5-12-2020

A Perceptual Evaluation of Short-Time Fourier Transform Window Duration and Divergence Cost Function on Audio Source Separation using Non-negative Matrix Factorization

Ryan J. Miller

Recommended Citation

Miller, R. J. (2020). "A Perceptual Evaluation of Short-Time Fourier Transform Window Duration and Divergence Cost Function on Audio Source Separation using Non-negative Matrix Factorization." *Master of Science in Audio Engineering (MSAE) thesis,* Belmont University, Nashville, TN. 6. https://repository.belmont.edu/msaetheses/6

A Perceptual Evaluation of Short-Time Fourier Transform Window Duration and Divergence Cost Function on Audio Source Separation using Non-negative Matrix Factorization

Master's thesis presented to the faculty of the
Audio Engineering Graduate Program
of
The Mike Curb College *of* Entertainment *&* Music Business
Belmont University, Nashville TN

In partial fulfillment of the requirements for the degree

Master of Science
with a major in
Audio Engineering

Ryan J. Miller
May 12th, 2020

Advisors

Wesley A. Bulla
Song Hui Chon
Scott Hawley
Doyuen Ko
Eric Tarr

ABSTRACT

Non-negative matrix factorization (NMF) is an established method of performing audio source separation. Previous studies used NMF with supplementary systems to improve performance, but little has been done to investigate perceptual effects of NMF parameters. The present study aimed to evaluate two NMF parameters for speech enhancement: the short-time Fourier transform (STFT) window duration and divergence cost function. Two experiments were conducted: the first investigated the effect of STFT window duration on target speech intelligibility in a sentence keyword identification task. The second experiment had participants rate residual noise levels present in target speech using three different cost functions: the Euclidian Distance (EU), the Kullback-Leibler (KL) divergence, and the Itakura-Saito (IS) divergence. It was found that a 92.9 ms window duration produced the highest intelligibility scores, while the IS divergence produced significantly lower residual noise levels than the EU and KL divergences. Additionally, significant positive correlations were found between subjective residual noise scores and objective metrics from the Blind Source Separation (BSS_Eval) and Perceptual Evaluation method for Audio Source Separation (PEASS) toolboxes. Results suggest longer window durations, with increased frequency resolution, allow more accurate distinction between sources, improving intelligibility scores. Additionally, the IS divergence is able to more accurately approximate high frequency and transient components of audio, increasing separation of speech and noise. Correlation results suggest that using full bandwidth stimuli could increase reliability of objective measures.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

1.1 Purpose of the Study

The purpose of this study was to determine if common parameters used in audio source separation algorithms affect the perceptual qualities of speech separated from noise. Non-negative Matrix Factorization (NMF) was used to perform audio source separation of speech with additive noise mixtures. Two experiments were conducted to examine different NMF parameters: first, the effect of the short-time Fourier transform (STFT) window duration on separated speech intelligibility was assessed. Second, the effect of the divergence cost function used on the level of residual noise present in separated speech was examined through both objective measurements and a subjective evaluation.

1.2 Research Question and Hypothesis

The research question for the first experiment was, “does the STFT window duration used for NMF speech enhancement affect the intelligibility of the enhanced speech?” The null hypothesis was, “there is no significant difference in intelligibility of speech when using different STFT window durations for NMF speech enhancement.” It was predicted that longer window durations with greater frequency resolution will improve separation and result in greater intelligibility. For the second experiment, the research question was, “does the divergence cost function used in NMF speech enhancement have a significant effect on the perceived level of residual noise?” The null hypothesis was, “the divergence cost function used will not cause perceptible changes in residual noise levels.” It was predicted that the Itakura-Saito cost function, which is commonly used for speech processing, will provide the most noise reduction.

1.3 Significance of the Study

Results from this study will provide insight as to the perceptual effects of the STFT window duration and divergence cost function within source separation algorithms on the separated sources. Previous studies implementing NMF focused on developing supplementary systems to improve separation performance but failed to evaluate the effects of NMF's internal parameters. Other studies have acknowledged the potential impact of using different window durations or cost functions from a purely objective standpoint without subjective quality evaluations. This study aims to reveal the influence of window duration and cost function on perceived speech enhancement performance and to provide clarity for parameter selection decisions in practical source separation applications.

2.0 PRIOR ART

2.1 Algorithm Description

Audio source separation is a method of decomposing a mixture of distinct audio sources into its individual components. NMF [1] has become one of the more popular algorithms for performing audio source separation and has been used in many applications including music remixing [2], noise removal [3, 4], automatic speech recognition [5, 6] and speech enhancement [7, 8]. Many studies have sought to improve separation performance by augmenting NMF with auxiliary data [9, 10] or deep neural networks [11 - 13]. However, there has been limited research exploring the effects of NMF's base parameters on the perceived quality of separation.

NMF approximates a non-negative matrix $\boldsymbol{V}$ of dimensions n x m as the product of two other non-negative matrices $\boldsymbol{W}$ and $\boldsymbol{H}$ where $\boldsymbol{W}$ has dimensions n x r and $\boldsymbol{H}$ has dimensions of r x m, such that:

$$\boldsymbol{V}_{n,m} \approx \boldsymbol{W}_{n,r}\boldsymbol{H}_{r,m} \tag{1}$$

In audio applications, $\boldsymbol{V}$ is typically the magnitude spectrogram of an audio mixture containing multiple sources with n frequency bins and m time intervals. The rank r is selected such that r < min(n, m) to create a compressed version of $\boldsymbol{V}$ that will reveal latent structure in the mixture [1, 14]. $\boldsymbol{W}$ contains the spectral basis functions for the individual sources within the mixture, while $\boldsymbol{H}$ contains the time-activation gains for each basis function. In this way, linear combinations of the rows of $\boldsymbol{W}$ and the columns of $\boldsymbol{H}$ can be used to extract individual sources within $\boldsymbol{V}$. Figure 1 illustrates the NMF approach to source separation. Here, the sources in $\boldsymbol{V}$ are two harmonic series which are each represented by a single basis function and corresponding time-activation gain. The rank r is 2 since the mixture is approximated by two pairs of basis functions and activation gains.

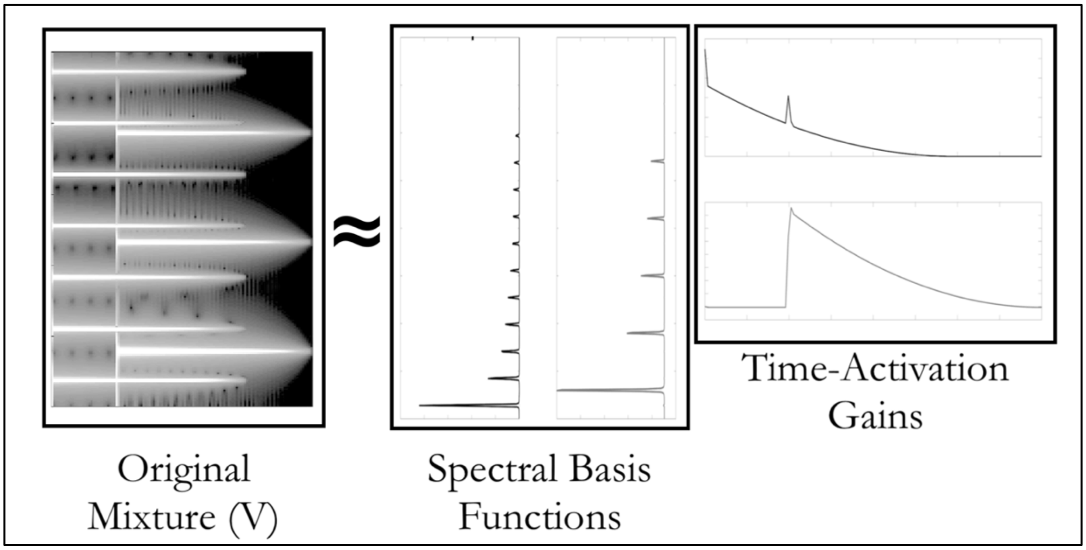


Figure 1. Visual example of NMF source separation.

Figure 2 shows a block diagram of the audio source separation process using NMF. ***V*** is obtained by taking the magnitude of the mixture's STFT. NMF can then be applied to generate ***W*** and ***H*** approximation matrices. The basis functions and activation gains for each source are then used to create a series of masking filters (mask) which can be used to extract each source from the original mixture. The phase information from the mixture is then used for the inverse STFT to convert the sources back into the time domain.

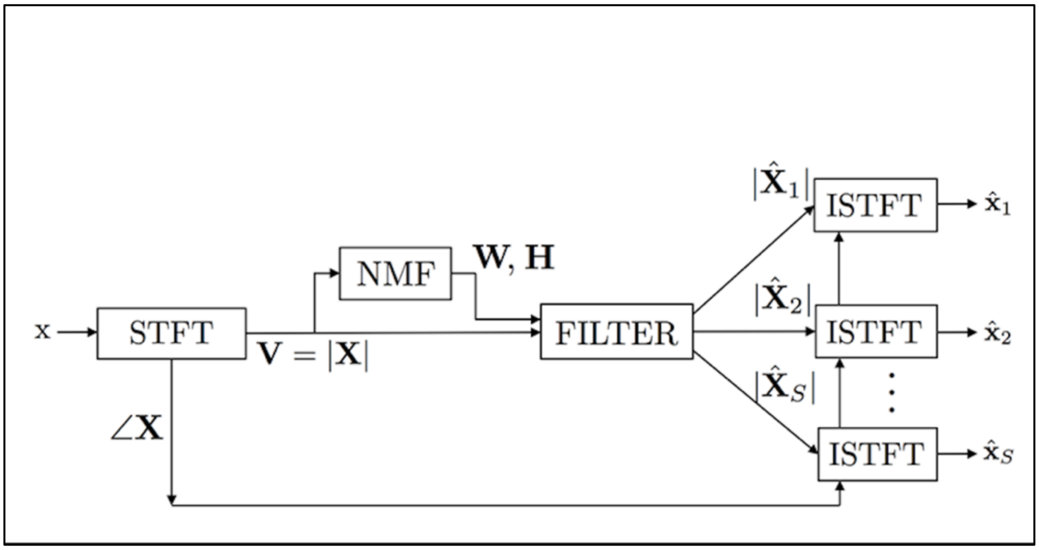


Figure 2. Block diagram of audio source separation using NMF [15].

W and ***H*** can be computed by minimizing the divergence of their product from ***V***. A divergence cost function is used to quantify the divergence. The objective of NMF then is to find values for ***W*** and ***H*** such that their product has minimal divergence from ***V***, more formally:

$$min_{\boldsymbol{W},\boldsymbol{H}\geq 0} D(\boldsymbol{V} \parallel \boldsymbol{WH}) \tag{2}$$

where *D* is the cost function evaluated between ***V*** and ***WH***. Three common cost functions based on the Bregman Divergence family are typically used in NMF source separation: the Euclidian Distance (EU), the Kullback-Leibler (KL) Divergence, and the Itakura-Saito (IS) Divergence [16]. These cost functions are calculated as shown:

$$D_{EU}(V \parallel WH) = \sum_{n,m} (V_{n,m} - WH_{n,m})^2 \tag{3}$$

$$D_{KL}(V \parallel WH) = \sum_{n,m} \left(V_{n,m} \log\left(\frac{V_{n,m}}{WH_{n,m}}\right) - V_{n,m} + WH_{n,m} \right) \tag{4}$$

$$D_{IS}(V \parallel WH) = \sum_{n,m} \left(\frac{V_{n,m}}{WH_{n,m}} - \log\left(\frac{V_{n,m}}{WH_{n,m}}\right) - 1 \right) \tag{5}$$

Here, "$\boldsymbol{WH}_{n,m}$" represents the product of ***W*** and ***H*** indexed at *n*, *m*, and "log" is the natural logarithm. In all three cases, element-wise calculations are summed over all *n*, *m* to produce the total divergence. These cost functions can be rearranged into multiplicative update rules which can be applied iteratively to ***W*** and ***H*** to minimize the total divergence from ***V***, provided that ***W*** and ***H*** are non-negative [17]. ***W*** and ***H*** are typically initialized with random non-negative values. The update rules for each cost function are shown below:

$$EU: \quad W \leftarrow W \cdot \frac{VH^T}{WHH^T} \quad , \quad H \leftarrow H \cdot \frac{W^T V}{W^T WH}$$

(6)

$$KL: \quad W \leftarrow W \cdot \frac{\frac{V}{WH} H^T}{1H^T} \quad , \quad H \leftarrow H \cdot \frac{W^T \frac{V}{WH}}{W^T 1}$$

(7)

$$IS: \quad W \leftarrow W \cdot \frac{\frac{V}{(WH)^2} H^T}{\frac{1}{WH} H^T} \quad , \quad H \leftarrow H \cdot \frac{W^T \frac{V}{(WH)^2}}{W^T \frac{1}{WH}}$$

(8)

Here, “1” denotes a matrix of ones with dimensions *n* x *m*, and “T” indicates matrix transposition. The dot operator “·” indicates element-wise multiplication. Remaining multiplications are computed by matrix multiplication. All division and exponential operations are performed elementwise. Derivations for these update rules can be found in [18]. Once ***W*** and ***H*** are optimized, a masking filter is applied for each source to extract them from the mixture. The generalized Wiener mask is commonly used and is calculated as follows:

$$M_i = \frac{W_i H_i}{WH}$$

(9)

Here, M_i is the mask for source *i*, W_i are the basis functions for source *i*, and H_i are the time-activation gains for source *i*. Multiplication is done by matrix multiplication and division is performed elementwise. Each masking filter can then be applied to ***V*** by performing elementwise multiplication before converting back to the time domain.

NMF can operate under three primary conditions: unsupervised, semi-supervised and fully supervised. In the unsupervised case, basis functions are generated blindly based on inherent structure in the mixture with no constraints based on prior information about the sources.

Complex sources such as speech or noise require multiple basis functions to represent them [19], and unsupervised separation provides no way to control which basis functions correspond to a specific source, making separation of complex sources difficult [20]. Supervised separation, on the other hand, uses isolated training information to develop a set of basis functions for the sources within a mixture prior to separation. Fully supervised separation provides training information for all sources within the mixture, while semi-supervised separation provides training information for some but not all sources. These basis function sets can then be concatenated together to create a trained $\boldsymbol{W}$ matrix for the entire mixture. The trained $\boldsymbol{W}$ is then left unchanged during the separation algorithm while the activation gains are updated using the multiplicative updates. Basis functions for any untrained sources are initialized randomly and are updated during algorithm convergence. Supervised training allows the number of basis functions for complex sources to be controlled in a way that leads to straightforward separation.

