2013

Perceptual Ruler for Quantifying Speech Intelligibility in Cocktail Party Scenarios

Kirstin M. Brangers
University of Kentucky, Kirstin.Brangers@uky.edu

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 and attached hereto needed written permission statements(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).

I hereby grant to The University of Kentucky and its agents the non-exclusive 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 a preapproved 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 dissertation including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Kirstin M. Brangers, Student

Dr. Kevin D. Donohue, Major Professor

Dr. Zhi David Chen, Director of Graduate Studies

PERCEPTUAL RULER FOR QUANTIFYING SPEECH INTELLIGIBILTY IN COCKTAIL PARTY SCENARIOS

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

Kirstin Marie Brangers

Lexington, Kentucky

Director: Dr. Kevin Donohue, Professor of Electrical Engineering

Lexington, Kentucky

2013

ABSTRACT OF THESIS

PERCEPTUAL RULER FOR QUANTIFYING SPEECH INTELLIGIBILTY IN COCKTAIL PARTY SCENARIOS

Systems designed to enhance intelligibility of speech in noise are difficult to evaluate quantitatively because intelligibility is subjective and often requires feedback from large populations for consistent evaluations. Attempts to quantify the evaluation have included related measures such as the Speech Intelligibility Index. These require separating speech and noise signals, which precludes its use on experimental recordings. This thesis develops a procedure using an Intelligibility Ruler (IR) for efficiently quantifying intelligibility. A calibrated Mean Opinion Score (MOS) method is also implemented in order to compare repeatability over a population of 24 subjective listeners. Results showed that subjects using the IR consistently estimated SII values of the test samples with an average standard deviation of 0.0867 between subjects on a scale from zero to one and R^2=0.9421. After a calibration procedure from a subset of subjects, the MOS method yielded similar results with an average standard deviation of 0.07620 and R^2=0.9181.While results suggest good repeatability of the IR method over a broad range of subjects, the calibrated MOS method is capable of producing results more closely related to actual SII values and is a simpler procedure for human subjects.

KEYWORDS: Quality Ruler, Cocktail Party, Speaker Intelligibility Index, Audio Intelligibility, Intelligibility Ruler

MULTIMEDIA ELEMENTS USED: WAV (.wav), MATLAB programs (.m), MATLAB GUI Figures (.fig)

Kirstin Marie Brangers

July 31, 2013

PERCEPTUAL RULER FOR QUANTIFYING SPEECH INTELLIGIBILTY IN COCKTAIL PARTY SCENARIOS

By

Kirstin Marie Brangers

Dr. Kevin D. Donohue
Director of Thesis

Dr. Zhi David Chen
Director of Graduate Studies

July 31, 2013

To my parents – Joe and Lisa

ACKNOWLEDGEMENTS

The development of this thesis would not have been possible without the support and guidance of several influential and knowledgeable individuals who helped contribute to the discipline needed in order to complete this study.

First, my highest gratitude to my advisor, Dr. Kevin Donohue, for not only his support and guidance in pursuing my M.S. in Electrical Engineering, but also for his immense knowledge in this field and ability to influence others to develop a sense of curiosity and passion for their work.

Dr. Hassebrook and Dr. Patwardhan for their involvement on my Defense Committee, as well as contributions in the curriculum part of my academic degree.

My friends who have been there to encourage, understand, and most importantly listen when it seemed there was no end in sight. I would like to thank Michael, who never hesitated to help me through a difficult time and provided consistent support.

My mom for her continuous care and constructive criticism to keep me focused, as well as lifelong motivation to always conquer the challenging rather than take the path of least resistance. My dad for being an additional resource and instilling the work ethic that was/is needed to successfully meet my goals. I want to thank my sisters – Corie and Hailey – who were always there to provide a fresh breath of air and a laugh, or two, during this process.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF FILES

Name of File	Type	Size
Man1	.wav	251 KB
Man2	.wav	251 KB
Man3	.wav	251 KB
Man4	.wav	251 KB
Man5	.wav	251 KB
Woman1	.wav	251 KB
Woman2	.wav	251 KB
Woman3	.wav	251 KB
Woman4	.wav	251 KB
IR_Noise	.wav	251 KB
SOI_Woman	.wav	167 KB
IR_RefSII_000	.wav	167 KB
IR_RefSII_010	.wav	167 KB
IR_RefSII_015	.wav	167 KB
IR_RefSII_020	.wav	167 KB
IR_RefSII_025	.wav	167 KB
IR_RefSII_030	.wav	167 KB
IR_RefSII_035	.wav	167 KB
IR_RefSII_040	.wav	167 KB
IR_RefSII_050	.wav	167 KB
IR_RefSII_060	.wav	167 KB
IR_RefSII_070	.wav	167 KB
IR_RefSII_080	.wav	167 KB
IR_RefSII_090	.wav	167 KB
IR_RefSII_100	.wav	167 KB
SOI_Man	.wav	251 KB
TestRec_Noise1	.wav	251 KB
TestRec_Noise2	.wav	251 KB
TestRec_Noise3	.wav	251 KB
TestRec_Noise4	.wav	251 KB
Test1	.wav	171 KB
Test2	.wav	171 KB
Test3	.wav	171 KB
Test4	.wav	171 KB
Test5	.wav	171 KB

Test6	.wav	171 KB
Test7	.wav	171 KB
Test8	.wav	171 KB
Test9	.wav	171 KB
Test10	.wav	171 KB
computeRMS	.m	2 KB
gong	.m	1 KB
intel	.m	4 KB
loadQR	.m	2 KB
loadSS	.m	2 KB
QR_UI	.fig	9 KB
QR_UI	.m	20 KB
rmsilence	.m	5 KB
RunExp	.m	2 KB
sii	.m	12 KB
spectrumlevel	.m	3 KB
SS_Radio	.fig	5 KB
SS_Radio	.m	11 KB
testscript_intel	.m	8 KB
wav2sig	.m	5 KB
weightsSOI	.m	6 KB

CHAPTER 1: INTRODUCTION

1.1 Cocktail Party Scenario

The cocktail party scenario is a familiar topic of interest in the area of Digital Signal Processing (DSP) involving a number of people engaging in simultaneous, mutually exclusive conversations in the same room or area. The term mutually exclusive is used to describe the notion that each individual is only participating in one conversation at any given moment rather than multiple conversations. Cognitive and auditory systems have the innate ability to focus on and extract a particular Speaker of Interest (SOI) from the interfering stimuli which is deemed noise, allowing one to hold conversations with another human in a noisy environment. In the area of DSP, steps have been taken to mimic this natural human process using arrays of microphones in a room, multichannel recordings, and advanced computer algorithms involving Sound Source Location and Auditory Scene Analysis. However, once the SOI is extracted from the background noise the intelligibility and quality of the final recording is of huge importance to the end user wanting to understand the removed information.

1.2 Perception of Quality

This raises curiosity as to understanding the perception of intelligibility of a speaker of interest in an audio signal. In this case, the original recording includes the SOI and multiple interfering speakers acting as background noises. The output recording includes the SOI signal extracted from the interfering noise where the interferences are masked in such a way that solely the SOI is audible. Thus, in the engineering

community, the perception of quality is important in evaluating and testing algorithm efficiency.

1.3 Image Quality

Research has been done that looks at trying to quantify a subject's perception of quality as it deals with an image. These studies aim at understanding a subject's perception of quality when assessing images where varying distortions were present. Multiple experimental methodologies have been implemented to assess and quantify a subject's perception, a subjective trait, in order to gain a better understanding into what attribute of quality could be altered in order to make an image more appealing, as well as provide insight to technological advancements in the area of imaging [1] [2] [3].

The use of a Quality Ruler (QR) has been successfully implemented which required subjects to perform a modified paired comparison test in order to assess the image quality. The QR consists of reference image samples, each with a known standard value, spanning a pre-determined, pre-calibrated scale. The subject is presented with a sample image and must make a decision as to where on the ruler the sample image belongs. This required the subject to compare the sample image with each individual reference image and then rank it according to the pre-determined, pre-calibrated scale which comprises the QR [1] [2] [3].

1.3.1 Subjective Methods of Measuring Image Quality

Several methods exist which use subjective test measures to measure a subject's perception of quality by having subjects directly participate in the testing methods. The results of these methods are different, whether the outcome is a categorical description of the test image or indicates the level of impairment present in the test image compared to

the reference. However, the Quality Ruler method aims at quantifying subjective perception and provides a calibrated scale for comparison amongst different experiments.

1.3.1.1 Single Stimulus Absolute Category Rating (SSACR) Method

The Single Stimulus Absolute Category Rating (SSACR) is a method used to measure the quality of an image relative to only itself. During this method, the subject is presented test images and is required to rate the quality of each individual image according to a categorical scale, as shown in Table 1.1. No specific information regarding the definition of 'quality' is given.

Table 1.1 - Categorical scoring scale for SSACR method [4]

Rating	Categorical Scale
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

Subjects are given no reference image and must rate each image independently; however, it is believed that subjects subconsciously reference the previous test sample and compare the quality of the current test sample before scoring [4]. The average numerical rating for each test image is computed, relaying the average categorical quality of each test image [3].

1.3.1.2 Double Stimulus Impairment Scale (DSIS) Method

The Double Stimulus Impairment Scale (DSIS) method presents the subject with a pair of images (or videos), one of which is the reference unknown to the subject. The

subject rates the quality of the second sample relative to the first sample according to the scale shown in Table 1.2.

