source. This is used for the following experiment to determine ideal masking performance.

4.2 Experimental Setup

For this experiment, the data created from both the virtual simulations and real recordings are utilized, and the experiment is run separately on each set of data for comparison. A total of 2, 3, or 4 sound sources are selected to be active in the audio scene for each individual trial. For each trial setup, the individual sources are randomly weighted to create varying SNRs between active sources, and the sources are randomly selected from the set of 8 human speech recordings. In each trial, an individual unique microphone response is created with only one active source. Delay-sum beamforming is applied at each of the source locations and the results are passed into the TF masking algorithm for analysis.

4.3 Results and Discussion

4.3.1 Overall ideal TF masking performance

The speech intelligibility (SII metric) results from this experiment were plotted in Figure 4.2. Both the DSB output intellgibility and the TF masking output intelligibility were plotted against the raw intelligibility of the response from the closest microphone (the best reference of the pre-processing target intelligibility). The simulated data and the real recording data were included to show any differences between the simulation technique and the real-data performance.

It can be seen from the figure that in every case, the TF masking technique outperformed the DSB in increasing the speech intelligibility from the unprocessed closest-microphone signal. Although the DSB beamformer did slightly improve the SII from that of the close mic (primarily in the simulated data), the TF masking was a great improvement, especially in the range where close-mic SII was less than 0.25 (considered to be a threshold for human speech intelligibility).

The delay-sum beamformer is seen to not perform as well with the real data as that of the simulated data. In fact, in the case of 4 active sources, the DSB actually *decreased* the intelligibility from the unprocessed signal (though only slightly). This is likely due to a combination of several reasons. First, the real recording data, by nature, has a higher noise content than the corresponding simulation. Room noise from computers, HVAC equipment and the like, along with equipment noise and distortion effects, all help to create a higher noise floor than what is possible in the simulation. Secondly, the DSB performance is dependent on the knowledge of microphone and source positions in 3D space. These positions are precisely and identically known in the simulation (because they are precisely defined); however, measurement errors are certainly possible and even likely to occur in the real setup. Finally, the propagation speed of sound is not a source of error in the simulation because the processing algorithms assume the same propagation speed that the data is simulated with. Although the propagation speed is calculated from measurements of room humidity, pressure, and temperature for the real data, any error in this calculation (for example, from measurement equipment inaccuracy) will be equivalent to a positional error in the DSB algorithm. The DSB results do not match precisely between the simulated and real data analyses, but the trends are very similar and are within an acceptable range to lend credibility to the simulation results.

Although there are small differences between the simulated and real results, it is seen that they do closely match overall. In each case, the TF masking and DSB results have matching trends and differ on an absolute scale by only a few percent. The close match between the simulation and real data results increases confidence in the simulation's ability to accurately predict real-life performance. Finally, it is noted that the upper-bound of TF masking performance does not appear to depend on whether the data is simulated or real, and may represent a true best-case performance goal

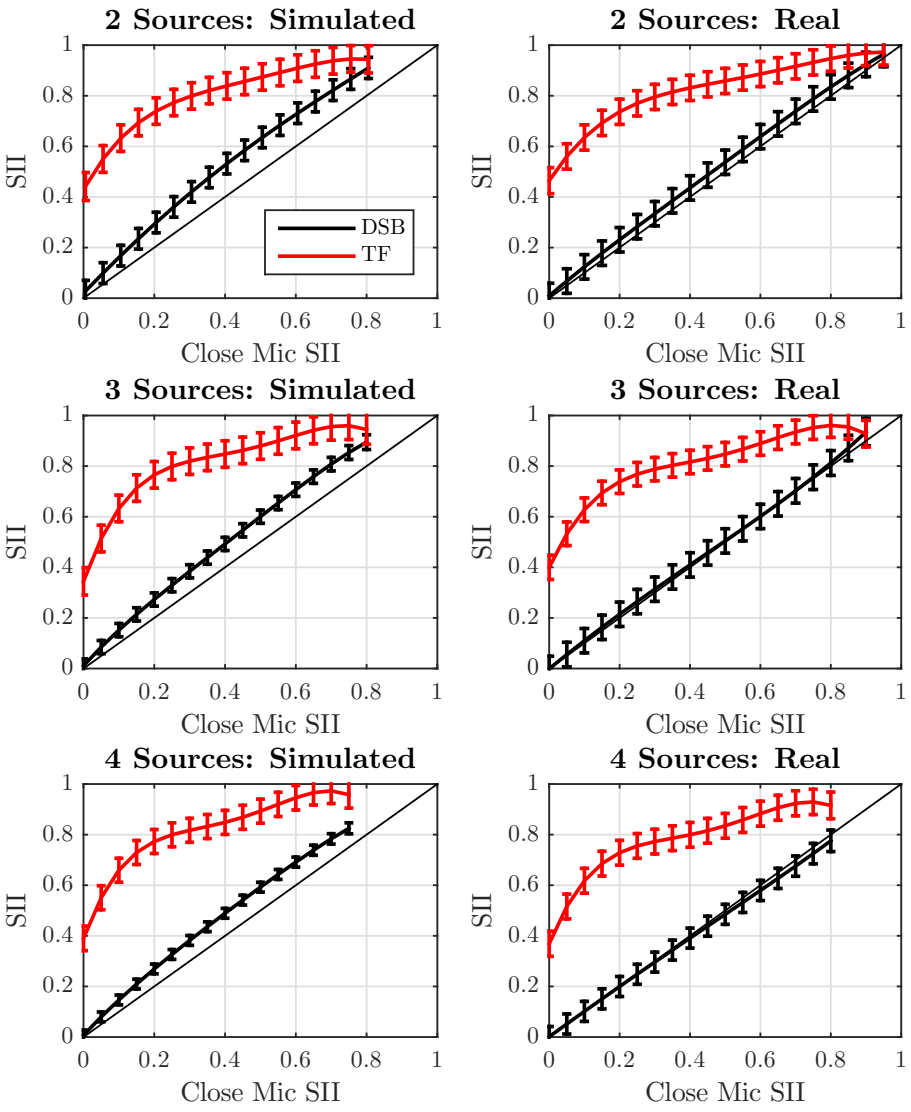

Figure 4.2: Intelligibility improvement of DSB and ideal TF masking for 2,3,4 active sources compared to closest microphone intelligibility (error bars depict depict one standard deviation to each side of the mean)

for the TF masking algorithm.

Lastly, the results show that the TF masking algorithm is most beneficial in the ranges of lower unprocessed intelligibility. First, the SII metric is bounded in $[0, 1]$ and, as such, the TF masking SII will approach the "unity line" (no improvement) as the original unprocessed SII approaches 1.0. Additionally, in the cases where the closest-mic SII is high (greater than approx. 0.6), the original unprocessed signal is already completely intelligible and the TF masking is unable to greatly improve on the

already-intelligible signal. Finally, there is a consequence of the ideal mask creation process: because the TF masking algorithm is operating on signals containing *only* a single target or interfering source, highly accurate binary masks can be created. These masks, then, contain information regarding the target signal and actually shape the interfering noise towards the target. In fact, subjective listening confirms that this ideal mask creation process can, consequently, create artifacts of the target signal in the noise signature used for SII calculation. Any amount of the target signal that leaks into the noise reference signal decreases the resulting SII metric. In this case, this effect is minimal, but may be cause for some artificial decrease in the TF masking upper-bound performance.

4.3.2 Ideal TF masking improvement over DSB

To show the improvement of TF masking relative to the DSB, the intelligibility results of both were plotted against each other in Figure 4.3. In this case, the SII of the DSB result is on the x-axis and the y-axis is the SII of the signal after TF masking processing. An SII of 0.25 is considered an approximate threshold where a speech signal (less than 0.25) is unintelligible and a signal (above 0.25) is intelligible to a human. The shaded areas of the figure denote where signals after only DSB processing were unintelligible (SII less than 0.25) but became intelligible (SII above 0.25) after TF masking.

These results show that the ideal TF masking technique offers significant improvement over the DSB, especially in cases where the DSB output was unintelligible. In fact, in every trial where the DSB output had an SII of less than 0.25, the ideal TF masking output was above 0.25. This means that TF masking improved every unintelligible DSB output to a result considered to be intelligible.