2.2 Objective Parameters

Considerable research has been conducted on the perceptual relevance of objective measurements for assessing separation performance [21 - 25]. Two predominant objective metric toolboxes for evaluating source separation performance are the Blind Source Separation Evaluation (BSS_Eval) toolbox [26] and the Perceptual Evaluation method for Audio Source Separation (PEASS) toolbox [21]. BSS_Eval measures performance using Signal-to-Noise Ratio (SNR)-based energy ratios to evaluate different types of separation quality degradation such as Source-to-Distortion Ratio (SDR), Source-to-Interference Ratio (SIR), Source-to-Artifacts Ratio (SAR), and source Image-to-Spatial distortion Ratio (ISR). PEASS uses an auditory perception model to generate perceptually based metrics including Overall Perceptual Score (OPS), Interference-related Perceptual Score (IPS), Target-related Perceptual Score, and Artifact related Perceptual Score (APS). Correlation tests between objective metrics and subjective ratings thus far have been inconclusive. Cano, FitzGerald, and Brandenburg [22] compared objective measures from both BSS_Eval and PEASS toolboxes with subjective ratings of separated music signals using two separation algorithms. Subjective results were obtained from 4 separate Multiple Stimulus with Hidden Reference and Anchors (MUSHRA) tests in which participants were asked to rate stimuli based on overall quality, artifact distortions, interference from other sources, and target source distortion. It was found that none of the objective metrics in either toolbox provided significant correlations with subjective scores. Additionally, a majority of the objective metrics had mild correlations. Ward, Wierstof, Mason, Grais, and Plumbley [24] conducted a similar study, comparing subjective results obtained from two separate MUSHRA tests against BSS_Eval and PEASS metrics. Stimuli were created from 23 different separation algorithms. Participants were asked to rate stimuli based on sound quality relating to the presence of artifacts and distortions, and on the interference relating to loudness of non-target instruments. APS was found to have strongest correlation with sound-quality while SIR had the strongest correlation for interference. The remaining metrics, however, exhibited a wide range of correlation. These inconsistencies indicate that current objective metrics do not sufficiently predict subjective separation performance and are therefore insufficient in determining perceptual separation quality.

2.3 STFT Window Duration and Cost Function

While research has been limited, previous studies have hinted at the STFT window durations effect on NMF performance. Smaragdis [14] found that increasing the STFT window duration from 8 ms to 64 ms using a sampling rate of 16 kHz can improve objective separation performance by up to 2.5dB, with a slight dip in performance above 64 ms. However, metrics used to evaluate performance are not standard in source separation literature and therefore do not compare directly to similar studies. Miller, Tarr, and Bulla [27] found that altering the STFT window duration used in NMF source separation produced significant detectable differences in a speech enhancement task using an ABX methodology. Window durations examined were 11.6 ms, 23.2 ms, 46.4 ms, and 92.9 ms at a sampling rate of 44.1 kHz. Shorter window durations were found to contain high levels of audible digital artifacts, while longer durations caused smearing of high frequency content. Further studies aimed to find ideal window duration lengths for optimal speech intelligibility and quality in speech synthesis applications [28, 29]. Ideal window durations were found between 15-64 ms at a sampling rate of 16 kHz, although these studies only used the magnitude component with random phase information for speech synthesis. These studies suggest that window duration may affect the intelligibility of separated speech when using NMF.

An objective analysis of NMF parameters including cost function was conducted in [30] to determine optimal parameters for separation of two simultaneous talkers. Sixteen combinations of overlapping male and female speech were tested. Bregman Divergence cost functions were evaluated using BSS_Eval metrics averaged over all talker combinations. It was found that the IS divergence had the best performance, although a subjective evaluation was not performed to verify results. Févotte, Bertin, and Durrieu [31] compared EU, KL, and IS divergences for pitch estimation of a short piano excerpt. IS basis functions were found to more accurately represent the pitches of individual notes, transient events, and piano pedal releases. It was also noted that the IS divergence is the most computationally expensive cost function. This simple application was intended to observe how basis functions evolve with different cost functions and shed light onto practical differences between them. However, the piano excerpt used had no interfering sources and the effect of cost function in a source separation application was not confirmed. Masking filters based on the KL and IS cost functions were proposed in [32] as an alternative to the generalized Wiener filter mask typically used. These filters were applied to three different source separation algorithms based on NMF. All algorithms were iteratively updated using both KL and IS cost functions and sources were separated using the KL, IS, and Wiener masks. PEASS metrics revealed that none of the masks were able to outperform the others across all test conditions, though the proposed masks did increase performance in some cases. These unclear results are further obscured by the fact that none of the existing evaluation metrics are formally related to the cost function [33].

2.4 Basis Functions

Another important parameter in NMF audio source separation is the number of basis functions used to define each source. Previous studies have indicated that having too few basis functions results in poor approximation and generalization of sources, while too many basis vectors can lead to overfitting [30]. Smaragdis [14] observed a tradeoff between separation performance and number of basis functions in a multiple talker separation task. More basis functions led to higher suppression of undesired signals while fewer basis functions produced a decrease in residual noise. It was also noted that anywhere from 100-500 basis functions will provide a good estimate for speech signals. Mohammed and Tashev [34] conducted an empirical study analyzing basis functions in the range of 10 to 2,000. Using the Perceptual Evaluation of Speech Quality (PESQ) methodology to evaluate separation performance, it was found that more basis functions produced higher quality speech approximations with decreased variance. However, these values were taken by representing clean speech signals using NMF dictionaries, and not from speech separated from a noisy mixture. Furthermore, in order to maintain a compressed version of $\boldsymbol{V}$as described in Section 2.1, the number of basis functions is restricted by the number of time frames and frequency bins in the spectrogram representation. In cases of signals with short duration or spectrograms created with short STFT window durations, 2,000 or more basis functions may not be practical.

2.5 Use of Full Bandwidth Stimuli

Speech enhancement studies typically employ stimuli sampled at 20 kHz or less [3, 14, 30, 35, 36], rejecting the upper octave of the audible frequency spectrum. Not only do these lower sampling rates reduce the amount of time-frequency data needed to approximate, but they also restrict subjective audio quality by band limiting the signal. Monson, Hunter, Lotto, and Story [37] presented a comprehensive review of studies investigating the effects of high frequency energy (HFE) on speech quality and intelligibility. These studies suggest that the inclusion of HFE between 8 and 22 kHz can improve both quality and intelligibility of speech, and that subjective evaluations of speech separation algorithms should use full bandwidth stimuli. Additionally, one of the few studies to find significant correlation between objective measures and subjective results used stimuli sampled at 44.1 kHz [25], which raises a question if full bandwidth audio increases the reliability of objective metrics to predict subjective results.

3.0 METHODS

Two experiments were conducted to determine the effects of NMF algorithm parameters on perceptual performance of source separation. Experiment 1 examined the effects of STFT window duration on speech intelligibility. Experiment 2 investigated how the cost function used affected the perceived level of residual noise after NMF speech enhancement. The following section describes the methodology for these experiments, including the participants, stimuli, procedures and experimental design.

3.1 Experiment 1 – Effect of STFT Window Duration on Speech Intelligibility

3.1.1 Participants

Eleven graduate audio engineering students participated in this study (eight male and three female). Participants previously received critical listening training for at least one semester, and all had self-reported normal hearing. All participants were native English speakers.

3.1.2 Stimuli

Test stimuli for this experiment consisted of speech extracted from a mixture of speech and additive noise using NMF [19] with different STFT window durations. Window durations used were 11.6 ms, 23.2 ms, 46.4 ms, and 92.9 ms (corresponding to 512, 1024, 2048, and 4096 samples at a 44.1 kHz sampling rate respectively). Two noise sources were examined: speech shaped noise (SSN) generated using talkers from an Institute of Electrical and Electronic Engineers (IEEE) speech corpus [38] and a conversation "babble" from the Connected Speech Test (CST) [39]. MATLAB code used to generate SSN can be found in Appendix A. A total of 80 unique sentences (none which were used to generate SSN) randomly selected from the same IEEE speech corpus [38] spoken by four American-English talkers (two male talkers and two female talkers) were included in the test set. All sentences contained five keywords. Forty of these sentences used male talkers while the other 40 sentences used female talkers. Twenty sentences were used for each individual talker. For each talker, 10 of the sentences were mixed with SSN at a $\mathrm{SNR}_{\mathrm{RMS}}$ of -6dB, and the remaining 10 were mixed with babble noise at the same SNR. Mixtures were then normalized to digital full scale. For each talker/noise source pairing, eight of the 10 sentences were processed by NMF using one of the window durations under test. Each window duration was used on two of the processed sentences in each group, such that each talker/noise source pairing had two sentences processed using a window duration of 11.6 ms, two using 23.2 ms, two using 46.4 ms, and two using 92.9 ms. The remaining two sentences in each talker/noise source pairing were left unprocessed in order to establish a baseline intelligibility level to compare each window duration against. In total, across all talker and noise source pairings, there were 16 sentences for each window duration, as well as 16 unprocessed sentences.

Ten additional unique sentences for training were included at the beginning of each test session. The training set included a combination of male and female speech mixed with SSN and babble equivalently to the test set described above. Two training sentences were processed with NMF using each of the test window durations while the remaining two training sentences were left unprocessed.

Additional NMF parameters used included fully supervised training, which most closely resembles the current state of the art in audio source separation [40]. The training used 30 sentences from both talkers of the target gender – none of which were included in the test or training stimuli set – to train a set of basis functions for the target speech. Fifteen sentences from each talker of the same gender as the target speech were used for training. The associated noise source from each sentence (SSN or babble) was also used to train a set of basis functions for the noise in each mixture. The trained speech and noise basis functions were then concatenated together to form the trained $\boldsymbol{W}$ matrix used during NMF. Two hundred basis functions in total were used, 100 for the speech and 100 for the noise. The number of basis functions was chosen to provide sufficient speech quality with relatively low computational cost. An IS cost function, which is known in the speech community for good perceptual properties [31], was used to perform iterative updates, and 100 update iterations were performed to create the time-activation approximation $\boldsymbol{H}$. STFT representations were created using 50% overlapping time frames and a Hann window filter. All speech and noise files were 16-bit/44.1 kHz resolution.

3.1.3 Experimental Design

Tests were conducted *via* a MacBook Pro laptop computer running a MATLAB version 9.7 [41] graphical user interface (GUI) and Shure SRH840 professional quality headphones in an acoustically controlled environment. A screenshot of the test GUI is shown in Figure 3. For each trial, participants listened to a noisy sentence and entered the sentence to the best of their ability in the "Response" text box. The independent variables for this experiment were the window duration used in NMF separation, the noise source (SSN or babble), and the different talkers and their gender (male or female). The dependent variable was the percent of correctly identified keywords in each sentence by the participant.

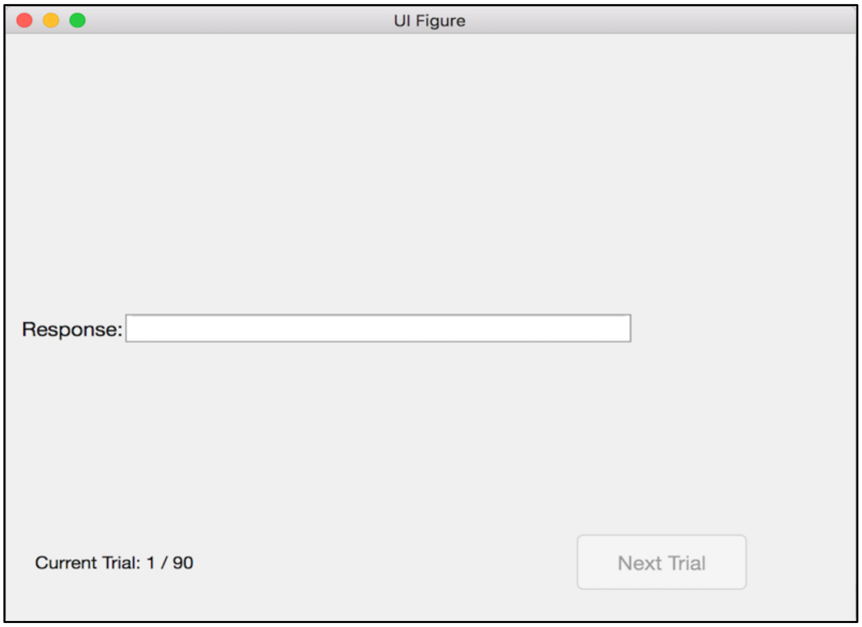


Figure 3. Experiment 1 Test GUI used during experiment.

3.1.4 Procedure

The experiment began with participants listening to a sample mixture of noisy speech in order to adjust their headphone sound level; once the participant set a headphone level, they were not allowed to change it during the experiment. The sample mixture was not included in the training or test stimuli set. Once they were satisfied with their headphone level, the experiment began. For each trial, a two-second pause occurred before a single automated playback of the test stimulus. Participants then typed the sentence they thought they heard into the text box on the GUI before proceeding to the next trial. Each trial used a unique sentence that only occurred once during the experiment to prevent participants from "learning" the sentences. The first 10 trials consisted of the training stimuli described above; the inclusion of training stimuli was unknown to the participants. Training was included to allow participants the opportunity to get acclimated with the test procedure and prevent erroneous errors caused by participants not being ready at the beginning of the experiment. A total of 80 test trials were conducted, each trial using one of sentences from the test set described above. Including the training, each participant completed 90 trials with a minimum 1-minute break after every 30 trials. The experiment recommenced after each break once the participant was ready to proceed by pressing a "Proceed" button. Testing was completed over 1 test session which lasted approximately 30 minutes.

3.2 Experiment 2 – Effect of Cost Function on Residual Noise Level

3.2.1 Participants

Participants included the 11 subjects from Experiment 1 plus an additional three participants (three female). While both experiments were evaluating NMF audio source separation, they employed unrelated psychophysical tests and were considered separate stand-alone experiments where no learning effect was expected between the two tests. All 14 participants were trained in the same manner. However, no post-hoc tests were conducted to confirm consistency between the original participant group and the three new participants.

3.2.2 Stimuli

Stimuli for this experiment were created by processing speech with additive noise through NMF using different cost functions. Two different talkers from an IEEE speech corpus [38] (one male and one female) were mixed with two noise sources (SSN and babble) at an SNR_{RMS} level of 0 dB and then normalized to digital full scale for a total of four mixtures. Sentences used in this experiment did not appear in the stimuli set from Experiment 1. SSN and babble noise were the same as described in Section 3.1.2. Each mixture was then processed through NMF using three different cost functions – EU, KL, and IS – for a total of 12 test stimuli. Fully supervised training as described in Section 3.1.2 was implemented. One thousand basis functions in total were used, 500 for the speech and 500 for the noise. Since this experiment used a small number of stimuli compared to Experiment 1, computational cost was not as much of a concern and therefore a greater number of basis functions were used to provide suitable speech quality. Additional NMF parameters consisted of a window duration of 46.4 ms, 50% overlapping frames with a Hann window filter, and 100 update iterations. All speech and noise files were 16-bit/44.1 kHz resolution. A webpage link to the stimuli used for this experiment can be found in Appendix B.