Table 1.2 - Categorical scale used for DSIS method [5]

Rating	Impairment Scale
5	Imperceptible
4	Perceptible but not annoying
3	Slightly annoying
2	Annoying
1	Very annoying

The average level of impairment is computed for each test image based on all subject response rating [5].

1.3.1.3 Double Stimulus Continuous Quality Scale (DSCQS) Method

The Double Stimulus Continuous Quality Scale (DSCQS) method requires the subject to individually rate the quality of two images (or videos) shown as a pair, both displaying the same scene. A categorical scale, as shown in Table 1.1, is mapped to the continuous scale as a general reference but users are not limited by these five options. The difference between the quality ratings is interpreted as a quantification of the amount of degradation in the test image compared to the reference image [5].

This method is used in subjective tests where subject ratings may be influenced by external factors, such as the order of presenting test samples and order and level of impairments, and all levels of quality are unknown or cannot be displayed in the test samples [5] [6]. The DSCQS method has shown accurate results despite the external factors present because subjects are required to view each image pair twice and bias

resulting from comparing present and past samples is reduced [6]. The average quality is computed for each impairment and test image using all subject responses; however, it is important to note that the scores are a reflection of difference in quality and cannot be interpreted according to the associated categorical scales [5].

1.3.1.4 Paired Comparison (PC) Method

The Paired Comparison (PC) method requires a pair of images to be presented to the subject. The scene in the images is the same, but the quality of each image differs depending on the type of degradation added. The subject is then asked to choose which image is of better quality. The source image, with no degradation, can be included in the test; however, the subject would have no knowledge as to which image is acting as the source image. Using the PC method provides further insight to a subject's preference between images where the quality of a number of test images are comparable; however, since the method requires all pair combinations possible in the set of test images to be considered, the process is sometimes viewed as too drawn-out and is often used to understand a subject's preference between images of nearly equal quality [4].

1.3.1.5 Degradation Category Rating (DCR)

The Degradation Category Rating (DCR) presents the subject with two images, a reference image and a test image. The pair presented to the subject displays the same scene, but the test image may have some type of slight distortion applied to it. The subject is then requested to rate the level of impairment present in the test image in relation to the reference image according to the scale in Table 1.3 [4].

Table 1.3 – Impairment scale used for DCR method [4]

Rating	Impairment Scale
5	Imperceptible
4	Perceptible but not annoying
3	Slightly annoying
2	Annoying
1	Very annoying

An average impairment score for each image is computed across all subjects to gauge the significance of the perceptual distortion.

1.3.1.6 Quality Ruler (QR) Method

The Quality Ruler (QR) method is defined by the ISO 20462 standard titled "Photography- Psychophysical Experimental Methods to Estimate Image Quality – Part 3." The basis of this method is the idea of just noticeable differences (JNDs), a measure depicting the range of perception. JNDs are used to calibrate the perception of quality so results can be compared between experiments. A JND unit is the smallest change between stimuli perceived and is defined by subject response outcomes of a paired comparison test [1] [2] [3].

JNDs were derived from the notion that perception is probabilistic in nature, tending to a normal distribution as the number of subject responses increases. As the degree of impairment between two images approaches zero (the images are nearly equal), subjects will randomly guess which image is of better quality, resulting in the number of correct responses approaching 50 percent. JNDs are used to define a multitude of perceptual attributes, including individual characteristics and overall quality, and provide a calibrated scale of quality for comparing different results [1] [3].

A QR is a series of same scene images that span a wide range of quality, but are individually close in quality. These images differ by a single attribute, measured in JNDs, and are used as reference images for the subject during testing. Subjects are given a test image to compare against the reference images on the QR, ultimately selecting which reference image the test image is closest in quality. The JND values are displayed on the scale accompanying the QR and the subject is asked to provide an integer rating of the test image. Many times multiple rulers are used, each depicting a separate scene, so that an average JND value can be computed for the test image in order [2] [3].

The QR method has shown reliable results compared to the levels of uncertainty present in other methods of image quality analysis. The QR method has also shown that regardless of the scene present in the image, perceptual assessment of quality remains consistent [1] [3].

1.4 Audio Quality and Speech Intelligibility

In cocktail party scenarios multiple people are speaking simultaneously, normally engaged in different conversations, providing a variety of noise distortions. Algorithms exist in which a SOI can be extracted from a recording where surrounding conversations act as noise in the area. Once the SOI is extracted from the interfering sources, the resulting recording is needed to gain information essential to the end user; thus, measuring the intelligibility of the signal has been and still is a main focus of research.

Signal and Image processing are very closely related, intriguing the idea that a QR measuring image quality could be expanded to assess the intelligibility of a SOI extracted from a noise signal. While the quality of the recording is important, more focus is

placed on the intelligibility of the SOI within the recorded sample. This resulted in the experimental design of an Intelligibility Ruler (IR) for SOI intelligibility analysis. Subjective and objective methods exist for measuring both quality and intelligibility of a signal; however, most objective methods require knowledge of both the speech signal before and after distortion, raising a problem when measuring the quality and intelligibility of a SOI in noise.

Currently, the Speech Intelligibility Index (SII) is one of the most widely used measurements of speech intelligibility [7] [8]. It provides a proportion of how intelligible the speech sample is relative to the noise signal. Computing the SII requires knowledge of the power present in both the noise and speech signal, preventing the use in real world scenarios. Since the SII measure is an accepted standard scale, these previously derived values were used in place of JND measures [7]. By creating the IR based on SII values, a subject's perception can be measured as it relates to speech intelligibility. This has the potential of bridging the gap between the SII scale being used for experimental and actual circumstances. Since subjective methods are a more accurate measure on intelligibility, reproducible and reliable results using the IR offers the opportunity of running small scale subjective tests on data in practical situations. This can potentially reduce the need of a large number of subjects to perform the tests, as well as eliminate the resources and costs associated with running subjective tests involving human subjects.

The objective of this research is to gain an understanding into a listener's judgment of intelligibility of a SOI audio signal extracted from interfering noise sources. By completing this study, the hypothesis that an audio intelligibility ruler can be used as a repeatable measure in evaluating the intelligibility of a SOI in live recordings will be

tested. This study will also examine how accurate the IR is in estimating the actual SII value of a SOI in live recordings. The outcome of the experiment conducted in this thesis will help determine to what degree consistent measures can be obtained over a broad range of subjects assessing the intelligibility of a SOI in cocktail party scenarios and compare to more common SSACR methods.

1.5 Organization of Thesis

In this chapter, the idea of how one perceives quality was presented along with a description of quality analysis in two different but similar areas. Research has been done in order to understand a subject's perception of quality as it relates to image analysis. By expanding the idea of a QR to the realm of audio, an IR is proposed as a subjective method of quantifying perception of intelligibility. Chapter Two describes previous research done in image and audio quality analysis, providing descriptions of both subjective and objective methodologies. Chapter Three describes the experimental design used to acquire subjective data from volunteer subjects and explains how the data will be analyzed. Chapter Four presents and examines the results from the methods used, as well as provides insight to any errors that could result in skewed data. Chapter Five draws overall conclusions to the study based on experimental results, clarifies whether the initial hypothesis was met, and offers future modifications and developments for measuring intelligibility.

CHAPTER 2: PREVIOUS WORK

2.1 Subjective Methods of Measuring Audio Quality

Subjective testing methods exist which require subject participation to directly measure a subject's perception of audio quality. The idea of quality can be interpreted differently. Older subjective tests involving opinion rating methods evaluate quality as the overall impression of the audio signal despite the types of distortions present. These methods leave the definition of quality to be interpreted by the subject, noting that listeners normally use a live person as a reference. The term quality is also used to describe the intelligibility of a person speaking in a recording. In this definition, a positive correlation exists between quality and intelligibility [9]. Unless otherwise stated, the definition of quality as it pertains to intelligibility will be used.

The results of these methods prove to be an accurate measure of subjective perception based on the fact that subjects' themselves are providing direct input to the results. Some of these methods encompass mean opinion scores, rhyme tests, and calculating percentage of intelligibility.

2.1.1.1 Mean Opinion Score

The most known method in signal quality analysis is the Mean Opinion Score (MOS), which evaluates the perceived quality of a sample by averaging the responses from multiple subjects. Either definition of quality can be used in performing this method; however, most experiments performed allow the subject to interpret the term as the overall perception of the signal. To apply this method, subjects will listen to a sample and then rate the quality of the sample on a categorical scale given numerical values ranging from one to five, as described in Table 2.1.

Table 2.1 - Subjective scoring scale used for calculating MOS

Numerical Rating	Categorical Scale
1	Bad
2	Poor
3	Fair
4	Good
5	Excellent

For accurate and repeatable results, multiple subjects are needed to perform this method of evaluation, as individual perception varies among listeners. If the MOSs agree across all subject ratings, a definitive subjective basis of quality results, as individual quality scales among listeners must correlate in some way [9].

2.1.1.2 Rhyme Tests

The Fairbanks, Modified Rhyme Test (MRT) and Diagnostic Rhyme Test (DRT) are examples of rhyme tests. The Fairbanks test presents the subject with 50 speech samples containing different stems, where a stem is the remainder of the word after the initial consonant is removed (i.e. *–ot* is the stem for the words *got, hot, lot*) and various levels of interfering noise. The listener listens to the word corresponding to the first stem and must write the initial consonant in the space provided. This procedure is repeated for the remaining 49 stems. The mean percentage correct is calculated for each level of noise present in the speech sample [10].