4.3.3 Subjective Performance

In addition to the SII performance metric, informal subjective listening was done by the author in the case of 4 simulated active sources. The intelligibility results are given in Table 4.1.

Table 4.1: Subjective listening assessment for 4 active sources in a simulated environment

Close Mic SII	**Close Mic**	**DSB**	**Ideal TF Masking**
0.1	No	No	Yes
0.2	No	Barely	Yes
0.6	Yes	Yes	Yes

The subjective assessment confirms that the ideal TF masking algorithm was able to improve an unintelligible signal to a completely intelligible result. Because the binary

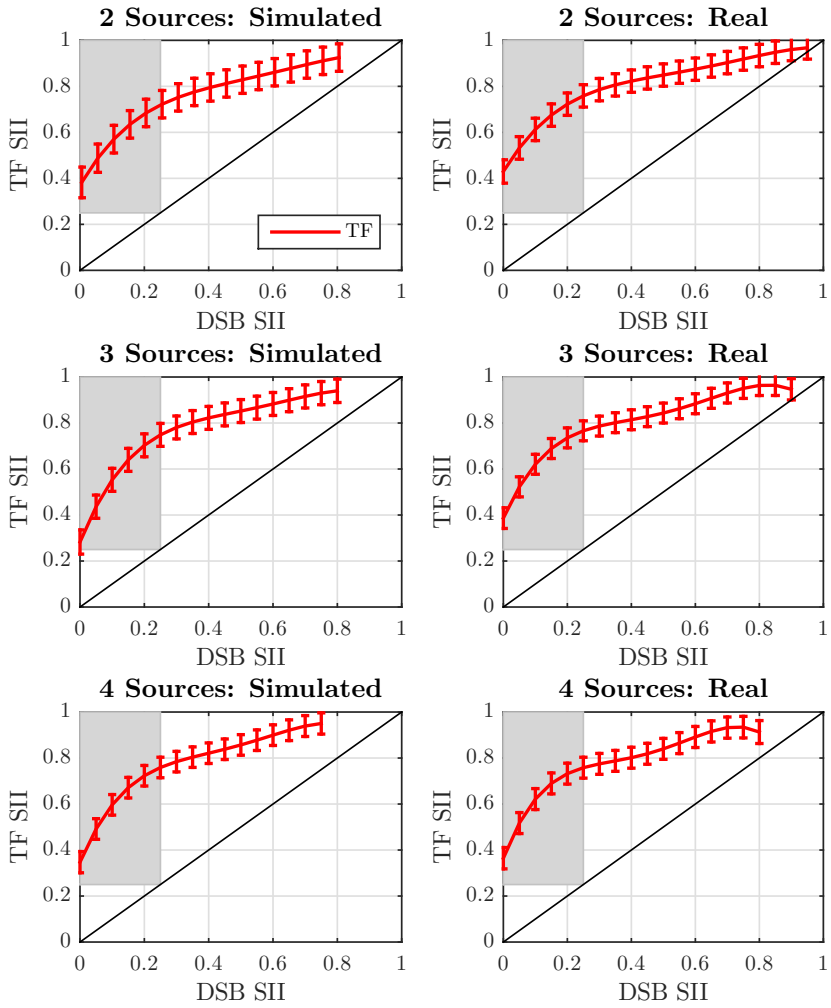

Figure 4.3: Intelligibility improvement from DSB to TF masking. Shaded area indicates where signal is unintelligible after DSB, but becomes intelligible after TF masking (error bars depict depict one standard deviation to each side of the mean)

TF masks were created ideally, the interfering noise in the TF masking output is essentially non-existent. Though there are some distortion artifacts in the TF masking output due to the sharp transitions of the mask, the speech remains completely intelligible and maintains the voice quality of the original speaker. The distortion artifacts created by the TF masker can best be described as making the voice occasionally "squeaky" with musical "blips" in the background.

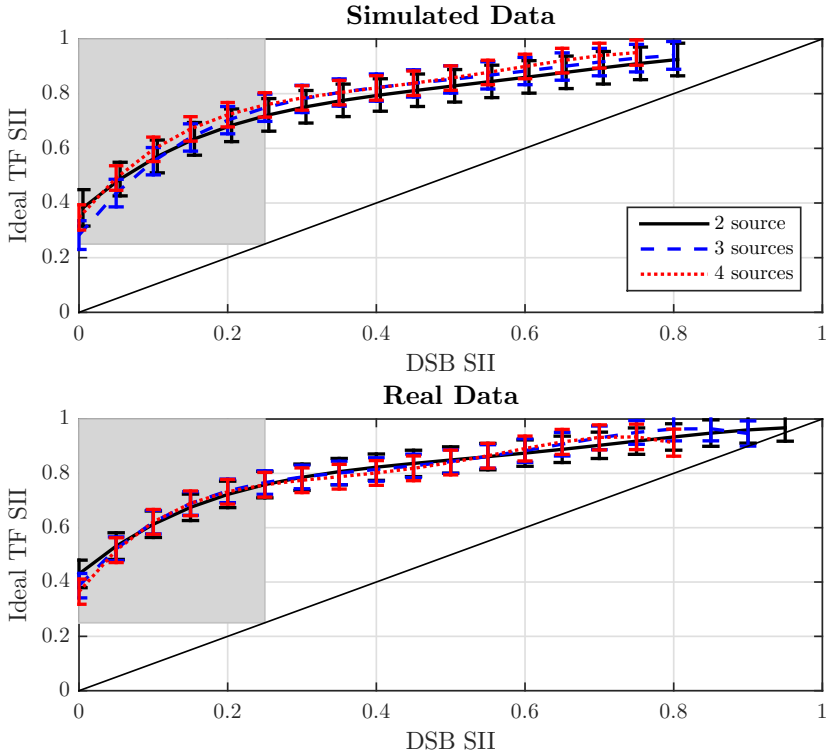

Figure 4.4: TF masking intelligibility improvement dependence on number of active sources (error bars depict depict one standard deviation to each side of the mean)

4.3.4 Ideal TF masking with increasing active sources

Our final analysis of the ideal TF masking results is an investigation of the effects of increasing active sources. We have performed experiments with 2, 3, and 4 active sources in the audio scene for both real and simulated data. The TF masking output SII is plotted against the DSB output SII for each quantity of active sources in Figure 4.4. As before, the shaded areas indicate improvement from unintelligible DSB output to intelligible TF masking output.

The results show that there is no clear dependence of the ideal TF masking performance on the number of active sources (for 4 or less sources). For the higher values of DSB SII, there is a slight direct relationship between increased active sources and increased TF masking performance, while the low values of DSB SII show a slight inverse relationship. The differences are not significant to draw any meaningful conclusions, though it can be stated that in all trials where the DSB output was unintelligible, the TF masking output *was* intelligible, regardless of the number of sources. This implies that the ideal TF masking algorithm is effective at both low numbers of active speakers, as well as increasing up to a full cocktail party scenario.

4.4 Conclusions

In this chapter, an experiment was presented to determine the upper-bound (or best-case) performance of the TF masking technique for 2, 3, and 4 active speakers. By adjusting the input signals to the TF masking algorithm to contain only a single active target or interfering source, the TF binary masks were created in an ideal fashion. The traditional delay-sum beamformer was used to create the input signals to the TF masking algorithm and its results were compared to the intelligibility improvement of the TF masking technique. The processing techniques were applied to both simulated data and real recorded data for comparison.

Through this experiment, it was determined that the ideal TF masking technique provided significant benefit for improving the intelligibility of a speech signal, as measured by both quantitative and subjective means. This was especially significant in the cases where performing a DSB operation yielded an unintelligible signal that was made intelligible by the TF masking processing. Though there were sources of error in this experiment (target signal leaking into interfering noise, positional errors with real recorded data, etc), these were determined to be negligible and the results showed consistent and expected behavior. It was additionally shown that the TF masking performance was not significantly affected by increasing the number of active sources, and was able to consistently improve the speech intelligibility in all cases. Further, we demonstrated a close match between the simulation results and those from real data and will continue to use the simulation technique alongside real recordings for the remainder of this study.