3.2.3 Experimental Design

Tests were conducted in two phases: a training phase and an evaluation phase. A screenshot of the training GUI is shown in Figure 4 and the evaluation GUI is shown in Figure 5. Tests were conducted *via* a MacBook Pro laptop computer running a MATLAB version 9.7 GUI [41] and Shure SRH840 professional quality headphones in an acoustically controlled environment. The independent variables for this experiment were the cost function used in NMF (EU, KL, or IS), the noise source (SSN or babble), and the gender of the talker (male or female). The dependent variable was the subjective residual noise score assigned to the processed test stimuli on each trial. Participants completed one testing session which lasted approximately 10 minutes.

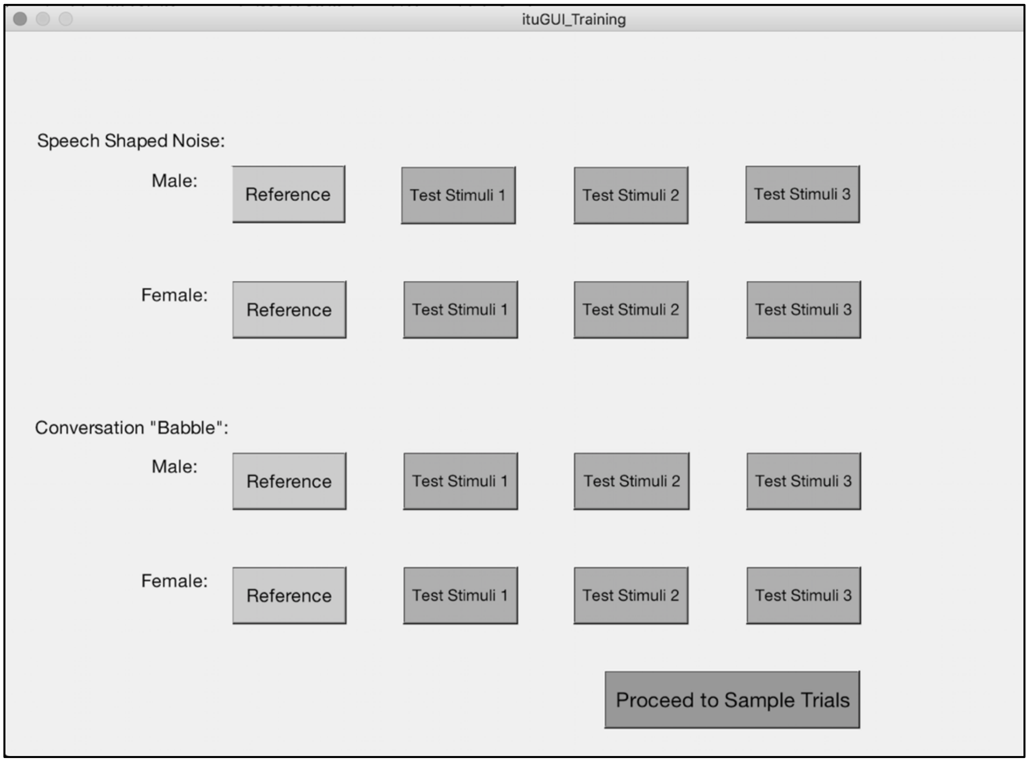
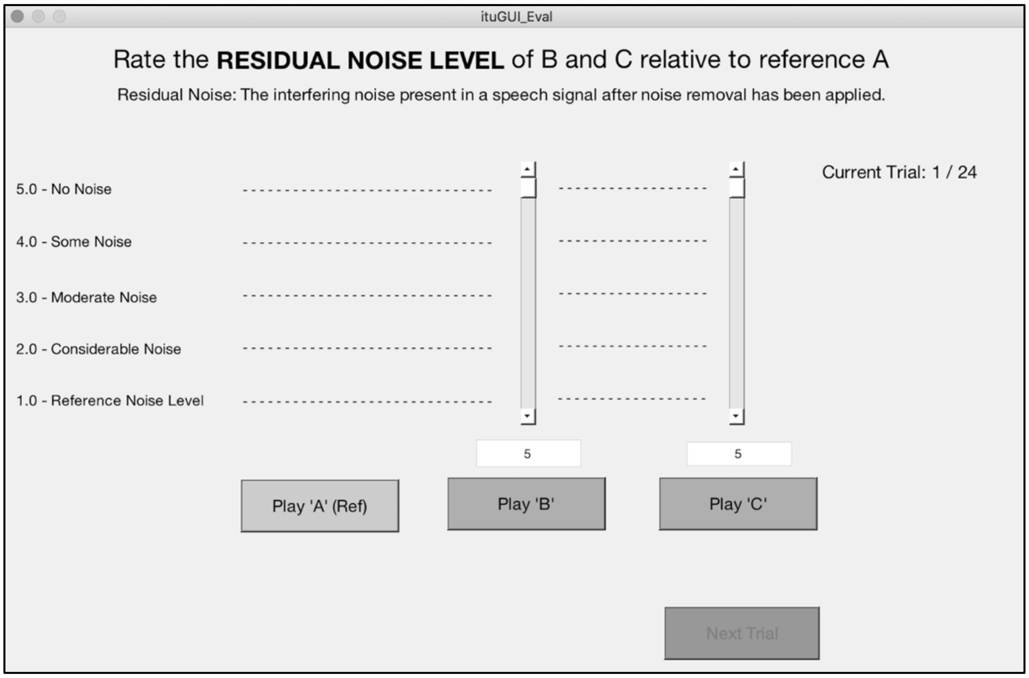


Figure 4. Experiment 2 training phase GUI.


Rate the **RESIDUAL NOISE LEVEL** of B and C relative to reference A

Residual Noise: The interfering noise present in a speech signal after noise removal has been applied.

Current Trial: 1 / 24

Figure 5. Experiment 2 evaluation phase GUI.

3.2.4 Procedures

The training phase aimed to familiarize participants with the stimuli in the experiment, the required task, and the test GUI. Training began by presenting participants with all of the test stimuli as well as their reference mixtures. Participants were allowed to play the stimuli as many times as they wanted to learn differences in the residual noise levels between stimuli and their references. When pressed, each playback button displayed a label at the top of the GUI indicating if that stimulus had "Low Residual Noise" or "High Residual Noise". These labels provided participants with general guidelines of what constituted residual noise and established what they should be listening for during the evaluation. It was emphasized to the participants that these were relative benchmarks used only to assist in understanding what constitutes residual noise, which was defined as "the interfering noise present in a speech signal after noise removal has been applied". Once the participant was satisfied, four sample trials were conducted under the guidance and supervision of the researcher. These sample trials ensured that the participant understood the purpose and functionality of the test. Any questions or misunderstandings by the participant were addressed by the researcher before proceeding to the evaluation phase.

The evaluation phase was based on the ITU-R BS.1116-3 standard subjective evaluation procedure [42]. For each trial, participants were presented with 3 stimuli labeled "A", "B", and "C". "A" was always the reference mixture while "B" and "C" were randomly a hidden reference and one of the test stimuli. Participants were asked to rate the level of residual noise in "B" and "C" relative to "A" based on the values shown in Table 1. Participants correctly identified the hidden reference by scoring it 1.0 (Reference Level Noise). The test stimuli were scored based on the perceived level of residual noise present, with higher scores corresponding to lower levels of perceived residual noise. Degradations to speech quality or intelligibility that may occur during

NMF processing were not to be considered when rating test stimuli. A total of 24 double-blind trials were conducted, with each test stimulus presented twice.

Table 1. Scale for residual noise level grading from ITU-R BS.1116-3 [42]

Residual Noise Present	Score
No Residual Noise	5.0
Some Residual Noise	4.0
Moderate Residual Noise	3.0
Considerable Residual Noise	2.0
Reference Level Noise	1.0

3.2.5 Objective Measurements

Additionally, BSS_Eval and PEASS measurements were taken to objectively measure quality of target speech. In particular, the SIR from BSS_Eval and the IPS from PEASS were taken, since these particular metrics are associated with the level of interfering sources present in the extracted target signal. Since the task of this experiment involved participants rating the level of residual noise (interference) in speech (target signal), the interference-related SIR and IPS metrics most closely relate to the subjective results. SDR and OPS metrics were also included as an overall performance comparison with other source separation literature. SDR and OPS are associated with the overall quality of the extracted target signal and are commonly used in literature such as [21] and [30] to indicate general separation performance.

4.0 RESULTS

4.1 Experiment 1

Figure 6 shows a histogram of the data distribution from Experiment 1. It can be seen that the data does not conform to a normal distribution. Consequently, non-parametric Friedman Analysis of Variance (ANOVA) and Wilcoxon signed-rank test were used to analyze the data from Experiment 1.

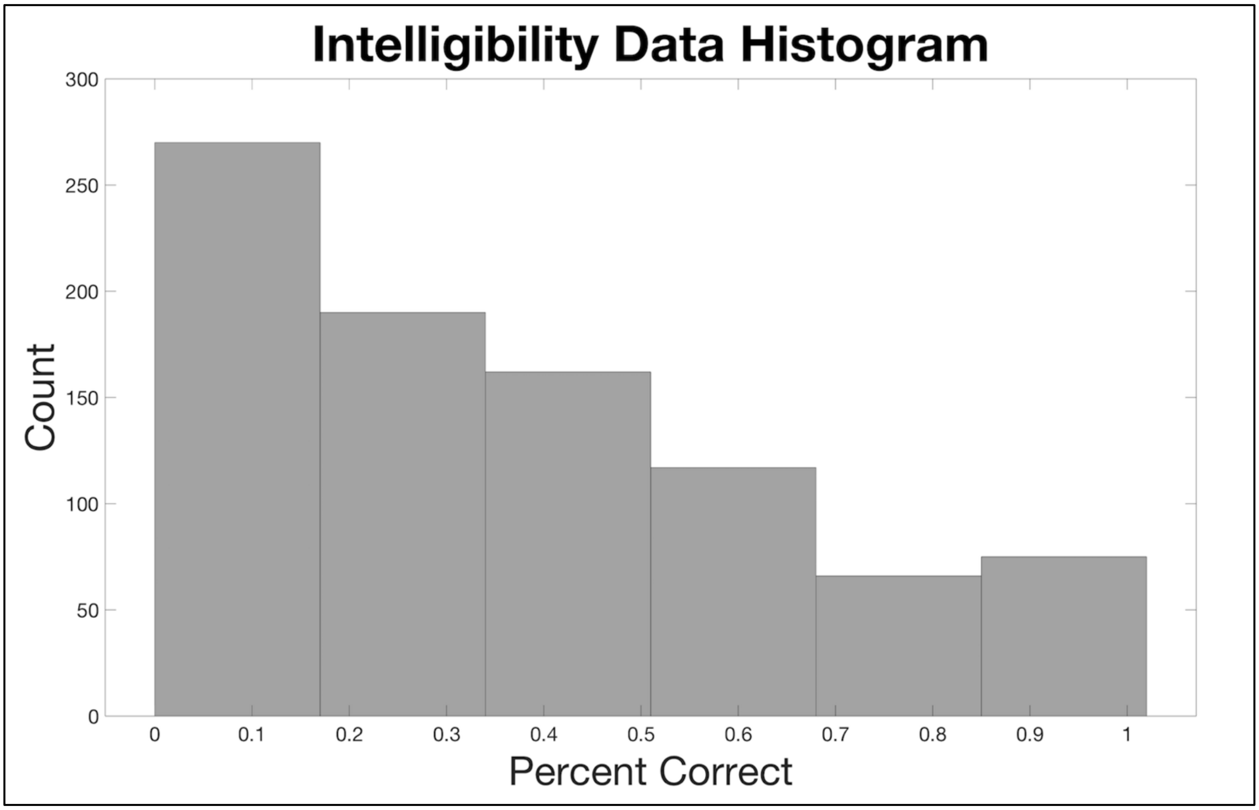


Figure 6. Histogram of Experiment 1 data.

Window duration was found to have a significant effect on the percent of correctly identified keywords, referred to herein as "intelligibility score" ($\chi^2(4,700) = 74.0$, $p < .001$, $\eta_p^2 = .11$). Figure 7 shows a comparison of mean intelligibility scores for the window durations evaluated. The intelligibility scores were significantly higher with longer window durations when using a

Bonferroni-adjusted alpha level of .005 (.05 / 10) due to multiple comparisons (unprocessed speech vs. 46.4 ms (Z = 3.63, p < .001, r = .10); unprocessed speech vs. 92.9 ms (Z = 5.77, p < .001, r = .06); 11.6 ms vs. 46.4 ms (Z = 4.38, p < .001, r = .08); 11.6 ms vs. 92.9 ms (Z = 5.73, p < .001, r = .08); 23.2 ms vs. 92.9 ms (Z = 4.48, p < .001, r = .09)). Differences between unprocessed speech and 11.6 ms (Z = 0.46, p = .646, r =.13), unprocessed speech and 23.2 ms (Z = 1.14, p = .255, r = .117), 11.6 ms and 23.2 ms (Z = 1.73, p = .084, r = .11), 23.2 ms and 46.4 ms (Z = 2.60, p = .009, r = .10) and 46.4 ms and 92.9 ms (Z = 2.23, p = .026, r = .13) were not significant.