The MRT test expands on the Fairbanks test and provides the listener with a limited list of rhyming words. The listener must decide which word on the list was spoken in the sample. The DRT expands even further on these previous methods and uses word-pairs, differing only by an initial consonant. The listener must select the

correct spoken word from the provided word pair. For both the MRT and DRT methods, the accuracy of selecting the correct word is computed, taking into account the effect of chance in choosing the correct answer [9].

2.1.1.3 Speech Intelligibility Percentage

Speech intelligibility refers to the accuracy of a listener understanding spoken content. This is measured by what the listener is able to detect under study and is calculated as the percentage of words the subject correctly identifies in a sample recording. This method proves to be reliable in measuring signal quality because environmental aspects can be altered to produce an atmosphere comparable to the listening situation. Experimental recordings can be created which contain the variety of factors which contribute to the overall noise experienced by the listener in an actual environment. This provides an accurate representation of the environment in which the listener will be required to detect spoken words/sentences, thus allowing for a better analysis of the intelligibility of the SOI in the signal [9].

2.2 Objective Methods of Measuring Audio Quality

Several objective methods have been derived which are used to provide a mathematically calculated estimate of the quality of speech samples. These normally require specific information about the original speech sample and are limited to controlled experiments in which this knowledge can be obtained. Several objective measures currently used are Signal-to-Noise Ratio, Linear Prediction based, Articulation Index models, Perceptual Evaluation of Speech Quality test, and Speech Intelligibility Index [9] [11] [12] [13] [14] [15] [16].

2.2.1.1 Signal-to-Noise Ratio (SNR)

Signal-to-Noise Ratio (SNR) is a commonly known objective measure of speech quality, comparing the power of the speech signal to the power of the noise signal. There are many variations to SNR measures, including segmental SNR and frequency-weighted SNR, but all require knowledge of the signal and noise power levels. While this calculation provides significant information regarding the quality of the speech signal, this objective method cannot be used in realistic scenarios where the speech signal is not initially separate from the noise signal. Also, this measure provides information regarding the quality of the speech signal rather than the intelligibility of the speech within the signal.

2.2.1.2 Linear Prediction (LP) Models

Linear Prediction (LP) models use linear prediction coefficients (LPC) to objectively measure the difference in quality between the original and distorted speech signals. Log-Likelihood Ratio (LLR), Itakura-Saito (IS), and Cepstrum Distance (CD) are all used to estimate the objective quality of distorted speech signals. Each of these measures requires knowledge of the original signal so the distance between the distorted signal and original signal can be calculated. This also proves to be of little use in realistic scenarios since the original speech signal cannot be obtained [9] [16].

2.2.1.3 Weighted Spectral Slope (WSS)

The Weighted Spectral Slope (WSS) measure is a weighted distance measure, at each frequency, between the spectral slopes of clean and distorted speech samples. In this measure, the spectral peaks are weighted more heavily, implying that spectral peaks

are more critical than the general spectral shape of the signal and have more of an effect on the perception of intelligibility [9] [16] [17] [12] [16] [17] [18].

2.2.1.4 Articulation Index (AI)

The Articulation Index is computed by averaging overall calculated SNR values across various frequency bands as shown in equation 2.1:

$$\boldsymbol{AI = \frac{1}{X}\sum_{j=1}^{X}\frac{min\{SNR(j),\ SNR_{max}\}}{30}} \qquad \textbf{2.1}$$

where *AI* is the articulation index, *X* is the number of subbands, *SNR(j)* is the j^{th} subband, and SNR_{max} is the maximum subband SNR level allowed. Weights can be applied to different frequency bands to account for the distortions present in each band. AI accurately estimates subjective quality as long as the distortions present in the signal come from either additive noise or signal attenuation; however, explicitly knowing the type(s) of noise present in the distorted signal creates an obstacle when dealing with realistic data [9]. Also, the original signal is needed in order to compute the SNR values for each frequency band, presenting problems in applying AI as a measure of signal quality in realistic scenarios.

2.2.1.5 Perceptual Evaluation of Speech Quality (PESQ)

The Perceptual Evaluation of Speech Quality (PESQ) is a widely accepted test, mimicking human perception, used to estimate the MOS of a signal by comparing differences between the original and distorted signals. Both signals are transformed to an internal representation equivalent to how the human auditory system perceives signals and then aligned with each other to account for any time delay due to degradation [14].

The difference between the representations is used to compute an estimated MOS which is then mapped onto the subjective MOS using a regression method specific to the particular study. This provides an objective MOS score which describes the quality of the signal after degradation has affected the original signal [12] [13].

In 2001, the International Telecommunication Union–Telecommunication Standardization Sector (ITU-T) established PESQ as an international standard for computing MOS in telephony applications. A PESQ test performed by ITU-T shows a strong correlation (r = 0.935) between the subjective MOS and the estimated objective MOS [9] [13]. Another study examining whether PESQ can be used as a measure of speech intelligibility also found high correlations (r=0.99, r=0.91) [10] [15]; however, minimal studies have been done outside of evaluating filtered telephone speech signals [16]. Studies have shown using PESQ to evaluate speech intelligibility of signals improved by noise suppression systems results in decreased correlation between objective scores and subjective scores [14] [12] [18]. This is believed to be the cause of an assortment of possible noise types creating the distortions rather than degradation due to solely additive noise.

2.2.1.6 Speech Intelligibility Index (SII)

A last objective measure of speech quality is an ANSI standard called the Speech Intelligibility Index (SII). SII values are computed using the individual power present in the speech and noise signal, and provides a value highly correlated with speech intelligibility [7]. The SII is an expansion of the AI, broadened to withstand degradations other than additive noise and signal attenuation only [19].

The SII is computed by dividing both the speech and noise signals into multiple bands and computing the SNR in each band. Each frequency band is weighted differently, giving more weight to bands contributing more to speech intelligibility. The SII measure is then calculated by averaging the weighted SNR values across all bands, resulting in a value between zero and one. An SII measure of zero represents no intelligibility while an SII of one indicates complete intelligibility [8].

Many studies show the validity of the SII method is limited to applications in which noise is stationary. For non-stationary noise sources, computing the SII over small sections of the speech sample and averaging over the entire sample results in an improved SII measure overall [8]. It is important to note that an experiment performed by Bronkhorst and Plomp (1992) showed as the number of interfering speech sources acting as noise increases, the resulting signal begins to portray stationary noise [20]. This can be applied to the cocktail party scenario in which multiple interfering sources act as a single noise signal.

While SII proves to be a widely used and accepted measure of speech intelligibility, calculating the SII requires knowledge of the power in the SOI, as well as the power in the interfering noise sources. Thus, for live experimental recordings, the SII is not helpful in computing intelligibility.

2.3 Limitations of Methods

The previously presented methods have produced accurate results when certain experimental conditions were satisfied while assessing audio quality. Major advancements have resulted from these methods, especially when using objective

methods to evaluate audio quality. However, limitations do exist, for both subjective and objective methods, which present problems when assessing intelligibility. For objective measures, many methods require knowledge of the clean speech signal before distortions degraded the recording. In live applications, this knowledge is normally not known and hard be obtain, resulting in obstacles and restrictions. For subjective measures, repeatability across experiments has shown limitations with current methods, as well as consistency among different subjects.

In the visual world, certain limitations similar to these existed for quantifying subjective perception. As a result of trying to understand and overcome these obstacles, the QR method was established. Seeing that this approach was implemented in the visual world and brought success in gaining an understanding of a subject's perception of quality, the notion developed that the same approach could be adapted for application in the audio realm and produce helpful insights. By using a standard to produce a predefined scale of intelligibility and requiring subjects to score test recordings according to the scale, the experimental focus of this thesis centered on providing a consistent measure of intelligibility over a wide range of subjects performing a subjective assessment and understanding the degree to which consistent measures can be obtained from this method.

CHAPTER 3: EXPERIMENT

3.1 Experimental Setup

The study was performed at the Davis Marksbury Building on the University of Kentucky campus. A computer and speakers were provided in the room the study took place in. Details of the equipment used are shown below in Table 3.1.

Table 3.1 - Details of equipment used in experiment.

Operating System	Windows 7
Sound Card	High Definition Audio Device
Speakers	BOSE Companion 2 Series II Multimedia Speaker System

Subjects participated in the study one at a time in order to keep a consistent environment for all subjects, as well as due to equipment limitations. The room where the experiment took place was approximately 10'x14' (149 sq. ft.) and was used by multiple research groups running experiments; however, the subject was the only person in the room for the duration of their participation in the experiment. The table at which each subject sat contained a computer monitor, the speakers, a keyboard, and a mouse. The placement of the equipment relative to the user is shown in Figure 3.1. The speakers were placed on each side of the monitor and approximately two feet from the subject.

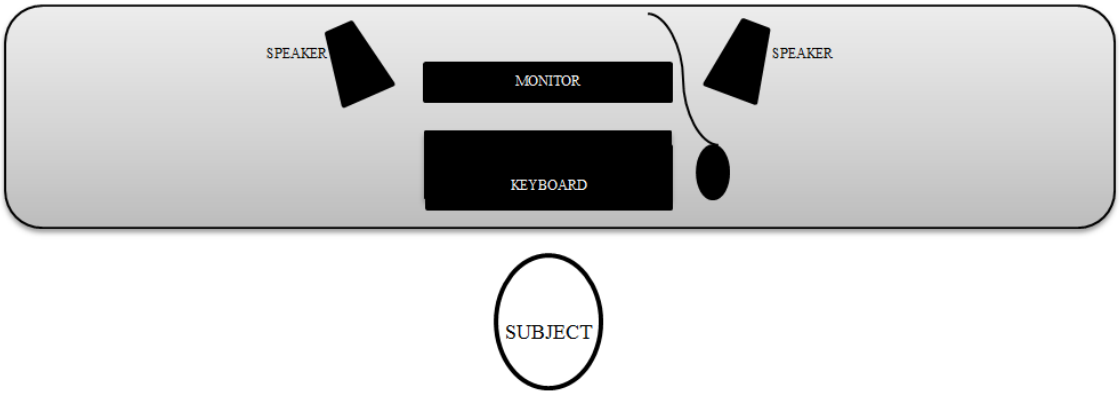


Figure 3.1 - Layout of equipment and subject used in experiment.