Finally, it should be restated that, because the TF masking algorithm was provided with pure single-active-source signals on which to operate, the TF masking results in this chapter represent an unrealistic best-case scenario and serve only to establish an estimated upper-bound on the expected TF masking performance. In the following chapter, we investigate the *practical* performance of the TF masking technique.

Chapter 5

TF Masking Improvements over Delay-Sum Beamforming

5.1 Introduction

Beamforming has been widely used to isolate a target source from amongst interfering noise, particularly with use of the traditional delay-sum beamformer. As such, we want to compare the practical performance of the delay-sum beamformer with that of the TF masking processing to understand the benefits of the masking technique. Unlike the previous experiment, we will be operating the TF masking script in a practical manner, meaning that the input data will contain both target signal and interfering noise which provides a realistic representation of what is possible in a real application.

5.2 Experimental Setup

As with the previous experiment, the data created from both the virtual simulations and real recordings are utilized, and the experiment is run separately on each set of data for comparison. A total of 2, 3, or 4 sound sources are selected to be active in the audio scene for each individual trial. For each trial setup, the individual sources are randomly weighted to create varying SNRs between active sources, and the sources are randomly selected from the set of 8 human speech recordings. For each trial, both delay-sum beamforming and TF masking is performed on the microphone response containing both target and noise to evaluate its practical performance. Note that the delay-sum beamformer is used to create the input signals to the TF masking algorithm, as described previously

5.3 Results and Discussion

5.3.1 Overall DSB and TF masking practical performance

The intelligibility results of this experiment are displayed in Figure 5.1. The TF masking output and DSB output SII values are plotted against the SII value of the microphone's signal closest to the target source. The results for 2, 3, and 4 sources are shown for both the real and simulated data.

In these results, TF masking again shows a clear performance enhancement over simple beamforming. The delay-sum beamformer was effective at increasing the SII some amount from that of the close mic (primarily with the simulated data), but the TF masking was able to further improve the intelligibility beyond the DSB.

In this experiment, the TF masking was provided input signals containing both the target signal and interfering noise, which is a realistic setup. The results show a rather significant decrease in TF masking performance from the upper-bound established in the previous experiment. Recall from the discussion on the TF masking algorithm that the beamformed response for each active source is used as the reference signal for binary mask creation. Because these beamformed signals contain some of the other active sources (recall eq.(1.25): beamformer gain at off-focal points is non-negligible), the binary masks cannot be ideally created as in the ideal experiment. This has a detrimental effect on the TF masking algorithm's ability to block out interfering noise, which is reflected in these results.

Finally, we again note a close match between the simulated and real data results in the practical application of the TF masking algorithm. Note that there is a limited test range for the case of 4 simulated sources: this is merely an artifact of the Monte Carlo technique used to create the range of input data.

5.3.2 Practical TF masking vs DSB performance

To show the relative performance of the TF masking technique and its improvement over the simple DSB analysis, the intelligibility results of both were plotted against each other in Figure 5.2. As before, the shaded areas indicate improvement from unintelligible to intelligible output as a result of applying TF masking after DSB processing.

It can be seen in the figure that, as discussed, the TF masking regularly improves the intelligibility of the signal past the DSB output. If compared to the corresponding plot from the previous experiment (Figure 4.3), it is seen that there is a significant decrease in the algorithm's performance. This is, again, due to the practical nature of the reference signals used for mask creation. Because the TF masking has as its input the results of DSB operations, the TF masking performance is dependent on the performance of the preceding beamformer.

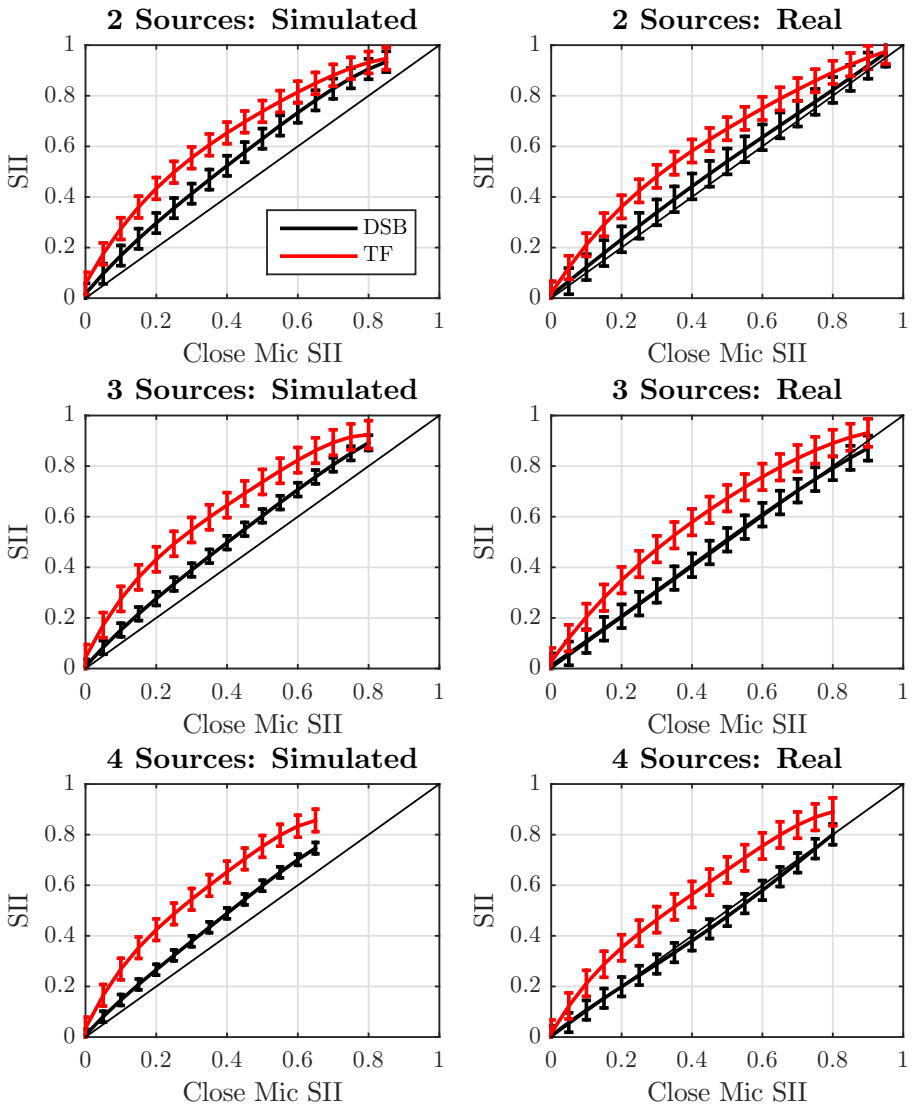

Figure 5.1: Intelligibility improvement of DSB and TF masking for 2,3,4 active sources compared to closest microphone intelligibility (error bars depict predicted standard deviation of results)

The primary results of interest are those that fall in the shaded area - scenarios where TF masking enhanced the signal from unintelligible to intelligible. In the results of this experiment, there are times when the TF masking did, in fact, improve the signal from unintelligible to intelligible. However, it is shown that there is a range of scenarios (DSB intelligibility values below approx 0.15) where TF masking did *not* improve the SII value to a point considered intelligible. In these cases, the TF masking did not improve the SII value above the estimated intelligibility threshold