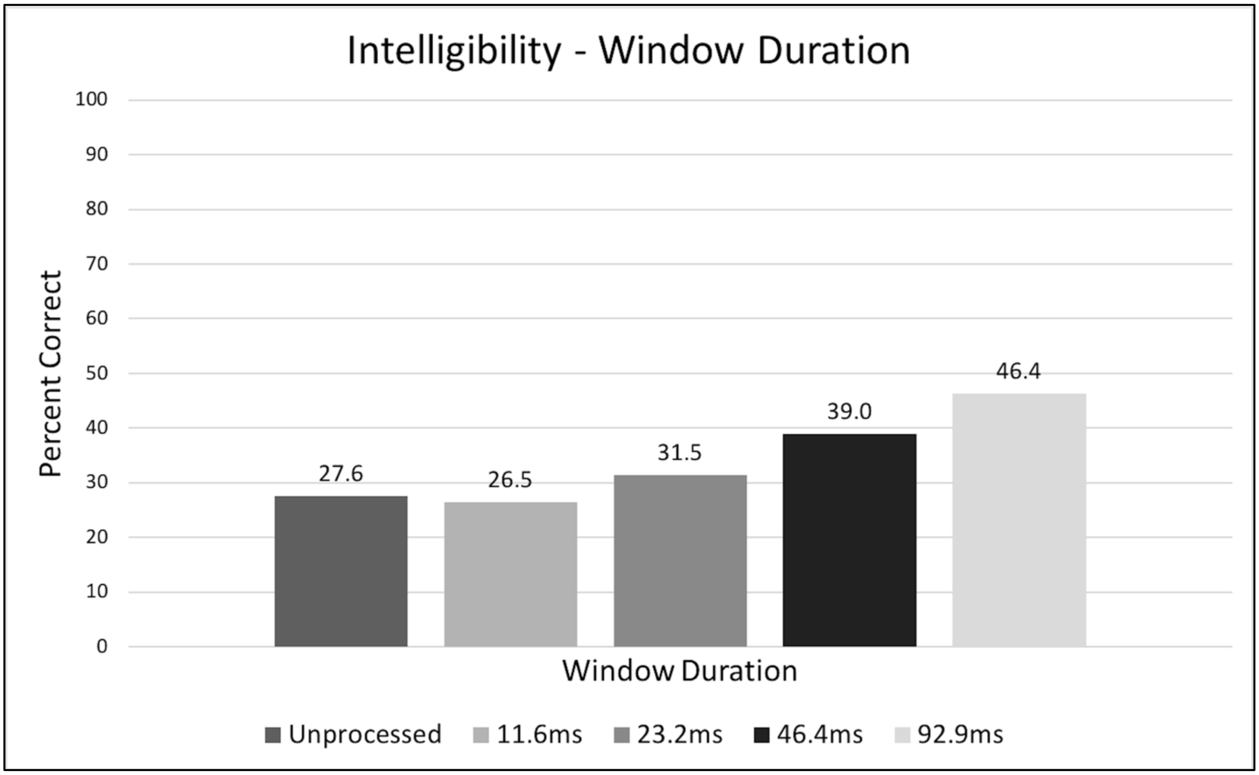


Figure 7. Effect of window duration on intelligibility scores.

Analyses were also performed to investigate the effects of the noise source on intelligibility scores. The type of interfering noise (SSN or babble) was found to have a significant effect on intelligibility scores (χ^2(1,439) = 90.1, p < .001, η_p^2 = .20). Figure 8 shows a comparison of mean intelligibility scores between noise sources across window durations. SSN mixtures showed steady improvement in intelligibility scores as window durations increased, while intelligibility scores of babble mixtures remained relatively consistent for shorter window durations with substantial increases for longer window durations. Babble sentences also had a higher intelligibility scores than SSN sentences across all window durations.

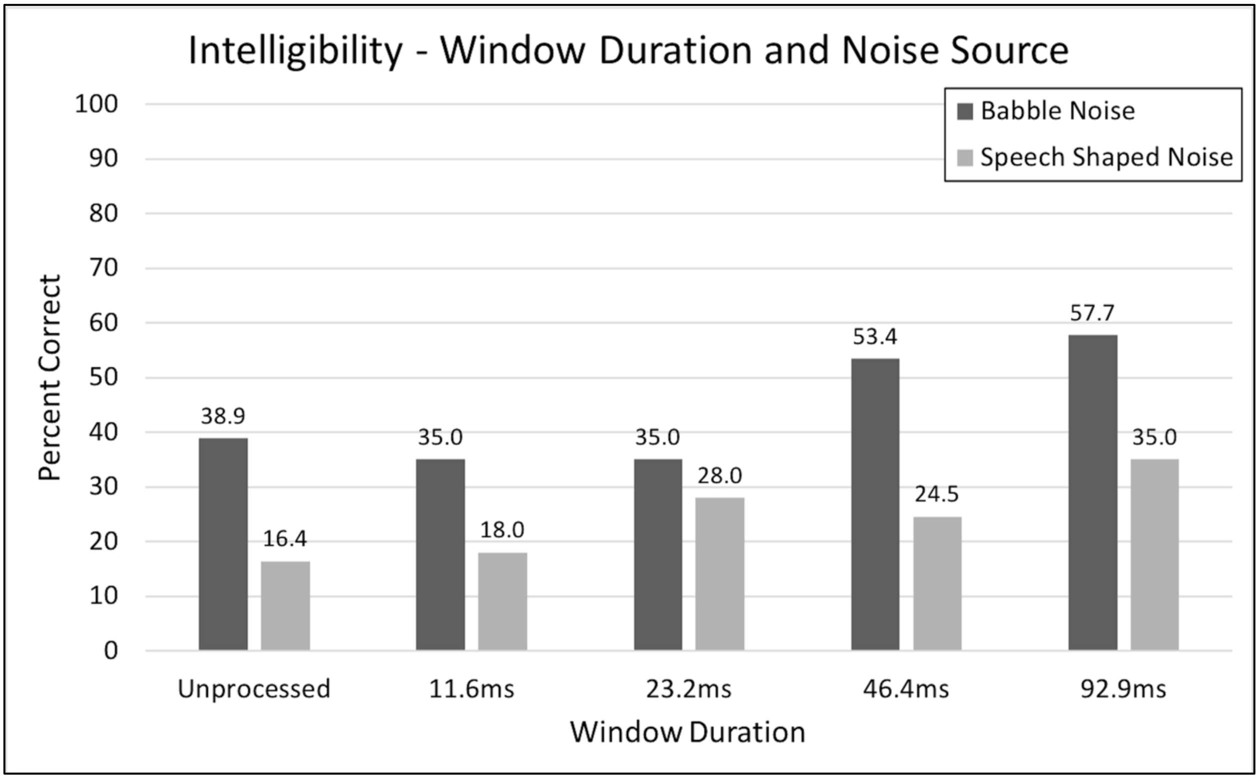


Figure 8. Effect of window duration and noise source on intelligibility scores.

Talker type also had a significant effect on intelligibility scores ($\chi^2(3,657) = 83.19$, $p < .001$, $\eta_p^2 = .13$). Figure 9 shows mean intelligibility scores of different talkers used in the experiment. Here, it can be seen that male talkers had significantly lower intelligibility scores than female talkers when using a Bonferroni-adjusted alpha level of .0083 (.05 / 6) due to multiple comparisons (Male 1 vs. Female 1 ($Z = 6.10$, $p < .0083$, $r = .13$); Male 1 vs. Female 2 ($Z = 7.80$, $p < .0083$, $r = .11$); Male 2 vs. Female 1 ($Z = 4.57$, $p < .0083$, $r = .20$); Male 2 vs. Female 2 ($Z = 7.21$, $p < .0083$, $r =$

.11). Differences between Female 1 and Female 2 were not significant (*Z* = 2.15, *p* = .031, *r* = .28). Male 1 scored the lowest mean intelligibility and Female 2 had the highest.

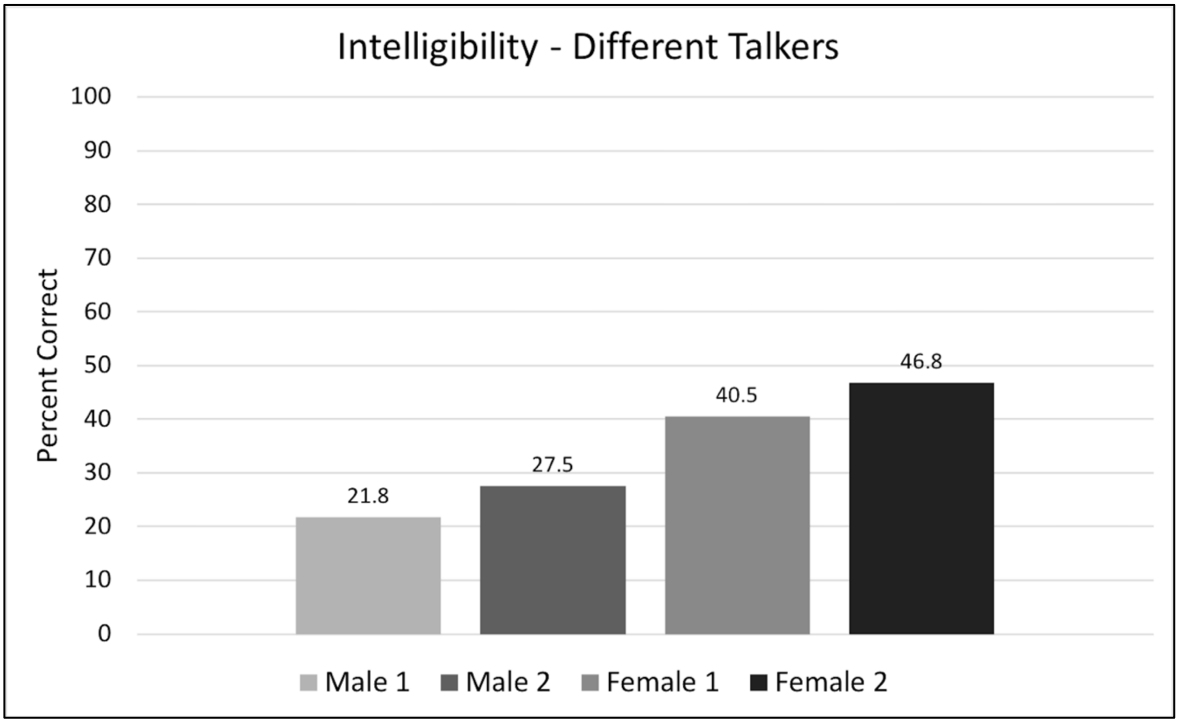


Figure 9. Effect of talker on intelligibility scores.

Figure 10 shows talker intelligibility scores across window duration. Male 1 was the most difficult to understand in the unprocessed case but achieved comparable intelligibility scores to other talkers with window durations above 23.2 ms. Female 2 had the highest intelligibility for all window durations except 11.6 ms, while Female 1 did not appear to have a clear relationship between window duration and intelligibility.

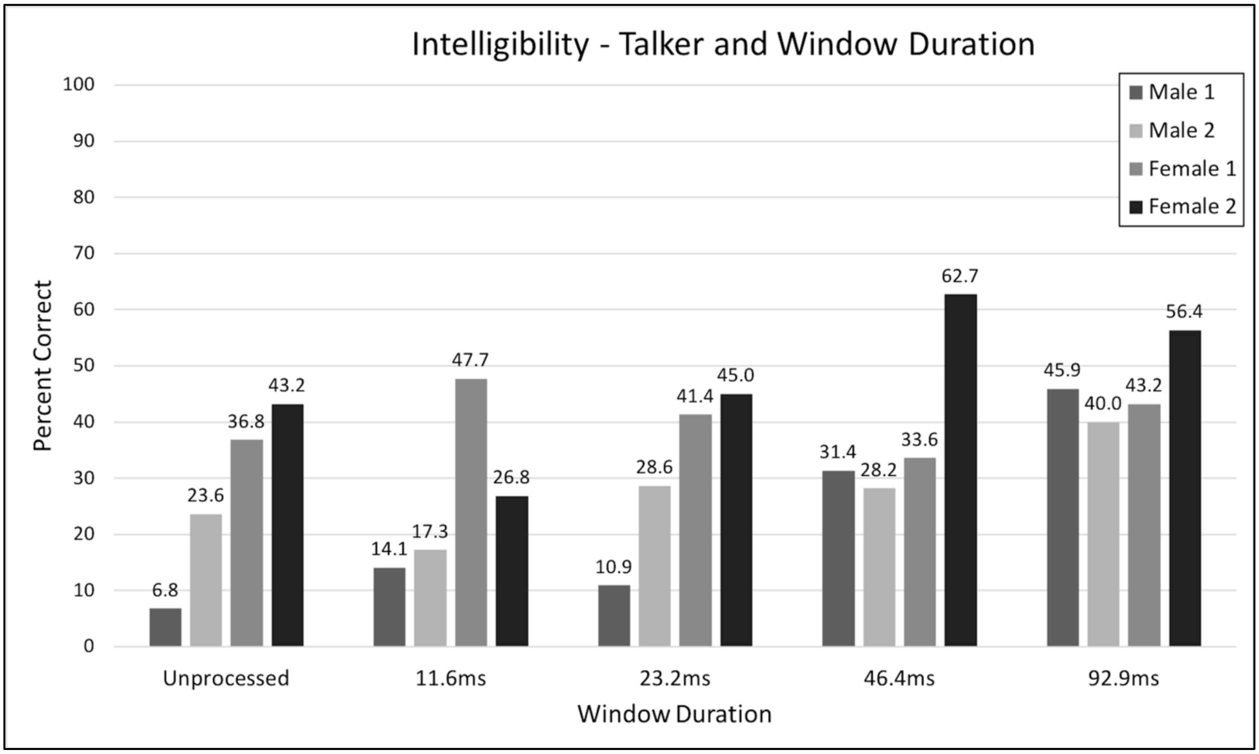


Figure 10. Effect of window duration and talker on intelligibility scores.

4.2 Experiment 2

Since this experiment used repeated trials, there was some inherent variability in participant responses. To account for any unreliable participants whose responses had higher than average variability amongst all participants, the difference in subjective residual noise score between repeated trials was calculated for all participants over all 12 test stimuli. Mean and standard deviations were then calculated for all participant differences across each of the 12 test stimuli. A given participant's responses were deemed to have high variability if the difference between their scores for a given stimulus was greater than two standard deviations of the differences for that stimulus. It was found that two participants (P10 and P13) had high variability in their responses for 50% or more of the test stimuli, while the remaining participants had high variability on 33% or less of the test stimuli. Therefore, the results from participants P10 and P13 were deemed unreliable and their data were excluded from the following data analysis. Table 2 summarizes differences between individual participant results. Differences with high variability as defined above are denoted with an underscore, while participants P10 and P13 are highlighted in **bold**.

Table 2. Difference in scores between repeated trials for all participants (columns) and test conditions (rows).

Stimulus	P01	P02	P03	P04	P05	P06	P07	P08	P09	**P10**	P11	P12	**P13**	P14
1	0.8	0.6	0.2	0.0	0.3	0.1	0.1	0.4	0.9	**0.4**	1.8	0.1	**1.3**	0.3
2	0.3	1.0	0.5	0.1	0.6	1.3	0.4	0.7	0.1	**1.2**	0.3	0.4	**0.9**	0.5
3	0.7	0.1	0.3	0.0	0.6	0.5	0.5	0.5	0.1	**0.2**	0.5	0.3	**1.7**	0.3
4	0.8	0.3	0.3	0.9	0.5	1.2	1.0	0.1	1.3	**0.9**	0.4	0.5	**0.8**	0.4
5	0.3	0.0	0.9	0.0	0.2	0.0	0.4	0.1	0.0	**2.1**	0.0	0.0	**0.2**	0.3
6	1.2	0.2	0.4	0.0	0.6	0.1	0.4	0.1	0.2	**0.8**	0.4	0.9	**0.1**	0.6
7	0.2	0.5	0.2	1.9	0.2	0.5	0.3	0.7	0.7	**0.6**	0.6	0.0	**0.3**	0.9
8	0.4	0.0	0.9	0.1	0.4	0.9	0.9	0.5	0.9	**0.3**	0.7	0.6	**2.6**	0.1
9	0.3	1.4	0.3	0.0	0.0	0.5	1.0	0.3	0.2	**0.7**	0.1	0.9	**0.7**	0.8
10	0.3	0.9	0.0	0.0	0.1	0.6	0.9	1.0	0.1	**1.1**	0.4	0.4	**0.8**	0.2
11	0.2	0.2	1.2	0.0	0.4	0.9	0.0	0.6	0.7	**0.0**	0.1	0.0	**1.2**	0.5
12	0.9	0.1	1.6	1.0	0.7	0.5	1.1	0.6	0.8	**1.3**	0.1	0.0	**1.1**	0.6

Figure 11 shows a histogram of the data distribution from Experiment 2. It can be seen that the data generally follows a normal distribution. Therefore, parametric ANOVA and t-tests were used to analyze the data.