A total number of 24 subjects volunteered to participate in this study (five were female). The mean age of all subjects was 28.95, ranging from 20 to 75 years of age. All subjects spoke English; although, English might not have been their native language. There were two groups, A and B, with subjects split equally between the two groups.

Two methodologies were involved in the study – Single Stimulus Absolute Category Rating (SSACR), adapted from the MOS method, and Intelligibility Ruler (IR), adapted from the QR method. Subjects in Group A performed the SSACR method first followed by the IR method. Subjects assigned to Group B performed the IR method first and then the SSACR method. This was done to reduce any bias that may have occurred from the order of completing the methods. It took subjects approximately 20 minutes to complete both methods.

This experiment was run in accordance with IRB approval (IRB Approval Number: 11-1009). After agreeing to participate via signing the informed consent form (Appendix B: Consent Form), the subject was assigned to either group A or group B and

then given an identification number which was used to organize data cumulated throughout the experiment. The identification number was different for each subject and in no way linked data to the subject's name, but was only used to separate data between subjects. After the subject was assigned an ID number and group, they received oral and written instructions (Appendix C: Written Instructions for Both Methods) of what they were expected to do during the study. Since there were two methodologies involved in the study, the subject received the written procedure for only the method they were participating in at that time. After reading through the written procedure, they were free to ask any questions regarding the study. The subject was then provided a demonstration of the study and any questions were answered. The subject then performed the first method on their own, in which only questions and/or problems regarding the user interface were addressed. After performing the first method, the subject received the written and oral procedures, along with the demonstration for the next method and then performed the second method of the study on their own. The procedure for each method follows in sections 3.3 and 3.4.

3.2 Creating the Recordings

Test recordings (10) and reference recordings (14) were generated for use in the experiment. All recordings were created the same way; however, the reference recordings had a single SOI and background noise combination, while the test recordings had a single SOI, different than the reference SOI, paired with four different background noise sources. Seven available WAVE files were used as noise sources in the recordings (4 male and 3 female) to create the cocktail party background noise. All were sampled at 16 kHz and trimmed such that their lengths were eight seconds long.

Two of the sources were chosen as SOIs in the recordings. A male was chosen as the SOI in the test recordings and a female was chosen as the SOI in the reference recordings to prevent any potential bias resulting from using same gender SOI sources. Table 3.2 shows the separate sentences spoken by the SOI in each of the recordings.

Table 3.2 - Sentences spoken by the separate SOIs in the test recordings and reference recordings

SOI	Sentence Spoken
Test Recording SOI - Male	"In my bedroom there's a baseball bat, a blue jay hat, a hairbrush."
Reference Recording SOI – Female	"Railroads are for catching trains. Sidewalks should be kept clean in winter."

To create the reference and test recordings, a SOI was combined with a noise signal consisting of multiple interfering speech sources. All signals were first scaled so their power levels were the same for consistent playback between WAVE files. A particular SOI signal was then weighted at various levels and combined with a noise signal, resulting in a cocktail party recording at a specific SII level.

The five remaining sources were added together in different ways to create four separate noise signals for use in the test recordings. These noise recordings contained both male and female sources and either three or four sources made up each noise recording. For the reference noise signal in the reference recording, all five sources were used to create the noise signal. Seven total WAVE files existed after the noise sources were configured and the SOI samples were chosen (2 SOI signals and 5 noise signals). These signals were then altered such that the power in each signal was normalized to a level of 0.300. The reference noise signal was normalized to 0.100 to allow the reference

recordings to span the entire SII scale. Scaling the signals to these values allowed the SII values computed from the SOI-noise pair to cover a wider range of the SII scale. The procedure for each signal was performed using MATLAB (*computeRMS.m*) and follows:

1. Read the signal into MATLAB.
2. Remove silence in the signal using the attached MATLAB file, *rmsilence.m*
3. Compute the RMS value of the signal using the MATLAB function, *std()*.
4. Normalize the signal by dividing the original signal (with silence) by the root mean square (RMS) of the signal.
5. Divide the normalized signal by the absolute maximum value in the normalized signal.
6. Scale the signal, by multiplying by 0.300, so the max power in the signal is equal to 0.300 (0.100 for the reference noise signal).
7. Write the scaled, normalized signal to a WAVE file.

The silence is removed before computing the power in the signal to ensure the RMS value is indicative of the density of the speaker's speech pattern. Considering a long period of silence present in the signal will result in a lower power rating of the source, even if the source is speaking very loudly during times of talking. Removing the silence allows a better comparison of the power levels between two signals, as well as reduces deviation in the computed SII values.

After each WAVE file was normalized to a constant value the SOI signals were combined with the noise signals, weighting the SOI signals in order to obtain different SII values of the recordings. In order to determine the weights needed for each SOI/noise

combination, the attached MATLAB file *weightsSOI.m* was used. This script requires the user to select the desired SOI WAVE file and noise WAVE file. The algorithm reads in both signals and removes any leading/trailing zeros on the SOI signal. A 5^{th} order high-pass Butterworth filter with a cutoff frequency of 100 Hz is applied to both signals in order to reduce room noise. Any silent intervals are removed from the signals to allow for improved calculated SII values.

By default, a vector of weights spanning from 0 to 500 by 0.01 increments is used. The SOI signal is then scaled by each weight and the SII is computed over 100 ms intervals to account for the non-stationarity of speech. The average SII value over all intervals is calculated and assigned as the SII value of the SOI-noise combination. Each individual weight produces a different average SII value. Also, the weights which create certain SII values will differ between different SOI/noise combinations. A MAT file is created with the same filename as the SOI WAVE filename that contains a matrix containing SII values in the first column and the corresponding weight in the second column.

The weights for each SOI-noise combination had to be determined in order to create the test and references recordings which had different SII values. Thus, the above procedure was carried out five times (test recording SOI with each of the four noise signals and then the reference recording SOI with its noise signal) in order to obtain the different weights associated with each SII value for each SOI-noise pair. Ten SII values, concentrating around the SII critical value, were chosen from the four SOI-noise pairs to make up the test recordings. The SOI-noise pairs are shown in Table 3.3, along with the

weights to apply to the SOI signal to obtain a specific SII value. The corresponding individual and pair WAVE files are attached.

Table 3.3 – SOI-Noise Pairs for Test Recordings with Corresponding SII Values and Scaling Weights

Test Recording Number	SOI - Noise Pair	Actual SII Value	Weight
1	SOI1 / Noise1	0.10	0.130
2	SOI1 / Noise2	0.15	0.220
3	SOI1 / Noise3	0.20	0.400
4	SOI1 / Noise4	0.20	0.460
5	SOI1 / Noise2	0.25	0.460
6	SOI1 / Noise4	0.30	0.810
7	SOI1 / Noise3	0.35	0.930
8	SOI1 / Noise1	0.40	0.890
9	SOI1 / Noise4	0.45	1.710
10	SOI1 / Noise1	0.50	1.480

To obtain the weights for the reference recordings, program *weightsSOI.m* was carried out once between the female SOI and the reference noise signal. Weights for the SII values corresponding to the layout of the IR, shown later in Figure 3.3 in section 3.4, were chosen from the results. The SOI-noise pairs for the reference recordings are shown in Table 3.4, along with the weights to scale the SOI signal in order to obtain specific SII values. The corresponding individual and pair WAVE files are attached.

Table 3.4 - SOI-Noise Pairs for Reference Recordings with Corresponding SII Values and Scaling Weights

Reference Recording SII Value	SOI - Noise Pair	Weight
0.00	SOI2 / Noise5	0.000
0.10	SOI2 / Noise5	0.150
0.15	SOI2 / Noise5	0.240
0.20	SOI2 / Noise5	0.340
0.25	SOI2 / Noise5	0.470
0.30	SOI2 / Noise5	0.610
0.35	SOI2 / Noise5	0.780
0.40	SOI2 / Noise5	0.990
0.50	SOI2 / Noise5	1.580
0.60	SOI2 / Noise5	2.540
0.70	SOI2 / Noise5	4.270
0.80	SOI2 / Noise5	8.050
0.90	SOI2 / Noise5	35.830
1.00	SOI2 / Noise5	used SOI2 signal by itself

**Note: for an SII value of 0.90, the weight of 35.830 provided an SII value of 0.8907 which was then rounded up to 0.90.*

Once these weights were determined, the MATLAB script *testscript_intel.m* was used to create the test and reference recordings with the desired SII values. This script required the user to select the desired SOI WAVE file and noise WAVE file. The algorithm read in both signals and removed any leading/trailing zeros on the SOI signal. A 5th order high-pass Butterworth filter with a cutoff frequency of 100 Hz was applied to both signals in order to reduce room noise. Any silent intervals were removed from the signals to allow for improved calculated SII values. The user was then prompted to enter in the weight(s) to apply to the SOI signal.