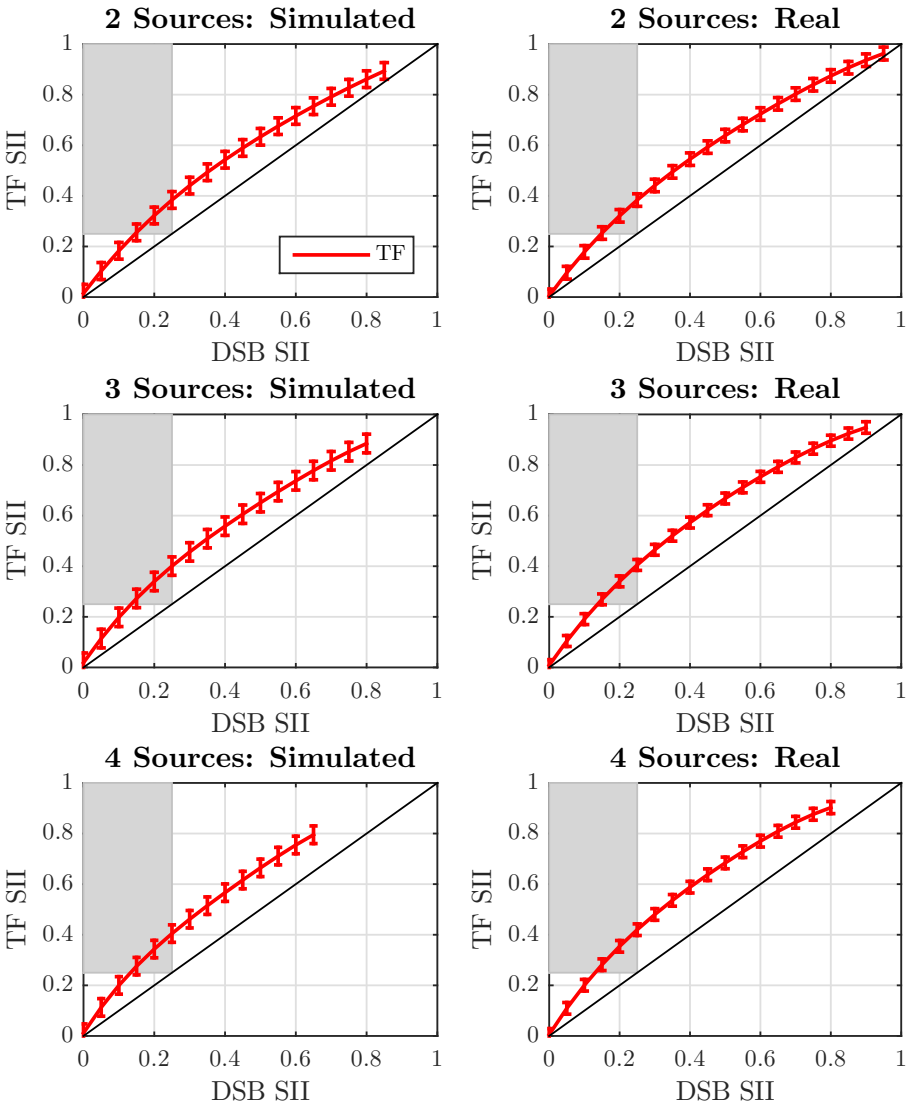

Figure 5.2: Intelligibility improvement from DSB to TF masking. Shaded area indicates where signal is unintelligible after DSB, but becomes intelligible after TF masking (error bars depict predicted standard deviation of results)

of 0.25, but did, however, improve the SII by close to 100% in some cases. That is, even in cases where the final result was not predicted to be intelligible, there was still a measurable intelligibility benefit of applying TF masking.

5.3.3 Subjective Performance

As before, subjective listening was done by the authors and the results for a simulated audio scene with 4 active sources are shown in Table 5.1.

Table 5.1: Subjective TF masking intelligibility improvement over DSB, with 4 active sources in a simulated environment

Close Mic SII	**Close Mic**	**DSB**	**DSB-TF Masking**
0.1	No	No	Barely
0.2	Barely	Moderately	Mostly
0.6	Yes	Yes	Yes

Although the SII predicted the inability of TF masking to improve an unintelligible DSB-output signal to an intelligible signal, the authors' subjective listening revealed consistent improvement by applying TF masking. In this case, the TF masking technique was not as effective as that in the ideal case, which is to be expected due to the practical implementation used in this experiment. The TF masking results are only shown for a single auditory scene (4 active sources, simulated data), but they are consistent with subjective performance from the other audio scenes from this chapter.

Subjectively, the TF masker enhances the speech intelligibility by increasing the perceived SNR of the target, and also by distorting the interference. While the interference of the simple DSB signal is clearly speech, the TF masking output contains noise that, while reminiscent of human speech, is highly distorted. In fact, it is very difficult to discern any words spoken by the interfering speakers (though the suggestion of sentence/phrase structure is maintained to some degree). The noticeable improvement in subjective speech intelligibility, in contrast with the marginal intelligibility increase as predicted by the SII, demonstrates the inability of the SII to predict human speech perception in all cases.

5.3.4 Practical dependence of TF masking on number of active sources

We finally investigate the effects of the number of active sources on the TF masking performance. The TF masking results (plotted against the DSB output SII level) are plotted for 2, 3, and 4 sources in Figure 5.3.

It was shown in Chapter 4 that, for the ideal case, there is not a strong influence on TF masking performance by the number of active sources in the audio scene. It is

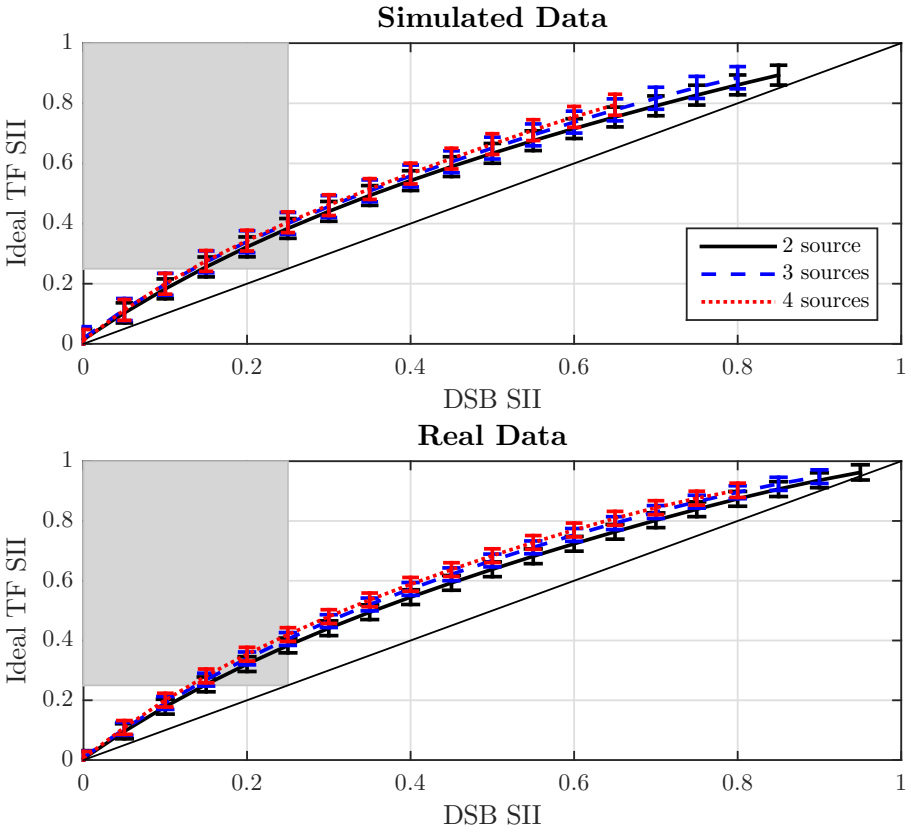

Figure 5.3: TF masking intelligibility improvement dependence on number of active source (error bars depict predicted standard deviation of results)

likewise seen here that in the practical case, no significant affects from the number of active sources are seen in the results. Thus, we can conclude that TF masking is practically effective with few active sources as well as "cocktail party" scenarios.

5.4 Conclusions

This chapter presented an experiment used to evaluate the practical performance of Time-Frequency masking on simulated and real data. By performing the processing in a way that is representative of a real-life setup, a realistic understanding of TF masking performance (and how it relates to the delay-sum beamformer) was achieved.