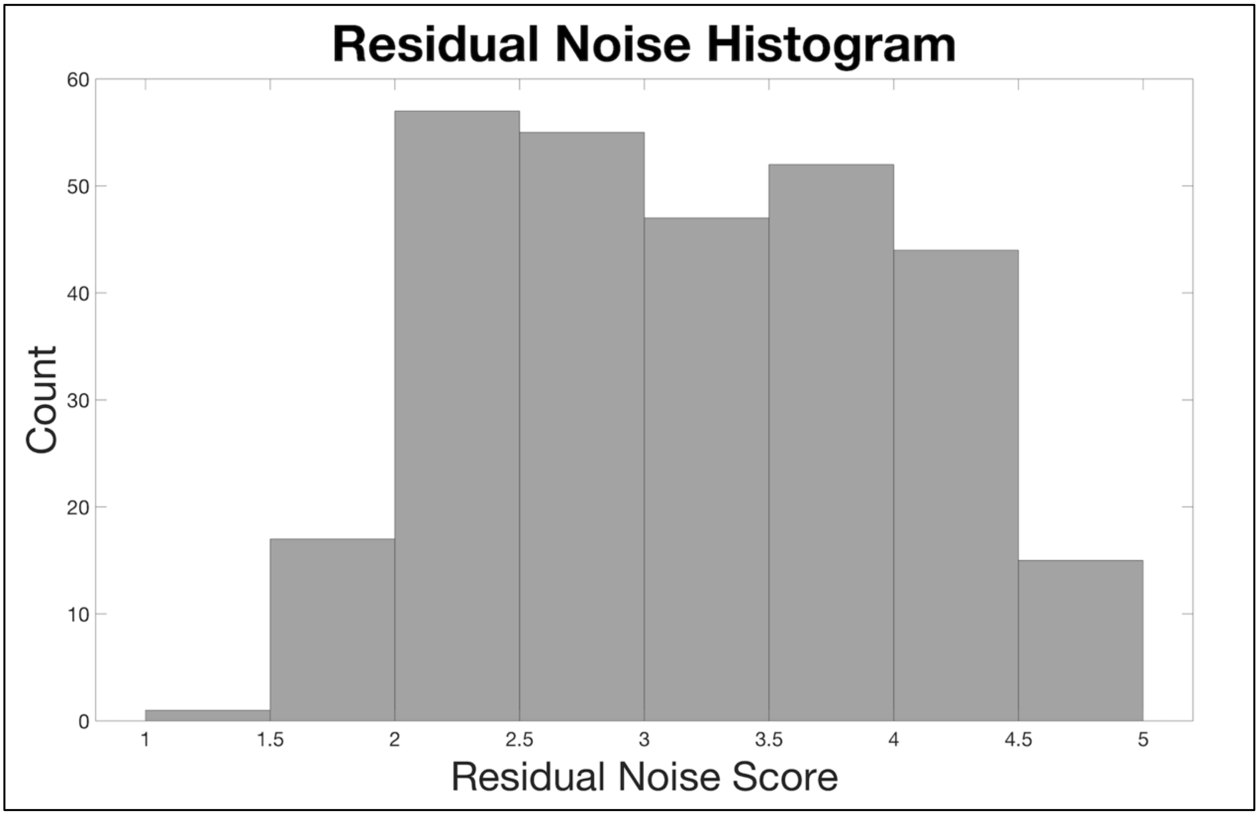


Figure 11. Histogram of Experiment 2 data.

Cost function had a significant effect on residual noise scores ($F(2,1) = 24.9$, $p < .001$, $\eta_p^2 = .15$). Figure 12 shows mean residual noise scores for different cost functions used in the experiment. Here, it can be seen that the IS divergence had significantly higher residual noise scores than the EU or KL divergences when using a Bonferroni-adjusted alpha of .017 (.05 / 3) due to multiple comparisons (IS vs. EU ($t(95) = 7.08$, $p < .017$, $d = .78$); IS vs. KL($t(95) = 6.35$, $p < .017$, $d = .71$)). Differences between EU and KL were not significant ($t(95) = 0.87$, $p = .386$, $d = .08$).

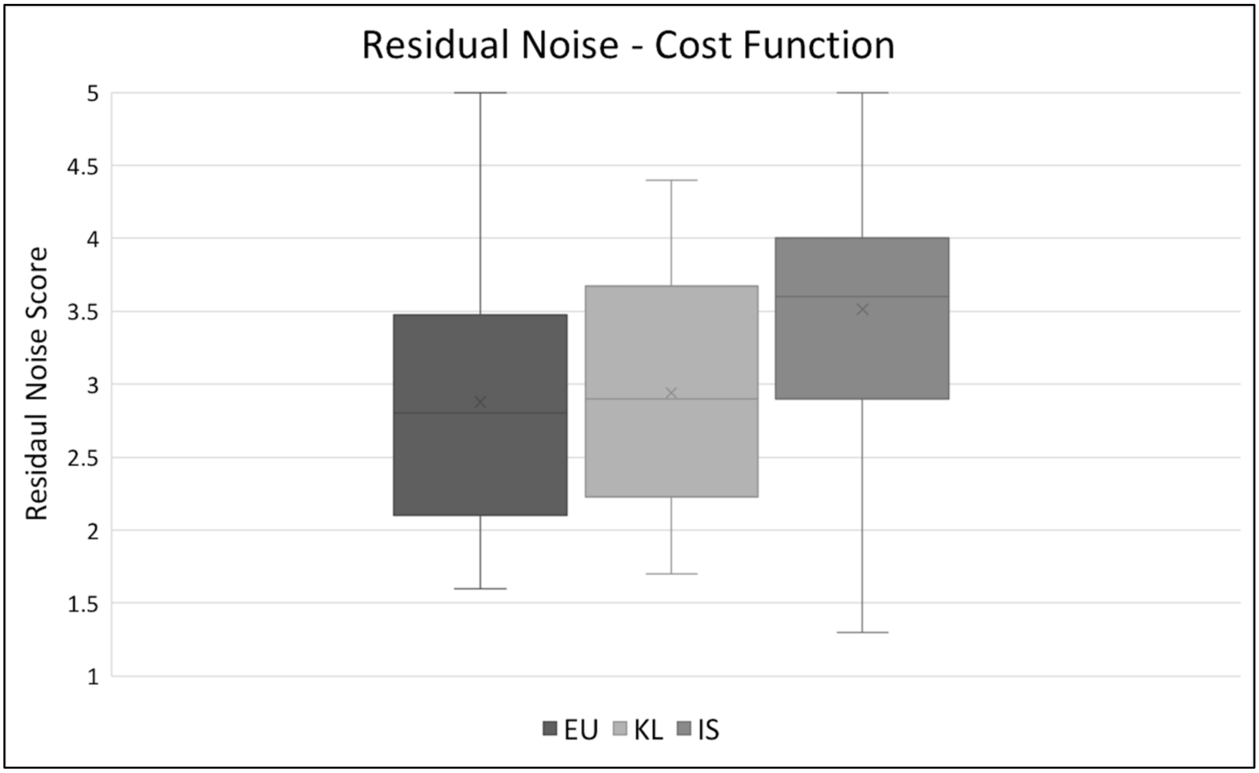


Figure 12. Effect of cost function on subjective residual noise score.

The type of interfering noise also had a significant effect on subjective residual noise scores ($F(2,1) = 100.3$, $p < .001$, $\eta_p^2 = .26$). Figure 13 shows residual noise scores of noise type across divergence cost functions. Here, we can see that babble noise mixtures were given higher residual noise scores for all cost functions, with the interaction effect between cost function and noise source not significant ($F(2,1) = 2.57$, $p = .078$, $\eta_p^2 = .02$).

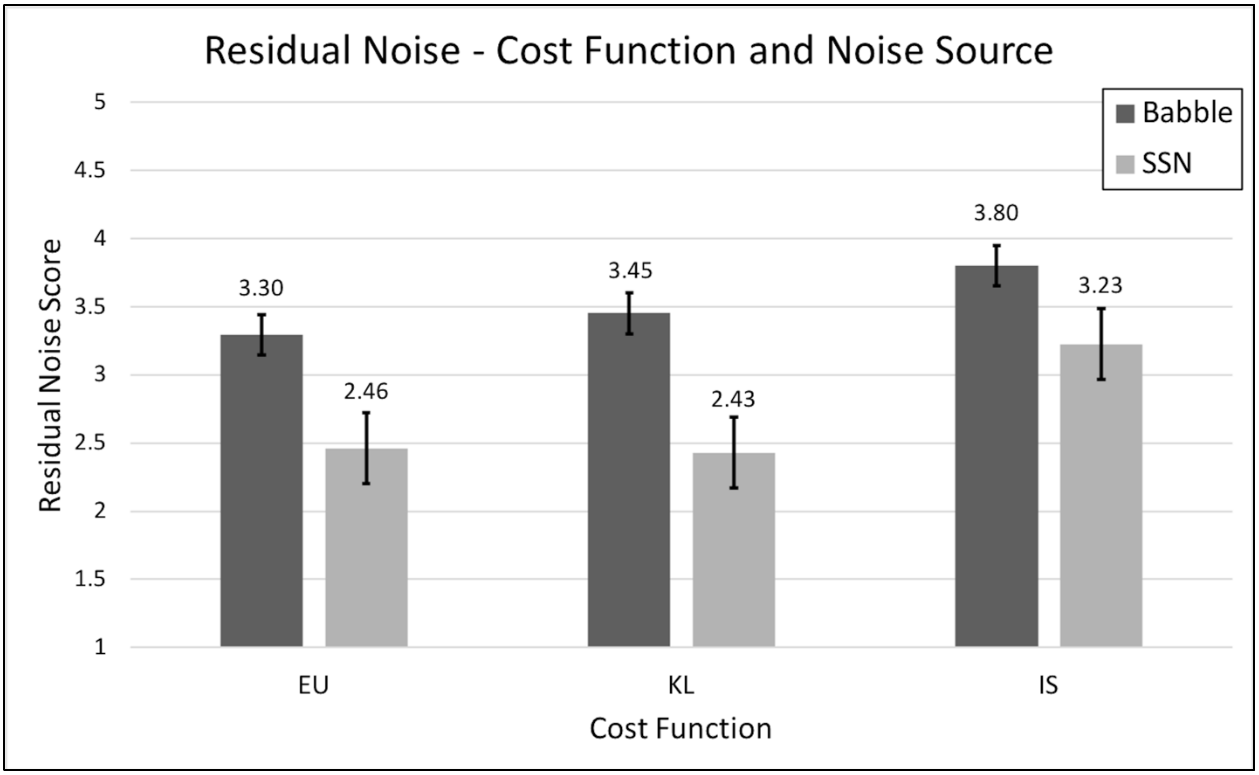


Figure 13. Effect of cost function and noise source on subjective residual noise score.

The effect of talker gender on residual noise scores was not significant ($F(2,1) = 2.90$, $p = .090$, $\eta_p^2 = .01$). Figure 14 shows talker residual noise scores across cost functions. Here, we can see that there is a significant interaction effect between cost function and talker ($F(2,1) = 6.86$, $p < .05$, $\eta_p^2 = .05$). Residual noise scores were higher for the Female talker using the EU and KL divergences, but the Male talker had higher residual noise scores when using the IS divergence.

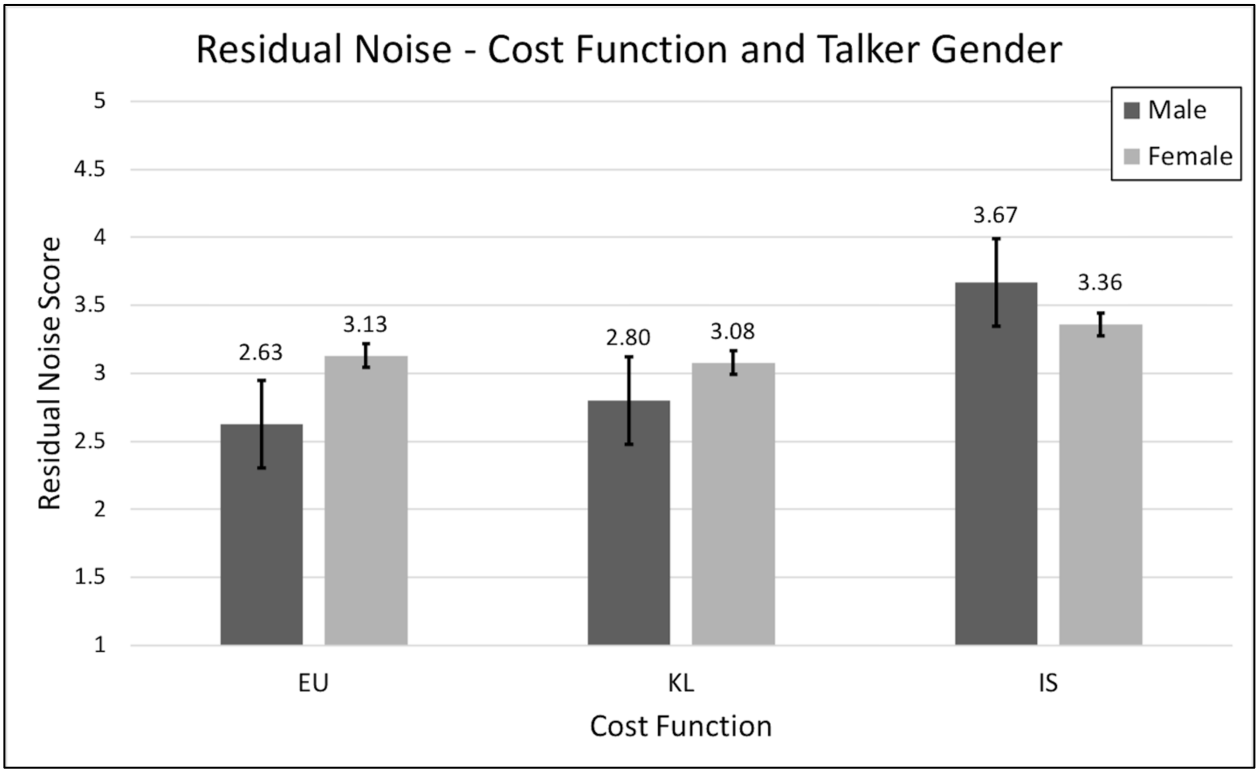


Figure 14. Effect of cost function and talker gender on subjective residual noise score.