The average SOI intelligibility was computed using the ANSI S3.5 1997 standard for calculating SII levels [7]. The script called the attached function *intel.m* which divided the SOI and noise signals into 100 ms segments and used *spectrumlevel.m* to estimate the individual spectrum power levels over 18 bands. The segments overlap by 50% to for a more accurate assessment. The spectrum levels were then passed to the program *sii.m*, written by Hannes Müsch, to compute the SII level over each segment to account for varying levels of intelligibility over a speech sample [21]. The overall SII value for the SOI-noise pair was then computed by averaging SII values over all segments of the signal. Once the average SII value was computed, the weighted SOI and noise signals were combined and written to a WAVE file for use in the experiment.

3.3 SSACR Method Procedure

For the SSACR method of analysis, the subject was first prompted to enter their assigned ID number. As stated above, this number was only used to separate data stored for each subject and was not used to identify any subject. During this method, the subject was presented a user interface that displayed a subjective scale, playback and stop buttons, and a score button, as illustrated in Figure 3.2 . The subject was instructed to press the button labeled ‘Speaker of Interest Sample’ to listen to the clean SOI recording without interfering background sources. They were then instructed to press the ‘Play’ button to listen to the test recording under question and then score it using the radio buttons.

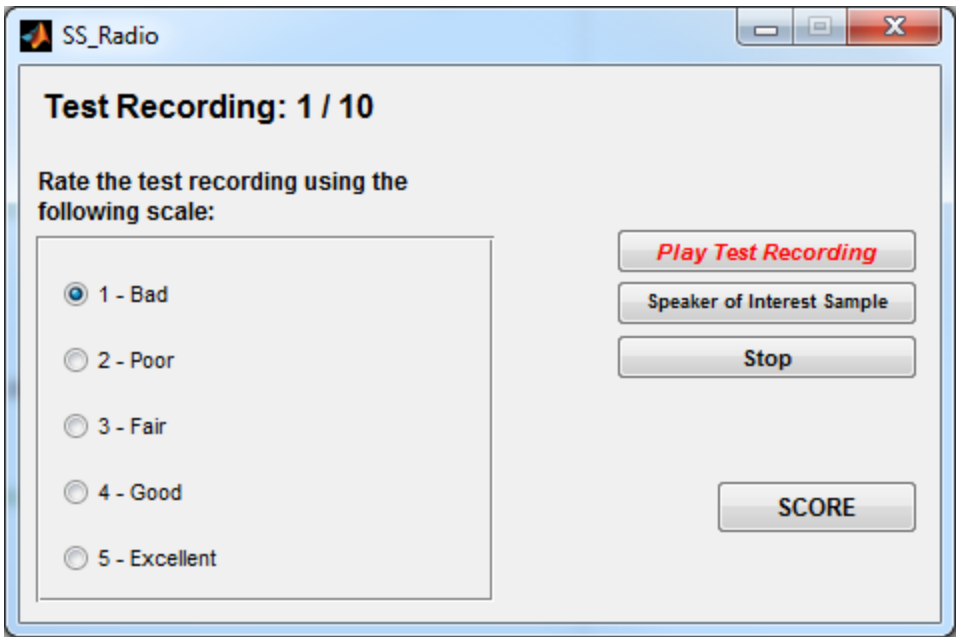


Figure 3.2 - User Interface for SSACR Method of Analysis

The test recordings simulated a cocktail party scenario, in which multiple sources speak simultaneously. Each test recording presented the same SOI talking, while different interfering noise sources acting as background noise played in the clip. The subject was required to rate the intelligibility of the SOI in the recording by selecting the button corresponding to the value of their choice. The scale is represented in Table 3.5.

Table 3.5 - Subjective scoring scale used in SSACR method

Numerical Rating	Categorical Scale
1	Bad
2	Poor
3	Fair
4	Good
5	Excellent

A score of 1corresponds to the SOI being completely inaudible; whereas, a score of 5 means that the SOI can be heard clearly with no questions regarding the intelligibility of the sentences spoken by the SOI. The scale was not continuous and required the subject to distinctively choose between subjective values. The subject was not provided any references as to what constituted 'Bad' or 'Excellent' in terms of SOI intelligibility, therefore, each subject judged SOI intelligibility based on their own interpretation of the subjective adjectives given.

Once the subject selected the radio button corresponding to the rating of their choice, they were instructed to select the 'Score' button to record the score for the given sample recording. The next recording sample was then made available for the subject to listen to and rate. This process repeated until the subject had successfully rated all 10 test recording samples in the SSACR portion of the study.

3.4 IR Method Procedure

For the IR method, the subject was first prompted to enter their assigned ID number. As stated before, this number was only used to separate data stored for each subject and was not used to as a means of subject identification. During this method, the subject was presented a user interface that displayed a ruler of reference recordings corresponding to SII values ranging from 0.0 to 1.0, playback and stop buttons, a text box to manually enter SII score ratings, and a score button, as shown in Figure 3.3.

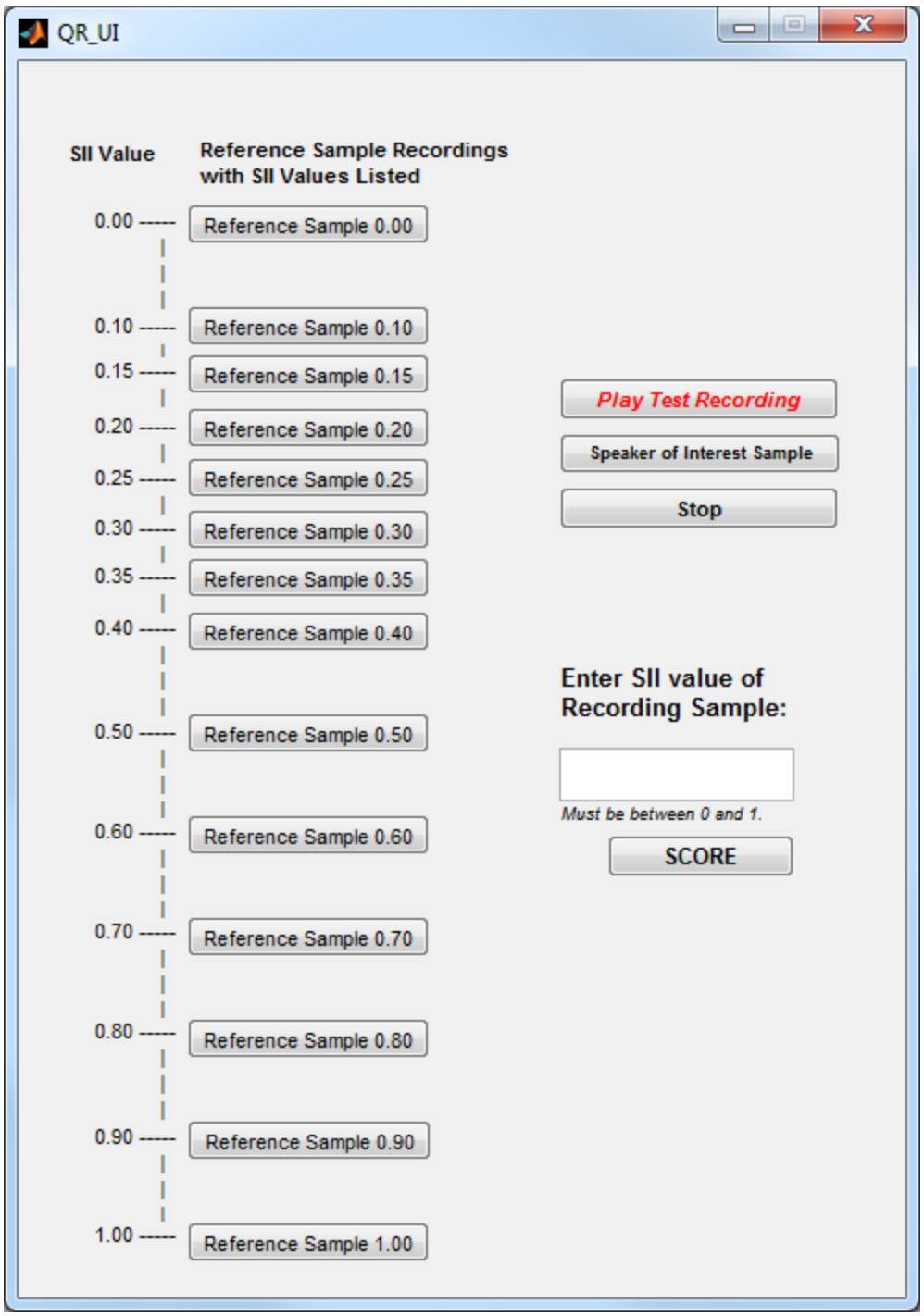


Figure 3.3- User Interface for IR Method of Analysis

The IR was pre-calibrated and divided into 14 sections with interval markings every 0.10 SII measure between 0.00 and 1.00. Between SII values of 0.10 and 0.40, interval markings appeared every 0.05 SII measure. The IR was divided in this manor

because intelligibility has a critical transition from mostly unintelligible to mostly intelligible with careful listening around a value of 0.30. After a SII value of approximately 0.60, the intelligibility of the SOI tends to level off and is deemed excellent [11] [22] [23].