It was found that, as expected, the practical performance of the TF masking technique was reduced from that of the ideal case established in Chapter 4. Even so, TF masking was shown to improve the intelligibility of an audio signal in every case from that with only beamforming applied. Although there were cases where the TF masking processing did not improve the speech intelligibility to an acceptable level as measured by the SII metric, there were still regions where TF masking *did* subjectively improve intelligibility from an unacceptable level to an acceptable one. Additionally, it was found through informal subjective assessments that many of the cases the SII metric identified as being unintelligible (while accurate for beamforming alone), were actually intelligible by a human listener in the case of TF masking. Further, it was shown that, as with the ideal case, the performance of TF masking was not strongly influence by the number of active sources in the audio scene. Finally, the reliance of TF masking on its beamformed inputs was discussed and how deficiencies in beamformer will adversely affect the TF masking algorithm's ability to create appropriate TF masks and improve intelligibility. In the following chapter, we investigate the effects of using an adaptive beamformer to create the requisite input signals for the TF masking algorithm.

Chapter 6

TF Masking with Adaptive Beamforming

6.1 Introduction

The TF masking process compares the target signal and interfering noise data to create the resulting binary masks. Up to this point, delay-sum beamforming has been exclusively used to create the target and interferer reference signals for this process. Recall from the discussion of Equation 1.25 that masker performance is dependent on beamformer gains when target and interferers are active simultaneously. If, then, the beamformer gain terms can be improved (increasing the ratio between focal point and off-focal beamformer gains), we hypothesize that TF masking performance will be enhanced. Because the Griffiths-Jim general sidelobe canceler (GSC) is a distinctly different approach to removing interfering speech and has been shown to improve speech intelligibility over that of the delay-sum beamformer [5], we attempt the novel approach of using the GSC beamformer to create the TF masking input signals. The improvement that the GSC provides over traditional DSB beamforming suggests that TF masking will perform better with GSC-created input signals. This experiment is designed to evaluate the improvement, if any, that GSC provides to the TF masking process.

6.2 Experimental Setup

The experimental audio environment is setup identically to that of the previous experiment (Section 5.2). For this experiment, however, not only is delay-sum beamforming performed and TF masking performed with DSB input signals, but GSC beamforming is also applied along with TF masking with GSC-beamformed input signals. This allows for direct comparison between the DSB and GSC beamformers, as well as the TF masking techniques with DSB or GSC inputs.

6.3 Results and Discussion

6.3.1 Overall performance of TF masking with and without adaptive beamforming

The intelligibility results from this experiment are shown in Figure 6.1. In this figure, the SII value of each processing output is plotted against the SII value of the closest microphone to the target source. As before, the DSB output and TF masking (using DSB signals as the input signals to the TF masking algorithm) SII values are plotted. However, the GSC adaptive beamforming output and TF masking (using GSC signals as the input signals) are plotted as well for comparison.

These results show that, for low starting SII values (close mic SII less than approximately 0.3), the GSC-TF masking algorithm matches or slightly outperforms the DSB-TF masking performance. However, for starting SII values higher than this, the DSB-TF masking outperforms the GSC-TF masking technique. This implies that for low-intelligibility starting signals, the GSC-TF has negligible or slight improvement from that of the DSB-TF masker. For higher values (where the close mic signal might already be intelligible), the results imply that DSB-TF masking is preferred.

There is a clear dip in GSC beamformer and GSC-TF masking performance as the starting SII value is increased. In particular, it can be seen that the GSC-TF masking results closely follow the trends of the GSC beamformer. Recall that for the GSC-TF masking algorithm, a GSC beamformer is first applied to all active sources and those beamformed signals are used as input signals to the GSC-TF masking algorithm. It follows, then, that a decrease in GSC beamformer performance (as we see for higher SII starting values) would yield a decrease in GSC-TF masking performance as well.

We can also discern from these results that the GSC beamformer noticeably outperforms the DSB beamformer at low starting SII values, particularly with increased number of active sources and also with the real recording data. The GSC beamformer has been shown to increase speech intelligibility over the delay-sum beamformer in existing literature and these results agree.

Again, we note that the simulated data is a good representation of the real recording data in these results. There is a noticeable exception, however, that the GSC beamformer and GSC-TF masking algorithm for the real data under-perform their corresponding performance with simulated data. This is partially due to the increased noise floor of the real recording data. The GSC beamformer adapts to minimize the power of its total output power, with the assumption that their is none of the target signal present in its noise reference signal. However, because some of the target signal will certainly leak into the noise reference, it is likely that the GSC actually attenuates some of the desired target signal in its output. This effect is primarily seen for high (greater than 0.5) SII starting values. Because the signals in this range are mostly intelligible to begin with, its effects may be negligible.

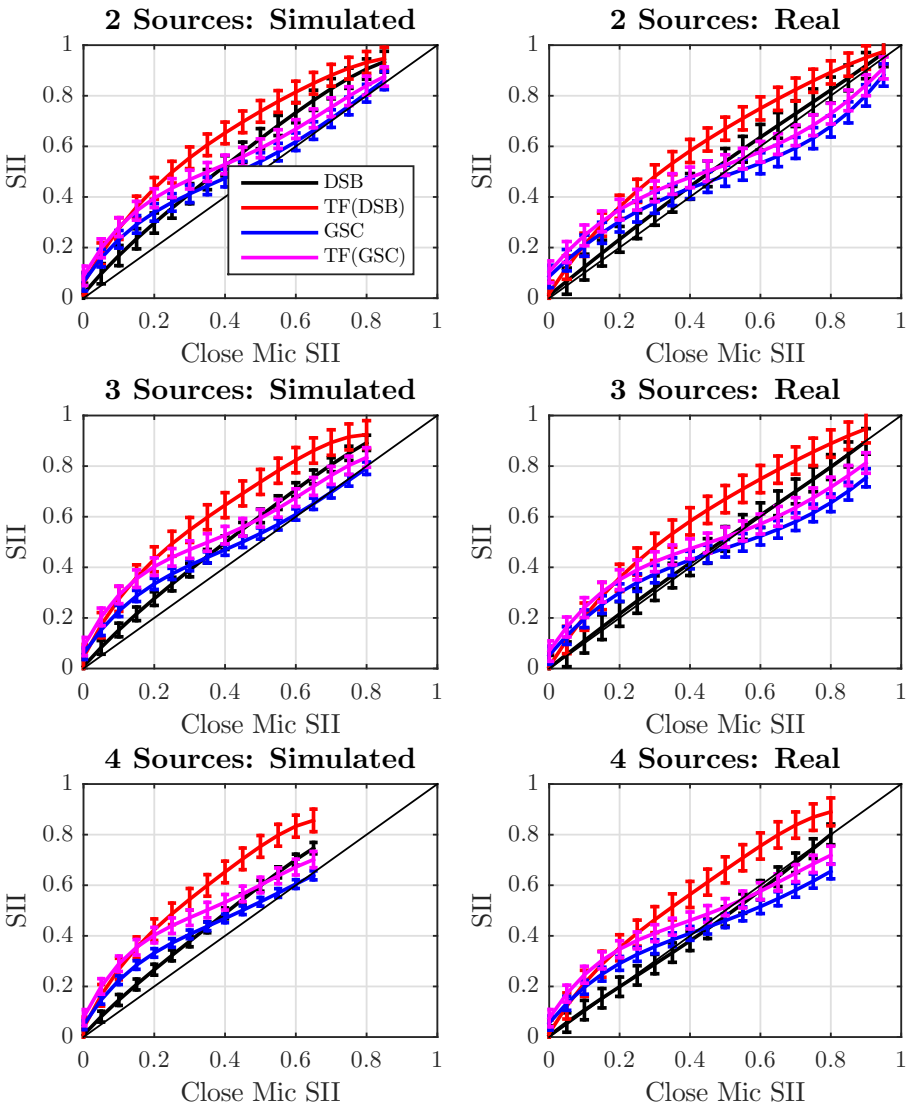

Figure 6.1: Intelligibility of DSB, TF masking using DSB inputs, adaptive beamforming (GSC), and TF masking using GSC inputs (error bars depict depict one standard deviation to each side of the mean)

Lastly, there is a consequence of our analysis technique that will decrease the measured performance of the GSC beamformer and GSC-TF masking processes. In our discussion in Section 3.3, we described a method for maintaining separability of target and noise signals through the adaptive beamforming process. Because the adaptive filters of the GSC use the total output signal as their reference signal, the actual filtering process shapes the filtered signals corresponding to the output signal. When the pure target and pure noise signals are passed through these adaptive filters, these

signals are likewise shaped corresponding to the total output reference signal. This is primarily noticed in the pure noise signal because the target signal is leaked in. Because the SII calculation relies on relative SNRs between signal and noise, any target signal that is found in the interfering noise signature will decrease the SII value, artificially lowering the SII metric. For cases where the target signal is stronger to begin with (ie higher starting SII values), the filters will more strongly shape the noise reference, exacerbating this effect. This is likely the primary cause for the dip in performance for higher close-mic SII values that is not seen for low close-mic SII values.