Analyses were also performed to examine the interaction of cost function, talker, and noise source. The interaction effect of noise and talker was significant ($F(2,1) = 1.79$, $p = .042$, $\eta_p^2 = .02$), while the cost function, talker, noise interaction effect was not significant ($F(2,1) = 1.79$, $p = .169$, $\eta_p^2 = .01$). Figure 15 shows the residual noise scores of each talker-noise source pair across cost functions. For all cost functions, the SSN mixtures have the lowest residual noise scores while Male SSN mixtures have the lowest scores for EU and KL divergences. The residual noise scores for Male SSN with the IS divergence appear to be comparable to both babble noise mixtures using an IS divergence.

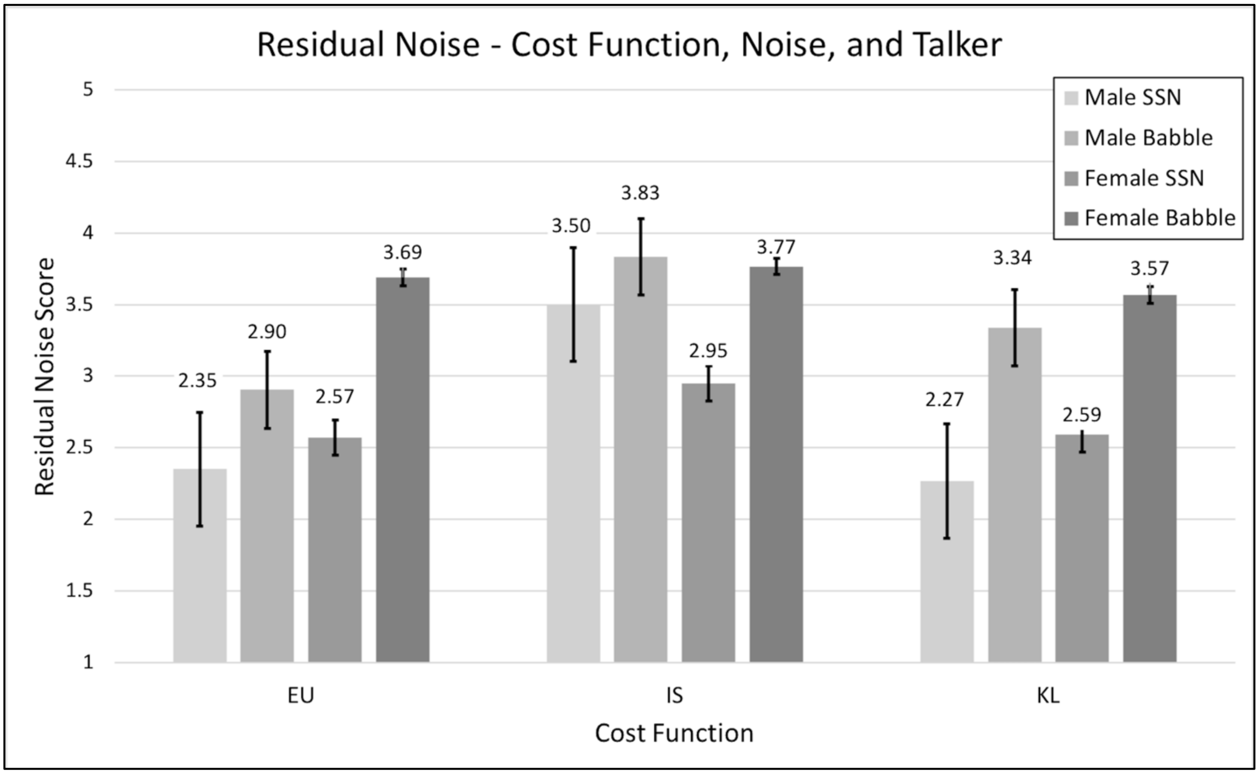


Figure 15. Effect of cost function, talker, and noise source on subjective residual noise score.

Objective measures from the BSS_Eval and PEASS toolboxes were additionally computed for the stimuli in this experiment. Table 3 lists the objective and subjective scores of all test stimuli used in the experiment. For all objective metrics, higher values indicate better performance. The highest values for each objective metric are indicated in **bold**. Mean subjective residual noise scores were included to perform correlation analysis between subjective and objective results.

Table 3. Comparison of BSS_Eval, PEASS, and mean Subjective scores for all stimuli

Stimuli	**BSS_Eval**		**PEASS**		Subjective Residual Noise Scores
	SDR (dB)	SIR (dB)	OPS (%)	IPS (%)	
Male SSN EU	2.8	3.8	26	25	2.35
Male SSN IS	**6.0**	8.2	**30**	43	3.50
Male SSN KL	3.1	3.9	25	23	2.27
Male BAB EU	3.6	13.0	24	**89**	2.90
Male BAB IS	3.4	16.2	24	**89**	**3.83**
Male BAB KL	5.0	14.8	23	88	3.34
Female SSN EU	3.4	2.4	8	12	2.57
Female SSN IS	3.8	3.1	**30**	25	2.95
Female SSN KL	3.3	2.5	8	10	2.59
Female BAB EU	5.1	15.5	22	88	3.69
Female BAB IS	5.3	**16.7**	21	86	3.77
Female BAB KL	5.4	15.5	15	87	3.57

Pearson correlations were performed between the subjective scores and the interference-related objective metrics SIR and IPS, as well as overall quality metrics SDR and OPS. Significant strong correlations were found between subjective scores and SIR ($r(10) = .86, p < .001$), between subjective scores and IPS ($r(10) = .80, p < .005$), and between subjective scores and SDR ($r(10) = .75, p = .005$). Correlation between subjective scores and OPS were mild and not significant ($r(10) = .18, p = .570$). Figure 16, Figure 17Figure 18, andFigure 19 show scatter plots and best fit correlation lines for subjective residual noise scores vs. SIR, IPS, SDR, and OPS metrics respectively. The SIR, IPS, and SDR metrics all had strong correlations with the subjective residual noise scores suggesting that these metrics were effective in predicting perceived quality based on objective scores. The OPS metric, however, had only a mild correlation with subjective residual noise scores and was not as effective at predicting perceived quality scores.

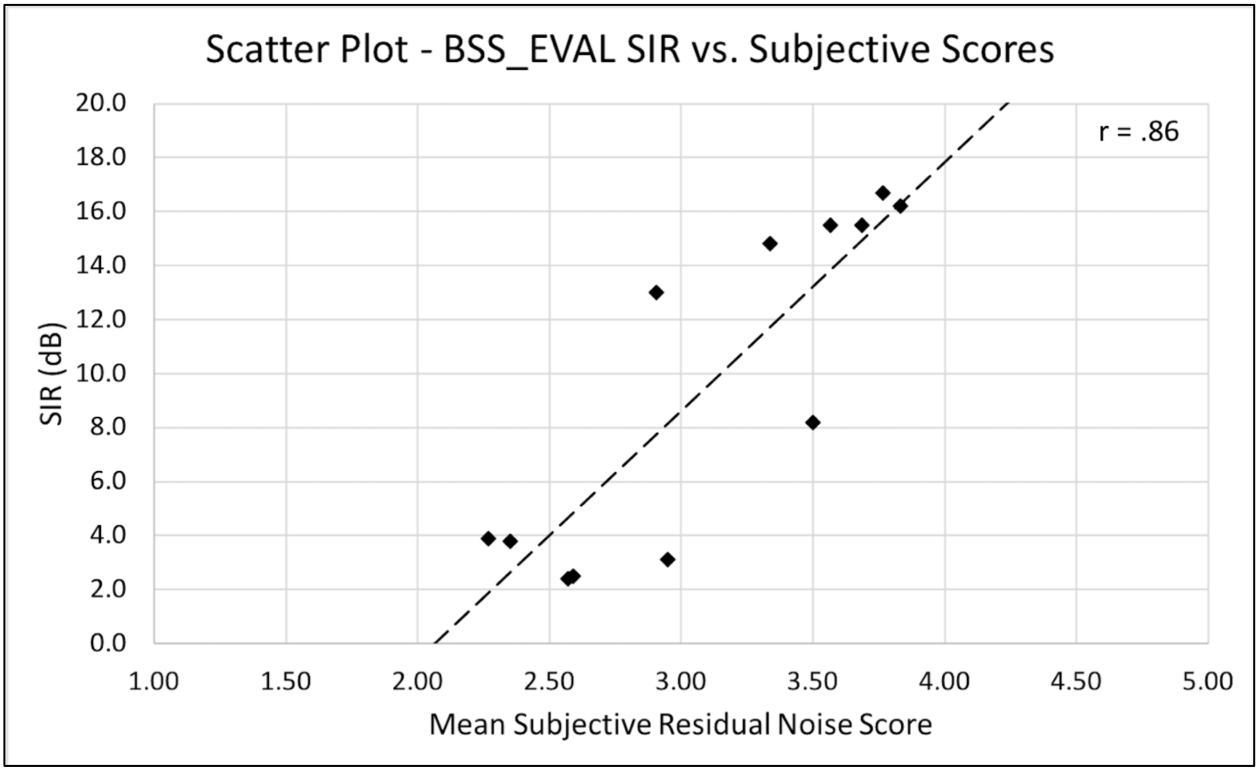


Figure 16. Scatter plot of subjective residual noise scores vs. SIR metrics.

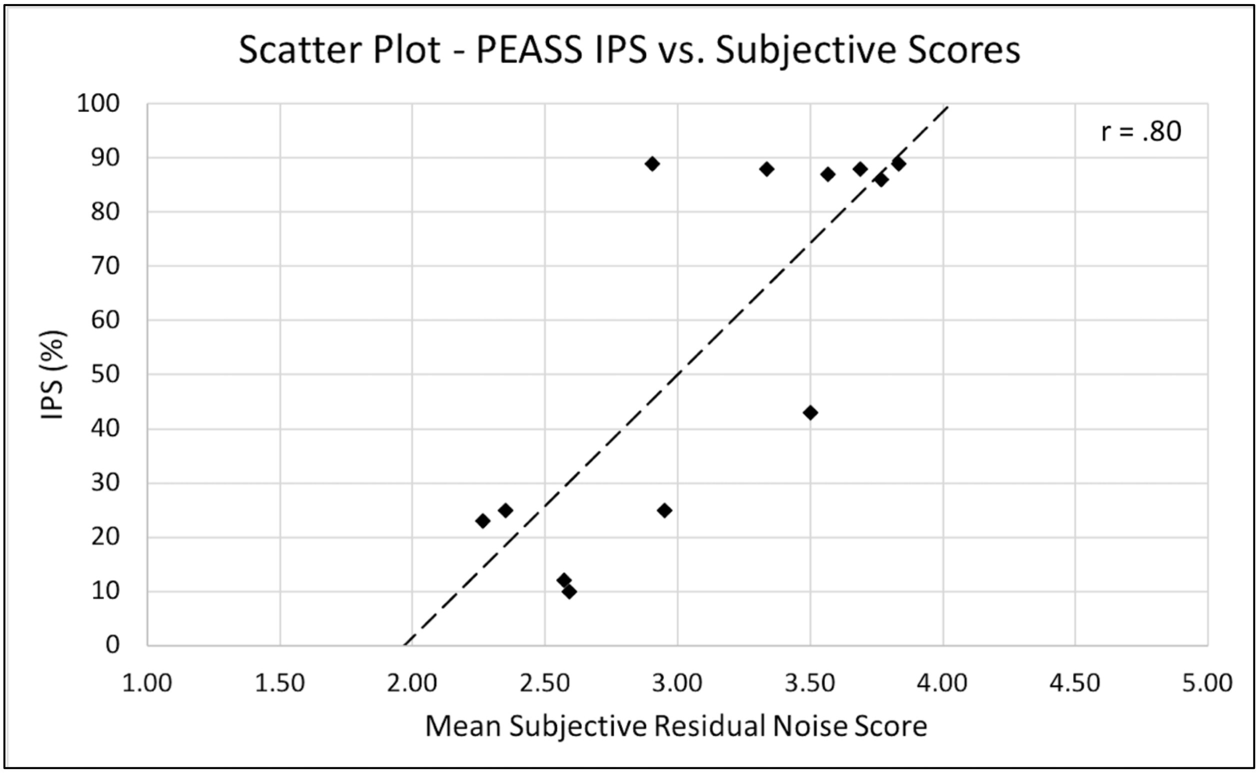


Figure 17. Scatter plot of subjective residual noise scores vs. IPS metrics.

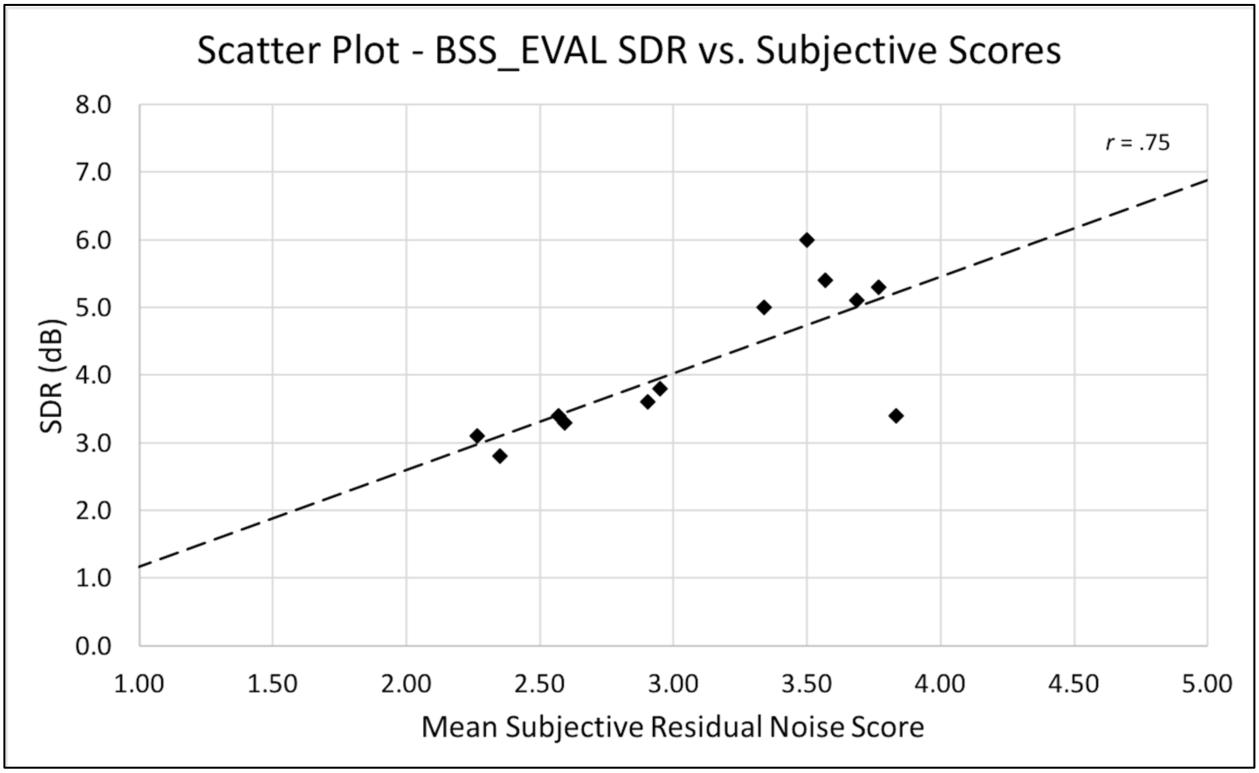


Figure 18. Scatter plot of subjective residual noise scores vs. SDR metrics.