Sample reference recordings containing a SOI were placed at every interval marking, with the SOI in the recording possessing the corresponding marked SII value. These reference recordings were obtained using the procedure in section 3.2. All reference recordings had the same SOI and noise sources acting as background conversation noise; however, the intelligibility of the SOI differed depending on the related SII level. The subject was instructed to press the button labeled 'Speaker of Interest Sample' to listen to the clean SOI recording without interfering background sources. They were to then press the 'Play' button to listen to the test recording under question and then compare the intelligibility of the SOI in the test recording to the intelligibility of the SOI in the reference sample recordings. Using the adjacent SII scale, the subject was instructed to rate the intelligibility of the SOI in the test recording using the ruler as a means of evaluation. The subject was also instructed to listen to all reference recordings for each SII level before rating the test recording, encouraging a comparative decision to be made.

To rate the test recording, the subject was instructed to enter the corresponding SII value for the reference sample they chose to best describe the intelligibility of both SOI samples into the text box on the user interface and press the 'Score' button to record the SII value for the given test recording. Upon pressing 'Score', the next test recording

was made available for the subject to listen to and rate. This process was repeated until the subject had successfully rated all test recordings in the IR portion of the study.

Since this method was more intricate than the SSACR method, the subject was allowed to first get accustomed to the IR scale via a demonstration mode. Once the subject felt comfortable with the interface and IR scale, they exited the demonstration mode and begin the actual experiment.

3.5 Analysis of Data

After all subjects performed both methods, a population mean and standard deviation were computed for each test recording. The mean was calculated for each test recording by taking the arithmetic mean of all subject ratings, excluding any outliers. For the SSACR method, the scores were normalized using equation 3.1:

$$SII_{SSACR} = \frac{X - 1}{4} \qquad \textbf{3.1}$$

where X is the score for the recording given by the subject using the SSACR method. This provided a score which ranged between 0.00 and 1.00 so comparisons could be made with the IR method. Therefore, initially a scoring of 5 corresponded to 1.00 and a scoring of 1 corresponded to 0.00 on the IR scale. Since the possible score values ranged from 0.00 to 1.00, the MOS and standard deviation for each sample was also in this range. The same test recordings were used for both methodologies; therefore, analysis was done which compared the resulting MOSs, standard deviations, and variances between the two methods.

In order to account for arbitrary scale and biases of the subjective ratings, a linear mapping was performed between the normalized ratings and the true SII values, as shown

below in equation 3.2. The method of least squares was used to perform this transformation, to minimize the squared errors between the observed data (SSACR subject scores) and the expected values (actual SII values). The following steps were taken in order to apply the linear transformation:

1. Compute **μ**, the mean of the vector containing the actual SII values.
   Compute **α**, the mean of the vector containing the subject scores for a particular test recording.
2. Make both **X** and **Y** zero mean vectors, where **X** contains the actual SII values of the test recordings and **Y** contains the individual subject scores using the SSACR method.
3. Compute the scaling factor, **β**.
4. Apply the scaling factor, **β**, to the individual scores in **Y**.
5. Restore μ to **Y** in order to shift the subject scores towards the SII values.

The final linear transformation equation is shown in equation 3.2 below and the derivation is included in the appendix (Appendix D: Linear Transformation):

$$\boldsymbol{SII_{new} = \beta(SII_{SSACR} - \alpha) + \mu} \qquad \textbf{3.2}$$

where SII_{new} is the SII score after applying the transformation and SII_{SSACR} is the score directly after normalizing the subject responses. A β equal to one (or close to one) describes a relative SII score which is near the actual SII value. The difference between α and μ describes the shift in the data towards the real SII value. A difference of zero (or close to zero) suggests no shift is needed and the subject score is near the actual SII value. Therefore, a β of 1 and a difference between α and μ equal to 0 is desired because the subject score would be the exact SII value.

Analysis was performed on both sets of data by examining the standard deviation and variance for each test recording, as well as the average standard deviation and variance for each method over all test recordings. T-tests were performed with 95% confidence for each test recording in order to determine the outcome of our null hypothesis, where the null hypothesis stated the mean of the population is equal to the actual SII value. A paired-sample t-test was performed with 95% confidence to determine whether any statistical difference between the two methods existed.

CHAPTER 4: RESULTS AND DISCUSSION

Linear mapping was applied to all scores acquired by subjects in both methods. We found that the scaling factor, β, in the IR method was approximately equal to unity (1.0051), demonstrating that subjects were able to match intelligibility levels between two different recordings and also distinguish the SII values between a sample and a series of references. A deviation from 1 can be contributed to experimental variations between subjects participating, as well as any systematic error that may be present. Therefore, further analysis requiring the linear transformation shown in equation 3.2 pertained to only the SSACR method.

4.1 SSACR Method

The data for the SSACR method before the transformation and with outliers removed is applied is shown in Table 4.1. Subject scores were considered outliers if they were more than ±3 standard deviations away from the population mean for that particular test recording. Outliers were determined for each test recording and not by subject. For example, if the score Subject 4 recorded for Test Recording 1 was considered an outlier, Subject 4's responses for the other 9 test recordings were kept and included in the calculations unless any of the remaining scores were also considered outliers. After normalization and before applying the transformation, the categories are separated by a score of 0.25 (i.e., 0, 0.25, 0.50, 0.75, and 1). While the categorical scale evenly describes intelligibility, the normalized scores span a wide range of the scale associated with intelligibility. For example, looking at Test Recording 10 we see the actual SII value for this recording is 0.50, a value typically corresponding to correct identification of all words with moderate listening effort based on listening efforts. However, the mean

normalized SII score (before applying the transformation) is recorded as 0.8021 ± 0.165. This range [0.637, 0.967] does not encompass the actual SII value of 0.5. Along with some ranges not encompassing the actual SII value, mean values reached upwards of approximately 0.8; whereas, the highest actual SII value in the test recordings was 0.5. Since a reference was not used for the test, deriving a linear transformation to map the categorical scale to the SII scale may result in better outcomes for use in comparisons.

Table 4.1 - Data from SSACR Method for All Subjects (with outliers removed) before Transformation Applied

SSACR Method before Transformation (outliers removed)				
Test Rec	**Actual SII**	**Mean of Scores**	**Standard Deviation**	**Variance**
1	0.10	0.0000	0.0000	0.0000
2	0.15	0.0761	0.1397	0.0195
3	0.20	0.2283	0.1833	0.0336
4	0.20	0.2174	0.1893	0.0358
5	0.25	0.0870	0.1217	0.0148
6	0.30	0.5000	0.2085	0.0435
7	0.35	0.5417	0.2170	0.0471
8	0.40	0.5938	0.2188	0.0479
9	0.45	0.7292	0.2074	0.0430
10	0.50	0.8021	0.1645	0.0271

Linear transformations were applied to the data using exclusion sets in order to determine whether a linear transformation over the entire data set would produce valid results. Four separate sets of data were grouped together as 'Exclusion Sets', which are shown in Table 4.2.

Table 4.2 - Exclusion Sets

Exclusion Set	Subjects in Set
1	Even Subjects
2	Odd Subjects
3	First Half of Subjects
4	Last Half of Subjects

Exclusion Set 1 excludes all members in Exclusion Set 2, and vice versa. The same applies for Exclusion Sets 3 and 4. Using only subjects in Exclusion Set 1, transformation factors were computed, as shown in Table 4.3 in the row labeled for Exclusion Set 1.

Table 4.3 - Transformation factors derived from each Exclusion Set

Exclusion Set	Transformation Factors		
	β *(scaling factor)*	α *(mean of subject scores)*	μ *(mean of actual SII values)*
1	0.4628	0.3396	0.2900
2	0.4510	0.4438	0.2900
3	0.4539	0.3917	0.2900
4	0.4356	0.3835	0.2900

Inserting these factors into equation 3.2, the transformation was applied only to subjects in Exclusion Set 2 (not members of Exclusion Set 1). Table 4.4 and Table 4.5 show data for Exclusion Set 2 before and after the transformation was applied, respectively. Note the same trend in decreasing standard deviation and variance for each test recording. When averaging over all variances in Table 4.5, a small value of 0.0062 is computed.

Table 4.4 - Exclusion Set 2 before transformation was applied

Exclusion Set 2 before Transformation				
Test Rec	**Actual SII**	**Mean of Scores**	**Standard Deviation**	**Variance**
1	0.10	0.0000	0.0000	0.0000
2	0.15	0.1250	0.1685	0.0284
3	0.20	0.2708	0.1982	0.0393
4	0.20	0.2083	0.1794	0.0322
5	0.25	0.1250	0.1306	0.0170
6	0.30	0.5625	0.1884	0.0355
7	0.35	0.5417	0.2087	0.0436
8	0.40	0.5833	0.1628	0.0265
9	0.45	0.7708	0.2251	0.0507
10	0.50	0.7292	0.1287	0.0166

Table 4.5 - Exclusion Set 2 after transformation was applied

Exclusion Set 2 after Transformation				
Test Rec	**Actual SII**	**Mean of Scores**	**Standard Deviation**	**Variance**
1	0.10	0.1328	0.0000	0.0000
2	0.15	0.1907	0.0780	0.0061
3	0.20	0.2582	0.0917	0.0084
4	0.20	0.2293	0.0830	0.0069
5	0.25	0.1907	0.0604	0.0037
6	0.30	0.3932	0.0872	0.0076
7	0.35	0.3835	0.0966	0.0093
8	0.40	0.4028	0.0754	0.0057
9	0.45	0.4896	0.1042	0.0109
10	0.50	0.4703	0.0596	0.0035

Using only subjects in Exclusion Set 2, transformation factors were computed, as shown in Table 4.3 in the row labeled for Exclusion Set 2. Inserting these factors into equation 3.2, the transformation was applied only to subjects in Exclusion Set 1 (not members of Exclusion Set 2). Table 4.6 and Table 4.7 show data for Exclusion Set 1 before and after the transformation was applied, respectively. Note the same trend in decreasing standard deviation and variance for each test recording between Table 4.6 and Table 4.7, as was observed Table 4.4 and discussed previously. When averaging over all

variances in Table 4.7, a value of 0.0059 is computed. This is within 0.0003 of the average variance computed for Exclusion Set 2.