6.3.2 DSB-TF and GSC-TF performance vs DSB

We next look at the difference of TF masking with traditional delay-sum and adaptive beamforming in Figure 6.2. We choose the SII value of a signal having only the traditional DSB applied as a common reference for both the DSB-TF and GSC-TF masking techniques. This allows for a consistent comparison of the two methods and provides for an understanding of the benefits, if any, of using adaptive beamforming in conjunction with TF masking.

We can see from this view of the results that the GSC-TF masking slightly outperforms the DSB-TF masking results in all cases where the DSB output was considered unintelligible (DSB SII < 0.25). This is most noticeable in the real recording data where the effect is more apparent. For values where the DSB output was considered intelligible to begin with (DSB SII > 0.25), the GSC-TF masking performs less well than the DSB-TF, though this is likely primarily an effect of SII calculation issues as described in the previous section. Additionally, this range where the GSC-TF under-performs the DSB-Tf has high initial intelligibility and, as such, the decrease in performance does not significantly matter.

Lastly, the adaptive beamforming technique, along with the GSC-TF masking technique, introduce more distortion effects than the DSB and DSB-TF masking does. This can cause a decrease in the SII measurement of intelligibility more so than the actual human intelligibility decreases. This is another cause for potential mismeasurements of the SII value. Formal subjective testing will help to clarify the amount of this error, but is beyond the scope of this work.

6.3.3 Subjective Performance

Subjective testing by the authors was performed to validate the SII intelligibility prediction of TF masking with adaptive beamforming. The subjective intelligibility measurements are shown below in Table 6.1, where the GSC column is a measure of the General Sidelobe Canceler (Griffiths-Jim) adaptive beamformer, and GSC-TF is the TF masking results using adaptive beamforming.

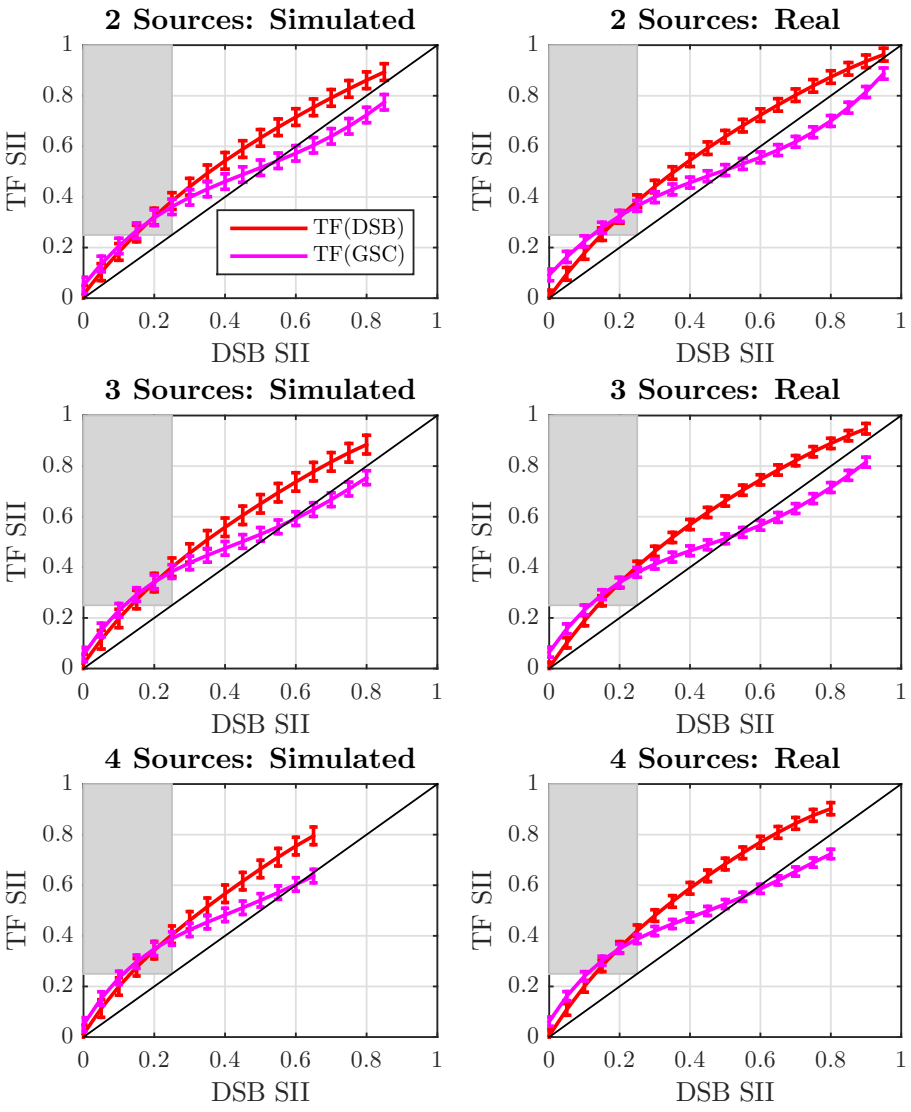

Figure 6.2: TF masking with DSB and adaptive beamforming inputs compared to DSB only intelligibility (error bars depict depict one standard deviation to each side of the mean)

It was determined that the adaptive beamforming TF masking technique did not noticeably improve the speech intelligibility in the case of close-microphone SII equaling 0.1. However, we noted that the interfering noise was more distorted using adaptive beamforming inputs, rather than simple DSB inputs; that is, it was more difficult to pick out the interfering speakers using the adaptive beamforming inputs. When using the DSB inputs, the interference is suggestive of human speech (though highly distorted), while the GSC inputs cause the interference to be non-suggestive of any

Table 6.1: Subjective TF masking intelligibility improvement (with and without adaptive beamforming), with 4 active sources in a simulated environment

Close Mic SII	**Close Mic**	**DSB**	**GSC**	**DSB-TF Masking**	**GSC-TF Masking**
0.1	No	No	No	Barely	Barely
0.2	Barely	Moderately	Moderately	Mostly	Yes
0.6	Yes	Yes	Yes	Yes	Yes

human speech.

Further, we found that when the close-microphone SII equaled 0.2, the simple beamforming TF masking inputs yielded a mostly intelligible signal, while the adaptive beamforming inputs yielded a fully intelligible signal. Finally, in the case where even the close-microphone was completely intelligible, the TF masking output using adaptive beamforming again contained interfering noise less perceivable than that of the TF masking using the simple DSB.

Although, there are cases where the adaptive beamforming inputs did not subjectively improve the speech intelligibility over that of the DSB beamforming inputs, the decrease in interferer intelligibility caused by adaptive beamforming suggests that, in other cases beyond those tested here, the adaptive beamforming technique may be beneficial in increasing target intelligibility.

6.3.4 TF masking with adaptive beamforming dependence on number of active sources

Our final analysis of the TF masking technique with adaptive beamforming is to investigate its dependence on the number of active sources. The performance results for 2,3, and 4 sources are shown in Figure 6.3.

As with the previous analyses, the GSC-TF masking technique is not shown to be significantly affected by the number of active sources in this study. We can again conclude that the TF masking with adaptive beamforming technique is equally effective for few active sources up to a full cocktail party scenario.