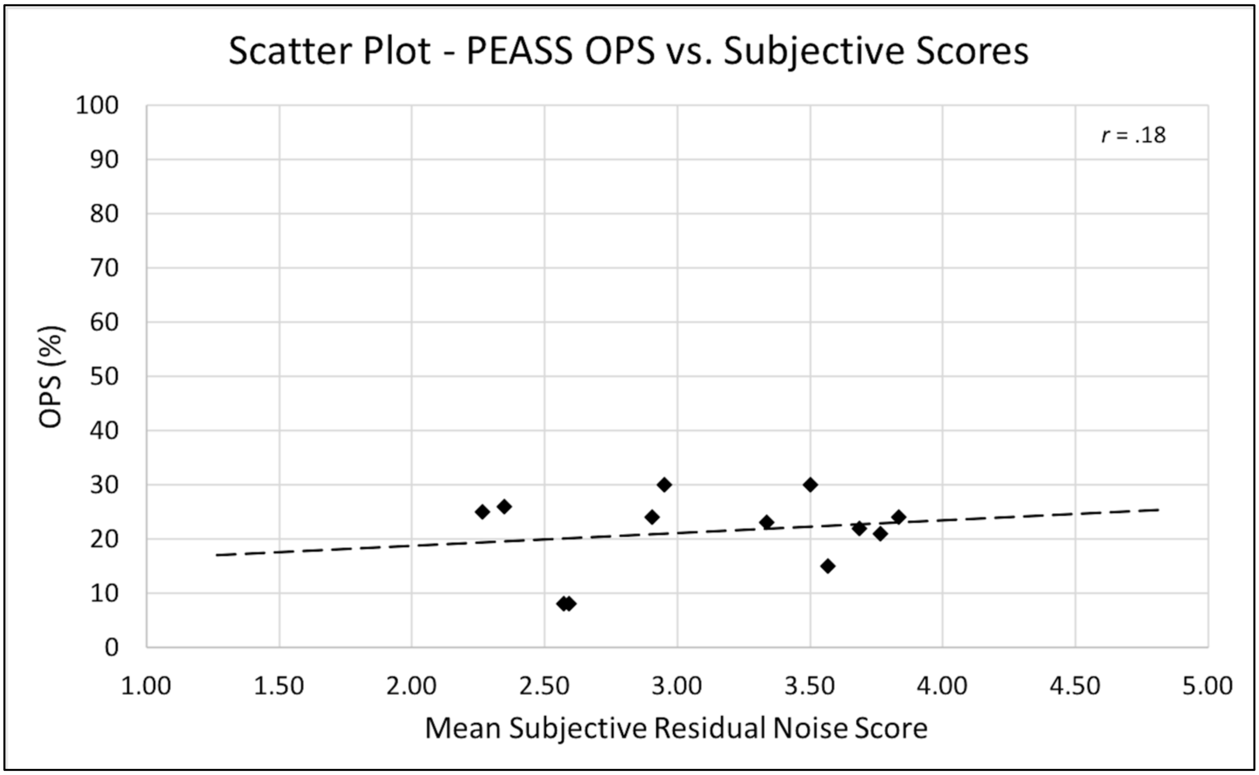


Figure 19. Scatter plot of subjective residual noise scores vs. OPS metrics.

5.0 DISCUSSION

5.1 Experiment 1

This experiment is a continuation of a previous study on the effect of STFT window duration on NMF performance [27]. The objective of the present study was to determine if the STFT window duration used in NMF speech enhancement had any effect on intelligibility of enhanced speech. Based on the preceding analysis, the null hypothesis can be rejected since the window duration was found to have a significant effect on intelligibility scores. Additionally, it was found that both noise source and the sentence talker had significant effects on intelligibility.

The STFT window duration was found to have a significant effect on mean intelligibility scores. Figure 7 reveals a general trend that as window duration increased so did the intelligibility scores. The STFT is known for a rigid trade off in time-frequency resolution, and the increased frequency resolution of longer window durations allows more precise spectral basis functions leading to improved separation and intelligibility scores. Shorter window durations also suffer from the presence of audible digital artifacts [27], resulting in a slight decrease of intelligibility scores from unprocessed speech when using a window duration of 11.6 ms. The observed trend raises the question: "if window durations continued to increase, would intelligibility scores likewise increase?" Longer window durations reduce temporal resolution and would likely begin to hinder intelligibility scores by smearing HFE often found in speech consonants. Paliwal and Alsteris [43] also noted that window durations longer than approximately 100 ms, when relying solely on magnitude information, lead to a reduction in intelligibility. The reduction in temporal precision with longer window durations suggests that eventually the window duration will reach a point where the intelligibility scores begins to decrease. The effect of window durations larger than 92.9 ms on intelligibility could be a topic of future research.

The type of interfering noise was also found to have a significant effect on intelligibility scores, with sentences in babble noise having significantly higher intelligibility scores than SSN. NMF is able to achieve improved separation performance when sources have minimal overlap in the time-frequency domain [14, 30]. For this experiment, SSN was generated using talkers from the same speech corpus used to create the test stimuli. Consequently, the spectrum of SSN more closely resembles the target speech than the babble noise does. The stationary characteristics of SSN also provide consistent masking effects, whereas the babble noise time-frequency characteristics are more variant with fewer masking effects, resulting in better intelligibility scores of target speech in the presence of babble noise even in the unprocessed case. Figure 20 shows spectrograms of the babble noise and SSN. It can be seen that the SSN has constant high energy up to 10 kHz, while babble noise has notable dips in amplitude between 2.5-10 kHz.

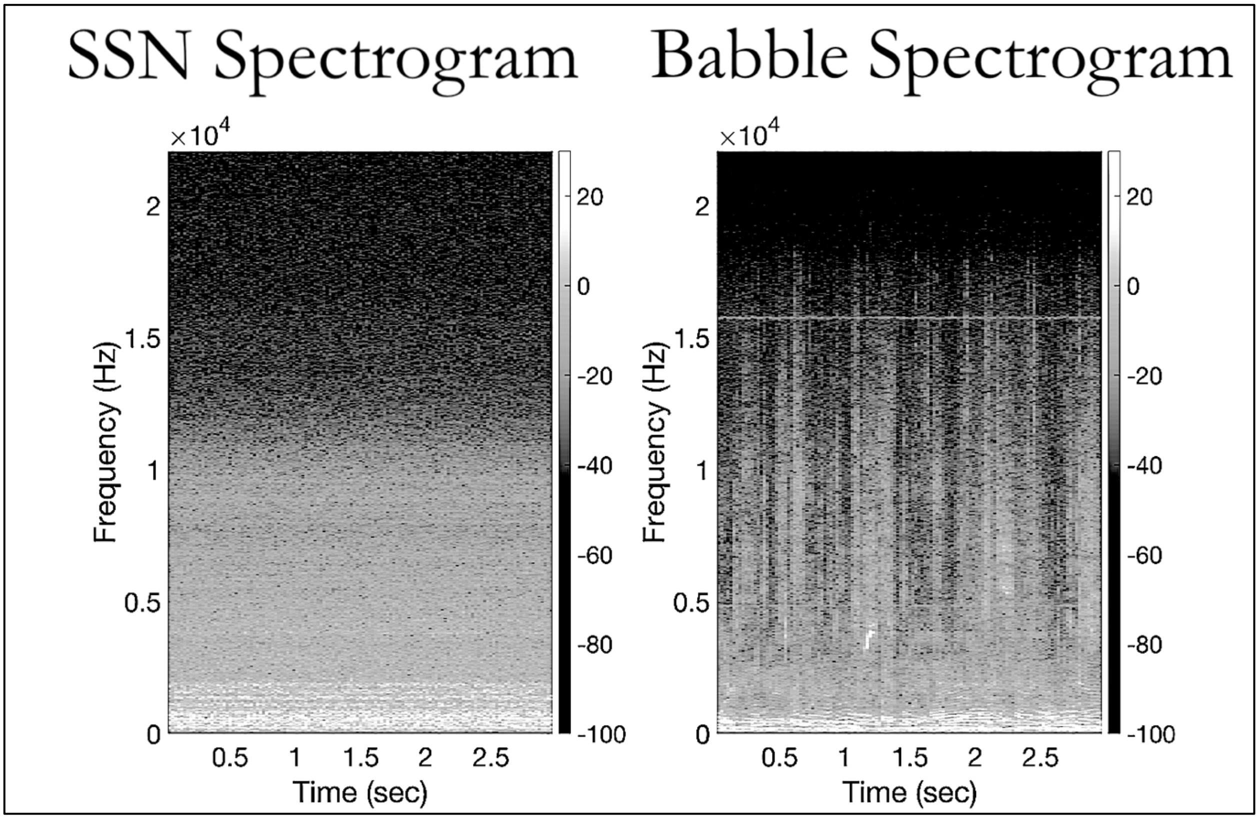


Figure 20. Spectrogram of SSN (left) and babble noise (right).

Finally, it was determined that sentences with female talkers had significantly higher mean intelligibility scores than those with male talkers. Extensive research has been conducted on intelligibility differences between male and female speech, with Bradlow, Toretta, and Pisoni [44] similarly finding that female speech is more intelligible than male speech. One possible reason could be differences in the fundamental frequency and fundamental frequency range of male and female speech, with female talkers typically having higher fundamental frequencies [44, 45]. However, it has been shown that these attributes have little correlation with listener intelligibility [46, 47]. The higher fundamental frequencies of female speech also produce higher harmonic frequencies. When looking at the noise spectrograms in Figure 20, most of the energy resides in the low frequencies, suggesting that harmonics in female speech may suffer fewer masking effects than in male speech, which could contribute to higher intelligibility scores. However, since the original mixtures were normalized to digital full scale prior to NMF processing, the RMS amplitude of each mixture varied slightly. The variation in RMS amplitudes may have had an inadvertent impact on the degree that masking effects impacted intelligibility scores.

Figure 10 reveals that female talkers produced relatively high intelligibility scores for window durations below 23.2 ms while male talkers had relatively low intelligibility scores for these shorter window durations. One explanation could be that the reduced temporal resolution of longer window durations causes smearing of speech high frequency harmonics, which are useful in maintaining intelligibility of consonants particularly for female speech [48]. With shorter window durations, the increased temporal resolution preserves these high frequency harmonics which could contribute to the increased intelligibility scores of female talkers, even in the presence of audible digital artifacts. Figure 10 also reveals that each talker had peak mean intelligibility scores at different window durations, for example Female 1 intelligibility scores were highest at 11.6 ms while Female 2 intelligibility scores were highest at 46.4 ms. The variation in best window duration for each talker could suggest that different optimum window durations exist for different talkers. These differences in optimum window duration could also be attributed to inherent variations in the stimuli used for the present study. Further research is needed to understand the interaction effect of different talkers and window duration on intelligibility of extracted speech.

5.2 Experiment 2

The objective of this experiment was to determine if the divergence cost function used to approximate audio sources in NMF had an effect on the level of residual noise in the target speech. Based on the preceding analysis, the null hypothesis for this experiment can be rejected since the cost function was found to have a significant effect on subjective residual noise levels. The type of interfering noise was also found to have a significant effect on subjective residual noise levels, while talker gender did not have a significant effect.

Three divergence cost functions were considered in the present experiment: the Euclidian Distance (EU), the Kullback-Leibler (KL) divergence, and the Itakura-Saito (IS) divergence. The IS divergence was found to provide the lowest residual noise levels amongst the three cost functions. While all cost functions investigated belong to the Bregman divergence family and share similar characteristics, the IS divergence is the only one that exhibits the property of scale invariance [31]. Scale Invariance indicates that both small and large amplitude time-frequency bins of the mixture $\boldsymbol{V}$ contribute equally to the total cost as calculated in Equation 5. Specifically, a poor approximation for low-amplitude bins will cost the same as a poor approximation for high-amplitude bins. Deviations between the input mixture and approximation for low-amplitude time-frequency bins are then optimized more effectively during algorithm convergence, resulting in greater precision of low-amplitude input components in the approximation. The EU and KL divergences do not exhibit scale invariance, and therefore poor approximations of high-amplitude bins are more heavily weighted in the cost calculations from Equations 3 and 4 than low-amplitude bins. In audio signals, the amplitude of harmonic frequencies in complex signals decrease exponentially as frequency increases [31, 49]. The IS divergence optimizes approximations of these high frequency components more than the EU or KL divergences, resulting in more effective separation and therefore lower residual noise levels.

One trade off of the scale invariance property is the production of audible digital artifacts. Since the IS divergence equally penalizes all differences between $\boldsymbol{V}$ and $\boldsymbol{WH}$, there is greater separation of speech and noise, and therefore a larger portion of noise is removed. However, speech and noise are both broadband signals with substantial overlap of time-frequency energy, which means that not all of the noise can be separated from target speech. The removal of additional noise means that there are more discontinuities in the time-frequency domain for the residual noise, which result in audible digital artifacts in the form of musical noise, introducing a distorted robotic effect in the extracted speech. The introduction of audible digital artifacts is not as prevalent when using the EU or KL divergence, however there is more residual noise present as the results of this experiment have shown.

Another difference between cost functions is their convexity properties. Both EU and KL divergences have been shown to be convex with respect to either $\boldsymbol{W}$ or $\boldsymbol{H}$, while the IS divergence does not have guaranteed convexity [50]. The IS divergence is then more prone to local minima and therefore not converging to the best approximation. From Figure 15, we see that the Male SSN, Male babble, and Female babble sentences with the IS divergence were able to achieve high mean subjective residual noise scores. For Female SSN, however, the IS divergence had similar mean residual noise levels as EU and KL, suggesting that it was unable to converge to the best approximation. The lack of a guaranteed global minima with respect to $\boldsymbol{W}$ or $\boldsymbol{H}$ implies that the IS divergence is prone to variability in separation performance.