Table 4.6 - Exclusion Set 1 before transformation was applied

Exclusion Set 1 before Transformation				
Test Rec	**Actual SII**	**Mean of Scores**	**Standard Deviation**	**Variance**
1	0.10	0.0000	0.0000	0.0000
2	0.15	0.1250	0.1685	0.0284
3	0.20	0.2708	0.1982	0.0393
4	0.20	0.2083	0.1794	0.0322
5	0.25	0.1250	0.1306	0.0170
6	0.30	0.5625	0.1884	0.0355
7	0.35	0.5417	0.2087	0.0436
8	0.40	0.5833	0.1628	0.0265
9	0.45	0.7708	0.2251	0.0507
10	0.50	0.7292	0.1287	0.0166

Table 4.7 - Exclusion Set 1 after transformation was applied

Exclusion Set 1 after Transformation				
Test Rec	**Actual SII**	**Mean of Scores**	**Standard Deviation**	**Variance**
1	0.10	0.0899	0.0000	0.0000
2	0.15	0.1462	0.0760	0.0058
3	0.20	0.2120	0.0894	0.0080
4	0.20	0.1838	0.0809	0.0065
5	0.25	0.1462	0.0589	0.0035
6	0.30	0.3436	0.0850	0.0072
7	0.35	0.3342	0.0941	0.0089
8	0.40	0.3530	0.0734	0.0054
9	0.45	0.4375	0.1015	0.0103
10	0.50	0.4187	0.0581	0.0034

Using only subjects in Exclusion Set 3, transformation factors were computed, as shown in Table 4.3 in the row labeled for Exclusion Set 3. Inserting these factors into equation 3.2, the transformation was applied only to subjects in Exclusion Set 4 (not members of Exclusion Set 3). Table 4.8 and Table 4.9 show data for Exclusion Set 4 before and after the transformation was applied, respectively. Note the same trend in

decreasing standard deviation and variance for each test recording between Table 4.8 and Table 4.9, as was observed and discussed for the previous Exclusion Sets. When averaging over all variances in Table 4.9, a value of 0.0060 is computed. This is within 0.0002 of the variances recorded for previous two Exclusion Sets.

Table 4.8 - Exclusion Set 4 before transformation was applied

Exclusion Set 4 before Transformation				
Test Rec	**Actual SII**	**Mean of Scores**	**Standard Deviation**	**Variance**
1	0.10	0.0000	0.0000	0.0000
2	0.15	0.1250	0.1685	0.0284
3	0.20	0.2708	0.1982	0.0393
4	0.20	0.2083	0.1794	0.0322
5	0.25	0.1250	0.1306	0.0170
6	0.30	0.5625	0.1884	0.0355
7	0.35	0.5417	0.2087	0.0436
8	0.40	0.5833	0.1628	0.0265
9	0.45	0.7708	0.2251	0.0507
10	0.50	0.7292	0.1287	0.0166

Table 4.9 - Exclusion Set 4 after transformation was applied

Exclusion Set 4 after Transformation				
Test Rec	**Actual SII**	**Mean of Scores**	**Standard Deviation**	**Variance**
1	0.10	0.1122	0.0000	0.0000
2	0.15	0.1690	0.0765	0.0059
3	0.20	0.2352	0.0900	0.0081
4	0.20	0.2068	0.0814	0.0066
5	0.25	0.1690	0.0593	0.0035
6	0.30	0.3675	0.0855	0.0073
7	0.35	0.3581	0.0947	0.0090
8	0.40	0.3770	0.0739	0.0055
9	0.45	0.4621	0.1022	0.0104
10	0.50	0.4432	0.0584	0.0034

Using only subjects in Exclusion Set 4, transformation factors were computed, as shown in Table 4.3 in the row labeled for Exclusion Set 4. Inserting these factors into equation 3.2, the transformation was applied only to subjects in Exclusion Set 3 (not

members of Exclusion Set 4). Table 4.10 and Table 4.11 show data for Exclusion Set 3 before and after the transformation was applied, respectively. Note the same trend in decreasing standard deviation and variance for each test recording between Table 4.10 and Table 4.11 as was observed and discussed for the previous Exclusion Sets. When averaging over all variances in Table 4.11 a value of 0.0055 is computed. This is within 0.0007 of the average variances computed for all other Exclusion Sets.

Table 4.10 - Exclusion Set 3 before transformation was applied

Exclusion Set 3 before Transformation				
Test Rec	**Actual SII**	**Mean of Scores**	**Standard Deviation**	**Variance**
1	0.10	0.0000	0.0000	0.0000
2	0.15	0.1250	0.1685	0.0284
3	0.20	0.2708	0.1982	0.0393
4	0.20	0.2083	0.1794	0.0322
5	0.25	0.1250	0.1306	0.0170
6	0.30	0.5625	0.1884	0.0355
7	0.35	0.5417	0.2087	0.0436
8	0.40	0.5833	0.1628	0.0265
9	0.45	0.7708	0.2251	0.0507
10	0.50	0.7292	0.1287	0.0166

Table 4.11 - Exclusion Set 3 after transformation was applied

Exclusion Set 3 after Transformation				
Test Rec	**Actual SII**	**Mean of Scores**	**Standard Deviation**	**Variance**
1	0.10	0.1229	0.0000	0.0000
2	0.15	0.1774	0.0734	0.0054
3	0.20	0.2409	0.0863	0.0075
4	0.20	0.2137	0.0782	0.0061
5	0.25	0.1774	0.0569	0.0032
6	0.30	0.3680	0.0821	0.0067
7	0.35	0.3589	0.0909	0.0083
8	0.40	0.3770	0.0709	0.0050
9	0.45	0.4587	0.0980	0.0096
10	0.50	0.4406	0.0561	0.0031

By performing the four tests discussed previously, we were able to support the idea that these factors are generalizable over broader populations. In each of the previous four exclusion cases, the percentage difference between the computed average variances of all Exclusion Sets was 11.97%. A one-way ANOVA test was performed to test whether there was any variation between the four Exclusion Sets. The null hypothesis, shown in equation 4.1, states that the population means of each Exclusion Set are all equal.

$$\boldsymbol{H_0}: \ \boldsymbol{\mu_i = \mu_2 = \mu_3 = \mu_4} \qquad \textbf{4.1}$$

The resulting p-value of the ANOVA test was 0.865 suggesting that that no variation exists between the Exclusion Sets. A p-value less than of .05 was considered for rejecting the null hypothesis. Therefore, using the entire data set to obtain transformation factors using the least squares method and applying the linear transformation to the data set presents no problems in producing valid results.

For the SSACR method, a scaling factor of β=0.4400 was found using the entire data set. Using the values in Table 4.12, the transformation in equation 3.2 was applied to the data attained in the SSACR method.

Table 4.12 – Transformation Factors derived from All Subjects

Transformation Factor	**Value**
β (scaling factor)	0.4400
α (mean of subject scores)	0.3775
μ (mean of actual SII values)	0.2900

Table 4.13 illustrates the mean, standard deviation, and variance of each test recording after the transformation was applied to subject data. Note the standard deviations and variances for each test recording in Table 4.13 are smaller than those

presented in Table 4.1; therefore, after applying the transformation to calibrate the scale, the subjective scoring more accurately follows the SII values of the test recording. The mean values in Table 4.13 more closely represent the corresponding actual SII values for each test recording, proposing that data are more accurate. The standard deviations for each test recording are smaller after applying the transformation, indicating high precision between subjects.

Table 4.13 - Data from SSACR Method for All Subjects (with outliers removed) after Transformation Applied

SSACR Method after Transformation (outliers removed)				
Test Rec	**Actual SII**	**Mean of Scores**	**Standard Deviation**	**Variance**
1	0.10	0.1239	0.0000	0.0000
2	0.15	0.1574	0.0615	0.0038
3	0.20	0.2243	0.0807	0.0065
4	0.20	0.2195	0.0833	0.0069
5	0.25	0.1621	0.0536	0.0029
6	0.30	0.3439	0.0917	0.0084
7	0.35	0.3622	0.0955	0.0091
8	0.40	0.3851	0.0963	0.0093
9	0.45	0.4447	0.0913	0.0083
10	0.50	0.4768	0.0724	0.0052

Averaging over all standard deviations and variances associated to each test recording, an overall average standard deviation of 0.0726 and average variance of 0.0060 was calculated. Such a small standard deviation and variance, in comparison to the scale being used (0-1), suggests that the SSACR method can be used to show promising results as long as a linear transformation is applied to the data. Since the transformation factors were obtained from the data itself, further steps using the Exclusion Sets proved the results were valid and the SSACR method was accurate and precise (reliable and reproducible between subjects).

A one-sample t-test was performed for each test recording to test a null hypothesis, with the results shown in Table 4.14. The null hypothesis is shown in equation 4.2 and states that the scaled mean of the subject scores for Test Recording *n* is equal to the actual SII value for Test Recording *n*.

$$H_0: \ \mu_i = SII_i \qquad 4.2$$

The mean, standard deviation, t-statistic, and 95% confidence interval are shown, as labeled in Table 4.14, for each test recording. The final column in the table states whether the null hypothesis, H_o, was rejected or not. The null hypothesis is rejected for a test recording if the 95% confidence interval does not include the actual SII value for that test recording. Figure 4.1 displays an overall clearer representation of the results derived from the t-test.