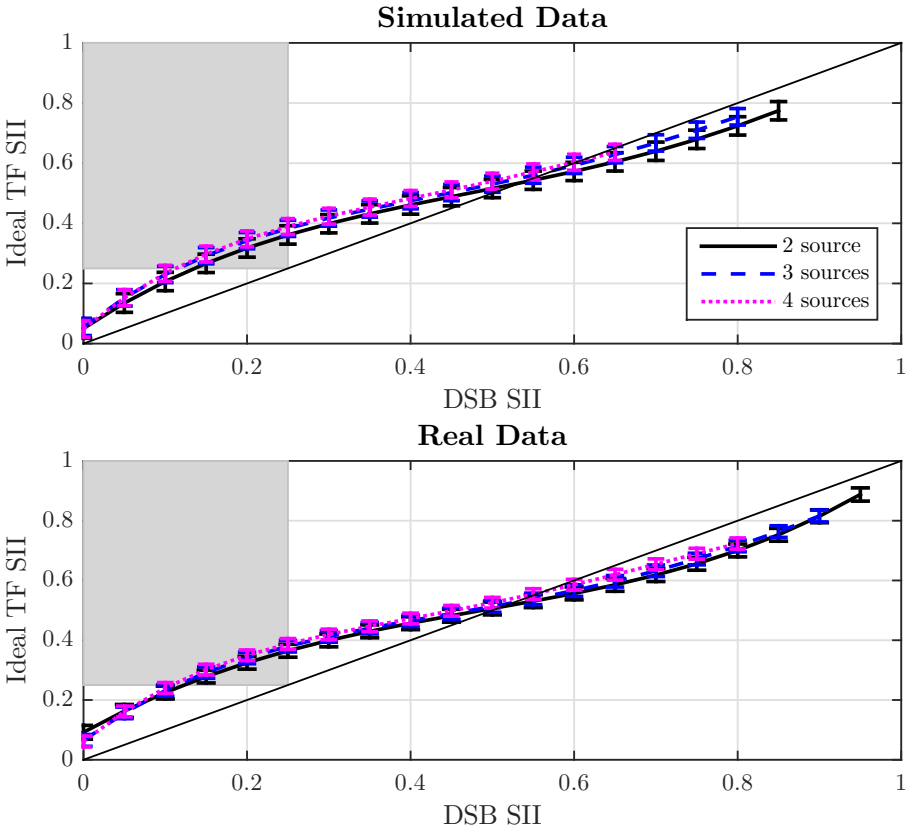

Figure 6.3: TF masking intelligibility using adaptive beamforming dependence on number of active sources (error bars depict depict one standard deviation to each side of the mean)

6.4 Conclusions

In this chapter, we presented an experiment to investigate the novel technique of combining Time-Frequency masking with established adaptive beamforming techniques. The results of this experiment were compared to those of the TF masking technique with traditional delay-sum beamforming. The experiment was performed using 2, 3, and 4 active sources in the audio scene, and the target isolation methods were applied to both the simulated and real recording data.

We found that for the important range of analysis (where the close-mic or traditional DSB outputs were unintelligible), the adaptive beamforming TF masking technique matched or outperformed the DSB-TF masking results. Additionally, it was shown that, as expected, the GSC adaptive beamformer performed better than the simple DSB beamformer for lower levels of initial intelligibility. Although the GSC and GSC-TF masking results were less effective at higher levels of initial intelligibility (as measured by SII), it was determined that this could be a consequence of non-perfect SII calculation due to target/noise leakage and distortion effects.

By comparing the GSC-TF and DSB-TF masking process to a common reference (DSB output intelligibility), we showed that the GSC-TF masking was beneficial over the DSB-TF masking for low levels of initial intelligibility. Additionally, informal assessment showed that the SII measurement was pessimistic in its evaluation of the improvements to low initial intelligibility levels, especially when the nature of the noise was highly distorted speech. Further, the effects of distortion and its effects on quantifying intelligibility with the SII metric were presented and discussed. Finally, it was determined that the TF masking used in conjunction with adaptive beamforming was not significantly affected by increasing the number of active sources in the audio scene which is consistent with the experimental results of the previous chapter.

Chapter 7

Final Conclusions and Future Work

The goal of this thesis and research was to further develop the application of Time-Frequency masking using distributed microphone arrays and to evaluate its performance. By creating a large amount of test data through recording and simulation techniques, overall trends were determined as to the improvement TF masking provides over other target isolation techniques. Using the automated Speech Intelligibility Index metric and informal subjective assessment allowed for the measurement of improvement by the TF masking technique. A series of experiments were presented and their results have led to the following:

- We developed an upper-bound for the intelligibility improvement caused by the TF masking algorithm. This was done through operating the TF masking in an ideal, yet unpractical, manner.
- TF masking was shown to improve speech intelligibility beyond that of delay-sum beamforming, as measured by both the SII and subjective ratings.
- TF masking, when used in conjunction with *adaptive* beamforming can positively impact speech intelligibility over using simple non-adaptive beamforming.
- Our technique of simulating an audio scene is effective at creating a representative dataset on which to test the various target isolation techniques presented in this work.
- The SII metric is useful for automated quantification of speech intelligibility, but does not perfectly predict a human's ability to discern speech from amongst interfering noise.
- Our techniques for analyzing the data while still maintaining target/interferer separation (as required for SII calculation) are effective and useful.

In this work, we have tested the performance of TF masking compared to other isolation techniques over a large dataset. By using Monte Carlo methods, we tested these techniques over an assortment of scenarios by varying relative source SNRs, number of active sources, source positions, and audio speech recordings. However,

our work did not include varying the physical environment (such as room reverberation), nor did we include a study of the dependence on microphone array positioning. Because other work has determined a dependence on the physical distribution of the microphones in the array [21], further work investigating its effects on TF masking may be beneficial. Additionally, these experiments studied TF masking performance exclusively with an array containing only 8 microphones; as such, future research of this topic may find value in studying the relationship between array size and isolation performance.

We determined in our experiments that the SII metric was valuable in predicting speech intelligibility (especially for use in Monte Carlo experiments), but did not perfectly predict human intelligibility. This issue was worsened particularly in cases where we were unable to maintain perfect target/interfering isolation throughout the TF masking or beamforming algorithms. The "leaking" of target or interfering sources into each other lowers the SII value, while not necessarily lowering the human intelligibility. Additionally, subjective assessments indicated that the SII metric was pessimistic in its evaluation of TF masker intelligibility, primarily due to its inconsideration of the nature of interference. For low values of intelligibility, the nature of interference became important wherein beamforming-only signals had speech-like noise, while TF masked signals contained highly distorted speech noise. Though the SII predicted similar intelligibility measures to these signals, the latter was subjectively more intelligible. To better understand the intelligibility of a speech signal, future work should include a formal subjective study or improved automated intelligibility metric.

Overall, we have shown that TF masking is an effective technique for improving speech intelligibility of a target source in a "cocktail party" environment, and that the technique of combining TF masking with adaptive beamforming can further increase intelligibility. Future study of this topic should investigate the effects of microphone array and audio environment parameters that this study did not, further describing the intelligibility benefits of Time-Frequency masking.

Bibliography

[1] W. C. Liao et al. "An effective low complexity binaural beamforming algorithm for hearing aids". In: *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. Oct. 2015, pp. 1–5.

[2] A. Mahmoodzadeh et al. "Binaural speech separation based on the time-frequency binary mask". In: *6th International Symposium on Telecommunications (IST)*. Nov. 2012, pp. 848–853.

[3] Y. Jiang, H. Zhou, and Z. Feng. "Performance analysis of ideal binary masks in speech enhancement". In: *2011 4th International Congress on Image and Signal Processing*. Vol. 5. Oct. 2011, pp. 2422–2425.

[4] Iain McCowan, Jason Pelecanos, and Sridha Sridharan. "Robust Speaker Recognition using Microphone Arrays". In: *IN PROCEEDINGS OF 2001: A SPEAKER ODYSSEY*. 2001.

[5] L. Griffiths and C. Jim. "An alternative approach to linearly constrained adaptive beamforming". In: *IEEE Transactions on Antennas and Propagation* 30.1 (Jan. 1982), pp. 27–34. ISSN: 0018-926X.

[6] Nicoleta Roman and John Woodruff. "Intelligibility of reverberant noisy speech with ideal binary masking". In: *The Journal of the Acoustical Society of America* 130.4 (Oct. 2011), pp. 2153–2161. URL: https://doi.org/10.1121%2F1.3631668.