The type of interfering noise was found to have a significant effect on residual noise levels, with babble mixtures having lower residual noise levels than SSN mixtures. As discussed in Section 5.1, babble noise had fewer masking effects than SSN, allowing easier separation of target speech and interfering noise. Since there was not a significant interaction effect between noise and cost function, improved separation was achieved for babble mixtures compared to SSN mixtures for all three cost functions.

Talker gender was found to not have a significant effect on residual noise, but the interaction of gender and cost function was significant. Figure 14 reveals fairly consistent subjective residual noise scores for the Female talker across all three cost functions, with IS having an improvement of only 0.23 over EU and 0.28 over KL. However, the effect of cost function on the Male talker is more substantial, with the IS divergence improving subjective residual noise scores by 1.04 over EU and 0.87 over KL. Févotte et. al. [31] showed that the IS divergence is able to more accurately represent transients and low amplitude time-frequency bins of the mixture which occur primarily at high frequencies. In the Female speech-noise mixtures, the high frequency harmonics and transients of the target speech are not masked as heavily and therefore all three divergence functions are able to minimize deviations between the mixture and trained speech basis functions, providing similar residual noise levels. Since the IS divergence is better at minimizing these deviations, the slight improvement in noise reduction compared to the EU and KL divergences is logical. With male speech mixtures, however, speech high frequencies and transients are more heavily masked by the noise and differences between the mixture and trained speech basis functions are larger. The EU and KL divergences do not penalize these high frequency differences as much, and therefore are unable to separate high frequency speech and noise as effectively as the IS divergence. The ability of the IS divergence to effectively separate the low-amplitude high-frequency components in speech leads to the large improvement in subjective residual noise scores observed for male speech using the IS divergence as observed in Figure 14. However, as discussed in Section 5.1, the variation of mixture RMS amplitudes prior to NMF processing due to full scale normalization may have had an inadvertent impact on the degree to which masking effects impacted residual noise scores.

Finally, a correlation analysis was performed between the SIR, SDR, IPS, and OPS objective metrics and the subjective residual noise scores. Significant strong positive correlations were found for SIR, SDR and IPS, indicating that for this experiment these metrics were good predictors of subjective scores. OPS only had a mild correlation which was not significant. As detailed in Section 2.2, source separation literature has provided mixed results on the validity of these metrics as tools to evaluate subjective separation performance. It was noted in Section 2.5 that significant correlations were found between objective metrics and subjective scores by Kornycky et. al. [25] when high-quality, full-bandwidth audio sampled at 44.1 kHz was used. The current study similarly used sampling rates of 44.1 kHz resulting in significant strong correlations between objective metrics and subjective scores. While OPS did not result in strong correlation with subjective scores, the current study asked listeners to rate stimuli based on the level of interfering noise present in target speech. OPS is intended to quantify overall quality of separated sources, and as discussed above the IS divergence is able to remove more noise but also introduces audible digital artifacts that can degrade overall quality. The fact that subjective scores in the current study were not based directly on overall quality of separated speech could explain why correlations were only mild. The strong significant correlations of the interference-related metrics SIR and IPS are a promising outcome suggesting that high quality audio may be necessary to improve the reliability of objective measures, however more research is needed to confirm the interaction of audio sampling rate with the correlation of subjective and objective metrics for source separation.

6.0 CONCLUSIONS

The goal of this study was to determine the perceptual effects of the STFT window duration and cost function used in NMF audio source separation. It was found that the STFT window duration had a significant effect on the mean intelligibility score of speech separated from noise. Additionally, it was found that the cost function had a significant effect on the perceived level of residual noise in separated speech. The results of the present study provide evidence to the importance of these parameters on source separation performance and suggest that researchers and engineers must use care when selecting parameters for a specific source separation task. The findings of this study can be used to improve the perceptual performance of robust source separation algorithms utilizing NMF for source decomposition. These findings may also apply to similar source separation techniques that utilize time-frequency signal representations and approximations generated from divergence cost functions. Furthermore, it was observed that common source separation objective parameters had significant positive correlation with subjective residual noise scores in Experiment 2 when full bandwidth audio sampled at 44.1 kHz was used. This suggests that, when using high quality audio stimuli, objective metrics could be more accurate in predicting subjective performance than when using audio at lower sampling rates. Further research should be conducted to verify the relationship between audio sampling rate and the correlation of objective metrics to subjective ratings.

BIBLIOGRAPHY

[1] D. Lee and H. Seung, "Algorithms for Non-negative Matrix Factorization," in *13th International Conference on Neural Information Processing Systems*, Denver, CO, 2000.

[2] J. Park, J. Shin and K. Lee, "Separation of instrument sounds using non-negative matrix factorization with spectral envelope constraints," 2018. [Online]. Available: arXiv:1801.04081.

[3] K. Wilson, B. Raj, P. Smaragdis and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Cambridge, MA, USA, 2008.

[4] F. Weninger, J. Le Roux, J. R. Hershey and S. Watanabe, "Discriminative NMF and its application to single-channel source separation," in *15th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Singapore, 2014.

[5] N. Moritz, M. R. Schadler, K. Adiloglu, B. T. Meyer, T. Jurgens, T. Gerkmann, B. Kollmeier, S. Doclo and S. Goetze, "Noise robust distant automatic speech recognition utilizing nmf based source separation and auditory feature extraction," in *2nd CHiME Workshop on Machine Listening in Multisource Environments*, Vancouver, BC, Canada, 2013.

[6] B. Raj, T. Virtanen, S. Chaudhuri and R. Singh, "Non-negative matrix factorization based compensation of music for automatic speech recognition," in *11th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Makuhari, Japan, 2010.

[7] Z. Duan, G. J. Mysore and P. Smaragdis, "Speech enhancement by online non-negative spectrogram decomposition in nonstationary noise environments," in *13th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Portland, OR, USA, 2012.

[8] B. Raj, R. Singh and T. Virtanen , "Phoneme-dependent NMF for speech enhancement in monaural mixtures," in *12th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Florence, Italy, 2011.

[9] G. Richard, "Informed Audio Source Separation," in *AES 53rd International Conference*, London, UK, 2014.

[10] P. Smaragdis and G. Mysore, "Separation by "humming": user-guided sound extraction from monophonic mixtures," in *EEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, 2009.

[11] J. Le Roux, J. R. Hershey and F. Weninger, "Deep NMF for speech separation," in *EEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Brisbane, Australia , 2015.

[12] T. Kang, K. Kwon, J. Shin and N. Kim, "NMF-based target source separation using deep neural network," *IEEE Signal Processing Letters,* vol. 22, no. 2, pp. 229-233, 2015.

[13] T. T. Vu, B. Bigot and E. S. Chang, "Combining non-negative matrix factorization and deep neural networks for speech enhancement and automatic speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Shanghai, China, 2016.

[14] P. Smaragdis, "Convolutive Speech Bases and their Application to Supervised Speech Separation," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 15, no. 1, pp. 1-12, 2007.

[15] N. Bryan, D. Sun and E. Cho, "Introduction to Non-negative Matrix Factorization," in *Single-Channel Source Separation Tutorial Mini-Series*, Stanford, CA, USA, 2013.

[16] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural Computation,* vol. 23, no. 9, pp. 2421-2456, 2011.

[17] W. Fonseca, Z. Peixoto, F. Magalhaes and R. Faria, "A New Recursive Semi-Supervised Non-Negative Matrix Factorization for Separation of Harmonic and Percussive Elements in Digital Sounds," *J. Audio Eng. Soc. ,* vol. 66, no. 10, pp. 779-790, 2018.

[18] J. J. Burred, "Detailed derivation of multiplicative update rules for NMF," Paris, France, 2014.

[19] N. Bryan, D. Sun and E. Cho, "Extensions and interpretations to non-negative matrix factorization," in *Single-Channel Source Separation Tutorial Mini-Series*, Stanford, CA, USA, 2013.

[20] S. K. Tjoa and K. J. R. Liu, "Multiplicative update rules for nonnegative matrix factorization with co-occurrence constraints," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, TX, USA, 2010.

[21] V. Emiya, E. Vincent, N. Harlander and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech and Language Processing,* vol. 19, no. 7, pp. 2046-2057, 2011.

[22] E. Cano, D. FitzGerald and K. Brandenburg, "Evaluation of Quality of Sound Source Separation Algorithms: Human Perception vs Quantitative Metrics," in *24th European Signal Processing Conference*, Budapest, Hungary, 2016.

[23] U. Gupta, E. Moore and A. Lerch, "On the perceptual relevance of objective source separation measures for singing voice separation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2015.

[24] D. Ward, H. Wierstorf, R. D. Mason, E. Grais and M. D. M Plumbley, "BSS eval or PEASS? Predicting the perception of singing-voice separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada, 2018.

[25] J. Kornycky, B. Gunel and A. Kondoz, "Comparison of subjective and objective evaluation methods for audio source separation," *Proceedings of Meetings on Acoustics,* vol. 4, no. 1, 2008.

[26] E. Vincent, R. Gribonval and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 14, no. 4, pp. 1462-1469, 2006.

[27] R. Miller, W. Bulla and E. Tarr, "Detection of the effect of window duration in an audio source separation paradigm," in *147th AES Convention*, New York, NY, USA, 2019.

[28] B. Jiang and J. Yang, "Preferred frame length for the short-time magnitude spectrum on speech intelligibility and speech quality," in *8th International Conference on Information, Communications, & Signal Processing*, Singapore, Singapore, 2011.

[29] K. Paliwal and K. Wójcicki, "Effect of analysis window duration on speech intelligibility," *IEEE Signal Processing Letters,* vol. 15, pp. 785-788, 2008.

[30] B. King, C. Févotte and P. Smaragdis, "Optimal Cost Function And Magnitude Power For NMF-Based Speech Separation And Music Interpolation," in *IEEE International Workshop on Machine Learning for Signal Processing*, Satander, Spain, 2012.

[31] C. Févotte, N. Bertin and J.-L. Durrieu, "Nonnegative Matrix Factorization with the Itakura-Saito Divergence: With Application to Music Analysis," *IEEE Neural Computation,* vol. 21, no. 3, pp. 793-830, 2009.

[32] D. FitzGerald and R. Jaiswal, "On the use of masking filters in sound source separation," in *15th International Conference on Digital Audio Effects (DAFx)*, York, UK, 2012.

[33] F. Germain and G. Mysore, "Stopping criteria for non-negative matrix factorization based supervised and semi-supervised source separation," *IEEE Signal Processing Letters,* vol. 21, no. 9, pp. 1-5, 2014.

[34] S. Mohammed and I. Tashev, "A statistical approach to semi-supervised speech enhancement with low-order non-negative matrix factorization," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, New Orleans, LA, USA, 2017.

[35] S. Wisdom, T. Powers, J. Pitton and L. Atlas, "Deep recurrent NMF for speech separation by unfolding iterative thresholding," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2017.

[36] N. Mohammadiha, P. Smaragdis and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 21, no. 10, pp. 2140-2151, 2013.

[37] B. B. Monson, E. J. Hunter, A. J. Lotto and B. H. Story, "The perceptual significance of high-frequency energy in the human voice," *Frontiers in Psychology,* vol. 5, no. 587, pp. 1-11, 2014.

[38] Rothauser, "IEEE Recommended Practice for Speech Quality Measurements," *IEEE Transactions on Audio and Electroacoustics,* vol. 17, no. 3, pp. 225-246, 1969.

[39] R. M. Cox, "Development of the Connected Speech Test (CST)," *Ear and Hearing,* vol. 8, no. 5, 1987.

[40] F.-R. Stoter, S. Uhlich, A. Liutkus and Y. Mitsufuji, "Open-Unmix - A reference implementation for music source separation," *Journal of Open Source Software,* vol. 41, no. 4, pp. 1-6, 2019.

[41] MATLAB Release 2019b, The MathWorks, Inc., Natick, MA, USA.

[42] ITU-R, "ITU-R BS.1116-3, "Methods for the subjective assessment of small impairments in audio systems," 2015.

[43] K. K. Paliwal and L. Alsteris, "Usefulness of phase spectrum in human speech perception," in *Eurospeech*, Geneva, Switzerland, 2003.

[44] A. R. Bradlow, G. M. Toretta and D. B. Pisoni, "Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics," *Speech Commun,* vol. 20, no. 3, pp. 255-272, 1997.

[45] Z. S. Bond and T. J. Moore, "A note on the acoustic phonetic characteristics of inadvertently clear speech," *Speech Commun,* vol. 14, no. 4, pp. 325-337, 1994.

[46] H.-B. Kwon, "Gender difference in speech intelligibility using speech intelligibility tests and acoustic analyses," *J. Adv. Prosthodont,* vol. 2, no. 3, pp. 71-76, 2010.

[47] B. S. Zinny and T. J. Moore, "A note on the acoustic-phonetic characteristics of inadvertently clear speech," *Speech Communication,* vol. 14, no. 4, pp. 325-337, 1994.

[48] R. P. Lippmann, "Accurate consonant perception without mid- frequency speech energy," *IEEE Transactions on Speech and Audio Processing,* vol. 4, no. 1, pp. 66-69, 1996.

[49] B. C. J. Moore, M. A. Stone, C. Füllgrabe, B. R. Glasberg and S. Puria, "Spectro-Temporal Characteristics of Speech at High Frequencies, and the Potential for Restoration of Audibility to People with Mild-to-Moderate Hearing Loss," *Ear Hear,* vol. 29, no. 6, pp. 907-922, 2008.

[50] N. Bertin, C. Févotte and R. Badeau, "A tempering approach for Itakura-Saito non-negative matrix factorization. With application to music transcription," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, 2009.

[51] R. Badri, J. H. Siegel and B. A. Wright, "Auditory filter shapes and high-frequency hearing in adults who have impaired speech in noise performance despite clinically normal audiograms," *J Acoust Soc Am,* vol. 129, pp. 852-863, 2011.

APPENDICES

Appendix A – Speech Shaped Noise

MATLAB Code used to generate Speech Shaped Noise can be found at the following link: https://www.mathworks.com/matlabcentral/fileexchange/55701-speech-spectrum-shaped-noise

SSN was generated using 2 sentences from each talker in the IEEE speech corpus used for Experiment 1 and Experiment 2 [38]. The same SSN was used for both experiments. None of the sentences used to create SSN were used in any of the training or test sets implemented in either of the experiments from the present study.

Appendix B – Experiment 2 MATLAB Code and Stimuli

MATLAB Code and Stimuli from Experiment 2 can be found at the following link: https://rjmiller927.github.io/research/mastersThesis.html