[7] Eric W. Healy et al. "An algorithm to improve speech recognition in noise for hearing-impaired listeners". In: *The Journal of the Acoustical Society of America* 134.4 (2013), pp. 3029–3038. eprint: http://dx.doi.org/10.1121/1.4820893. URL: http://dx.doi.org/10.1121/1.4820893.

[8] A. Mahmoodzadeh et al. "Binaural speech separation based on the time-frequency binary mask". In: *6th International Symposium on Telecommunications (IST)*. Nov. 2012, pp. 848–853.

[9] D Wang. "Time–frequency masking for speech separation and its potential for hearing aid design." In: *Trends in Amplification* 12.4 (2008), pp. 332–353. ISSN: 1084-7138.

[10] H. Unnikrishnan, K. D. Donohue, and J. Hannemann. "Time-frequency masking for speaker of interest extraction in an immersive environment". In: *IEEE SOUTHEASTCON 2014*. Mar. 2014, pp. 1–8.

[11] Phil Townsend. "Enhancements to the Generalized Sidelobe Canceller for Audio Beamforming in an Immersive Environment". Master's Thesis. University of Kentucky, 2009.

[12] O. L. Frost. "An algorithm for linearly constrained adaptive array processing". In: *Proceedings of the IEEE* 60.8 (Aug. 1972), pp. 926–935. ISSN: 0018-9219.

[13] J. P. Townsend and K. D. Donohue. "Stability Analysis for the Generalized Sidelobe Canceller". In: *IEEE Signal Processing Letters* 17.6 (June 2010), pp. 603–606. ISSN: 1070-9908.

[14] A. Hussain, K. Chellappan, and S. Z. M. "Evaluation of multichannel speech signal separation with beamforming techniques". In: *2014 IEEE Conference on Biomedical Engineering and Sciences (IECBES)*. Dec. 2014, pp. 766–771.

[15] D. Bagchi et al. "Combining spectral feature mapping and multi-channel model-based source separation for noise-robust automatic speech recognition". In: *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. Dec. 2015, pp. 496–503.

[16] J. Dennis and T. H. Dat. "Single and multi-channel approaches for distant speech recognition under noisy reverberant conditions: I2R'S system description for the ASpIRE challenge". In: *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. Dec. 2015, pp. 518–524.

[17] Adel Hidri, Souad Meddeb, and Hamid Amiri. "About Multichannel Speech Signal Extraction and Separation Techniques". In: *CoRR* abs/1212.6903 (2012). URL: http://arxiv.org/abs/1212.6903.

[18] Joonas Nikunen. "Microphone Array Post-Filtering Using Supervised Machine Learning for Speech Enhancement". In: *In: Proc. 15th Annual Conference of the International Speech Communication Association (Interspeech)*. 2014.

[19] Jitong Chen et al. "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises". In: *The Journal of the Acoustical Society of America* 139.5 (2016), pp. 2604–2612. eprint: http://dx.doi.org/10.1121/1.4948445. URL: http://dx.doi.org/10.1121/1.4948445.

[20] Y. Murase et al. “On microphone arrangement for multichannel speech enhancement based on nonnegative matrix factorization in time-channel domain”. In: *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*. Dec. 2014, pp. 1–5.

[21] Jingjing Yu. “Microphone Array Optimization in Immersive Environments”. PhD thesis. University of Kentucky, 2013.

[22] T. Yoshioka and T. Nakatani. “A microphone array system integrating beamforming, feature enhancement, and spectral mask-based noise estimation”. In: *2011 Joint Workshop on Hands-free Speech Communication and Microphone Arrays*. May 2011, pp. 219–224.

[23] Shuyang Cao, Liang Li, and Xihong Wu. “Improvement of intelligibility of ideal binary-masked noisy speech by adding background noise”. In: *The Journal of the Acoustical Society of America* 129.4 (2011), pp. 2227–2236. eprint: `http://dx.doi.org/10.1121/1.3559707`. URL: `http://dx.doi.org/10.1121/1.3559707`.

[24] Kirstin M. Brangers. “Perceptual Ruler for Quantifying Speech Intelligibility in Cocktail Party Scenarios”. Master’s Thesis. University of Kentucky, 2013.

[25] Jianfen Ma, Yi Hu, and Philipos C. Loizou. “Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions”. In: *The Journal of the Acoustical Society of America* 125.5 (2009), pp. 3387–3405. eprint: `http://asa.scitation.org/doi/pdf/10.1121/1.3097493`. URL: `http://asa.scitation.org/doi/abs/10.1121/1.3097493`.

[26] American National Standards Institute. *American National Standard Methods for Calculation of the Speech Intelligibility Index, Std S3.5*. Acoustical Society of America, 1997.

[27] A. Schlesinger. “Transient-based speech transmission index for predicting intelligibility in nonlinear speech enhancement processors”. In: *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Mar. 2012, pp. 3993–3996.

[28] T. Fukumori et al. “Estimation of speech recognition performance in noisy and reverberant environments using PESQ score and acoustic parameters”. In: *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*. Oct. 2013, pp. 1–4.

[29] Ulrik Kjems et al. “Role of mask pattern in intelligibility of ideal binary-masked noisy speech”. In: *The Journal of the Acoustical Society of America* 126.3 (2009), pp. 1415–1426. eprint: `http://dx.doi.org/10.1121/1.3179673`. URL: `http://dx.doi.org/10.1121/1.3179673`.

[30] S. Vladimir et al. “Intelligibility Assessment of Ideal Binary-Masked Noisy Speech with Acceptance of Room Acoustic”. In: *Journal of Electrical Engineering* 65 (Jan. 2015), pp. 325–332.

[31] Chengzhu Yu et al. “Evaluation of the importance of time-frequency contributions to speech intelligibility in noise”. In: *The Journal of the Acoustical Society of America* 135.5 (2014), pp. 3007–3016. eprint: http://dx.doi.org/10.1121/1.4869088. URL: http://dx.doi.org/10.1121/1.4869088.

[32] M. Schaefer et al. “Numerical Near Field Optimization of Weighted Delay-and-Sum Microphone Arrays”. In: *IWAENC 2012; International Workshop on Acoustic Signal Enhancement*. Sept. 2012, pp. 1–4.

[33] Kevin D Donohue. *Distributed Audio Lab*. URL: http://vis.uky.edu/distributed-audio-lab/.

[34] O. Yilmaz and S. Rickard. “Blind separation of speech mixtures via time-frequency masking”. In: *IEEE Transactions on Signal Processing* 52.7 (July 2004), pp. 1830–1847. ISSN: 1053-587X.

[35] Dorothea Kolossa et al. “Independent Component Analysis and Time-Frequency Masking for Speech Recognition in Multitalker Conditions”. In: *EURASIP Journal on Audio, Speech, and Music Processing* 2010.1 (2010), p. 651420. ISSN: 1687-4722. URL: http://dx.doi.org/10.1155/2010/651420.

[36] Donal G. Sinex. “Recognition of speech in noise after application of time-frequency masks: Dependence on frequency and threshold parameters”. In: *The Journal of the Acoustical Society of America* 133.4 (2013), pp. 2390–2396. eprint: http://dx.doi.org/10.1121/1.4792143. URL: http://dx.doi.org/10.1121/1.4792143.

[37] Yipeng Li and DeLiang Wang. “On the optimality of ideal binary time-frequency masks”. In: *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. Mar. 2008, pp. 3501–3504.

Vita

Joshua Morgan graduated in May 2016 from the University of Kentucky Summa Cum Laude with honors and receiving a Bachelor of Science in Electrical Engineering and Minor of Mathematics. He attended UK supported by the Otis A. Singletary and external scholarships. During his undergraduate career, Joshua held numerous leadership positions in engineering student organizations, most notably including Electrical Team Lead and overall Team Manager of the UK Solar Car Team. He additionally completed 3 co-op experiences before graduating. While studying at UK, he was awarded the honors of UK College of Engineering Dean's List for all semesters, the Eta Kappa Nu Outstanding ECE Junior Award, H. Alex Romanowitz Award, Robert L. Cosgriff Senior ECE student award, UK College of Engineering Alumni Associate Senior Leadership award, and the Takacs Co-op of the Year Award.