Table 4.14 – Results from one-sample t-test performed for SSACR Method

T-Test (α=0.05) for SSACR Subject Data after Transformation with Outliers Removed						
Test Rec	**Actual SII**	**Mean**	**Std. Dev.**	**T**	**95% CI**	**Reject or Fail to Reject H_o**
1	0.10	0.1239	0.000	1.6E+15	**(0.124,0.124)**	*Reject*
2	0.15	0.1574	0.061	0.574	(0.131,0.184)	Fail to Reject
3	0.20	0.2243	0.081	1.446	(0.189,0.259)	Fail to Reject
4	0.20	0.2195	0.083	1.125	(0.184,0.256)	Fail to Reject
5	0.25	0.1621	0.054	-7.866	**(0.139,0.185)**	*Reject*
6	0.30	0.3439	0.092	2.343	**(0.305,0.383)**	*Reject*
7	0.35	0.3622	0.095	0.627	(0.322,0.403)	Fail to Reject
8	0.40	0.3851	0.096	-0.756	(0.344,0.426)	Fail to Reject
9	0.45	0.4447	0.091	-0.283	(0.406,0.483)	Fail to Reject
10	0.50	0.4768	0.072	-1.570	(0.446,0.507)	Fail to Reject

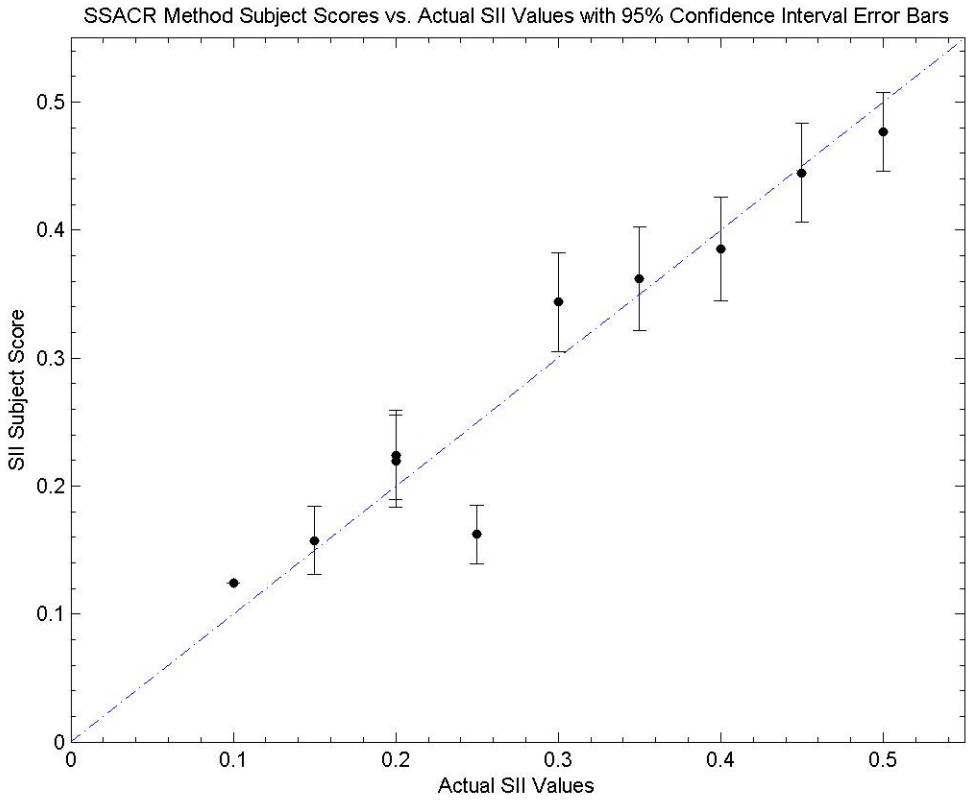


Figure 4.1 – Mean values from SSACR Method with Error Bars showing 95% Confidence Limit

In Figure 4.1, the x-axis corresponds to the actual SII value of the test recording and the y-axis represents the mean score computed from all subject responses. The population means for each test recording are denoted by the filled in circles (*Note: There are two test recordings located at an actual SII value of 0.2*). The error bars for each mean value show the 95% confidence limit derived from the t-test. The dotted line across the plot illustrates perfect subject score responses, having a slope equal to one. The overlap of the error bars with the dotted line shows that the SSACR method produced results near the desired outcome, with the exception of the test recordings with actual SII values at 0.10, 0.25 and 0.30. It is also noted that the confidence limits corresponding to

the two recordings possessing actual SII values of 0.2 overlap very closely, further supporting the precision observed between subjects performing this method.

The outcome of the t-test for the SSACR method suggests that the method is somewhat reliable in predicting SII scores of test recordings with unknown values. Results of the t-test show rejection of the null hypothesis for three of the ten recordings. By looking at which recordings were rejected in, it is seen that two of the three rejections were located in the critical region of intelligibility levels. All subjects (with the exception of one outlier) rated Test Recording 1 as 'Bad', which was mapped to a score of 0.1239. Since there was no variation in subject responses, the confidence interval only contained one value which is not the actual SII value of the recording. At a level of 0.10, the SOI is deemed unintelligible by SII standards; thus, a scoring of 'Bad' was a valid response for this recording and is rejected due to the values of the transformation factors used for the data set.

Both Test Recording 5 and Test Recording 6 were rejected. The actual values for these two recordings fall around the critical range of SII values, between 0.25 and 0.30. SII values between 0.25 and 0.30 generally describe the SOI as slightly intelligible where most words can be recognized with a high level of listening effort. At SII values above 0.30, subjects deem the sample as intelligible and are able to hear almost all words spoken by the SOI with moderate effort. For Test Recording 6, the average subject score for the recording was 0.3439 ± 0.092, an overestimate of the actual SII value of 0.30. This is expected since at SII values of 0.30, the subject begins to hear all words spoken by the SOI.

For Test Recording 5, the average score for the recording is 0.1621 ± 0.054, an underestimate of the actual SII value of 0.25. This is expected since speech only starts becoming slightly intelligible around a value of 0.25. Due to the limited options on the subjective scale used for the SSACR method, it is expected that the scores of recordings possessing SII values in this critical area be somewhat erroneous; however, since the errors fall in the expected directions, the SSACR method proves to be accurate. With further improvements, such as increasing the number of options on the subjective scale, issues concerning the validity of data in the critical region may be diminished.

Using Excel, a line of best fit was applied to the mean subject scores with a y-intercept at zero as shown in Figure 4.2. The resulting R^2 value of 0.9181indicates a high level of precision over the population. The slope of the line of best fit is equal to 0.9878, a value very close to 1. A slope of 1 indicates perfect subject responses. The computed value of the slope indicates a high level of accuracy between subjects performing the test. Linear regression analysis was performed on the entire data set and an R^2 value of 0.6578 was computed, indicating large variance between subjects, as shown in Figure 4.3.

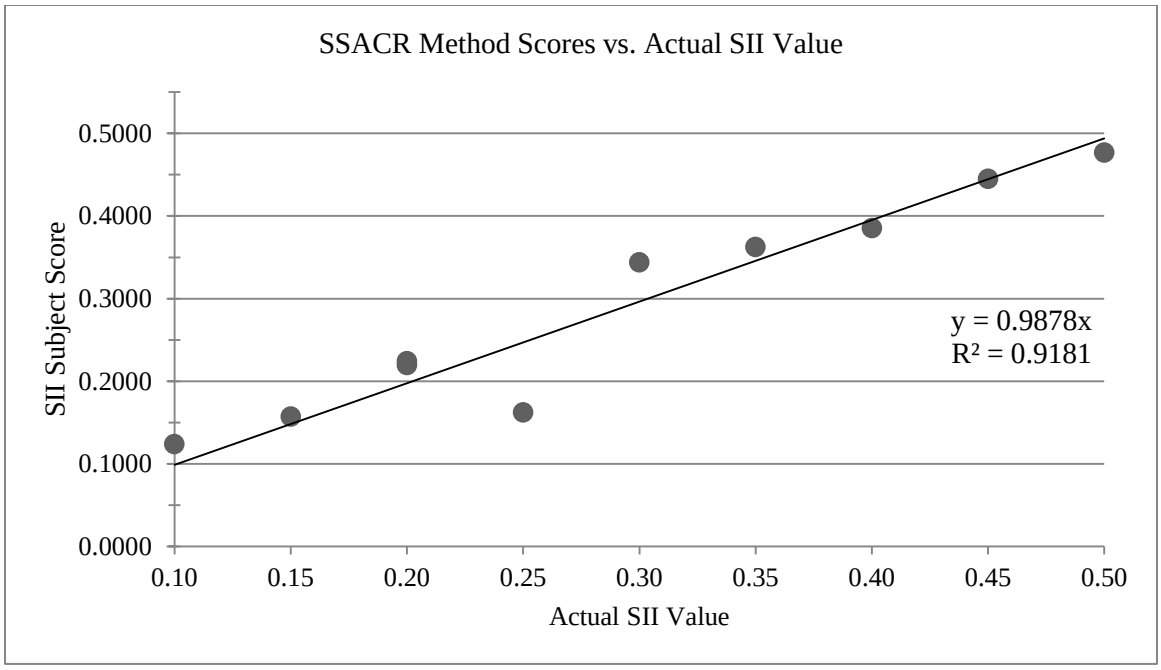


Figure 4.2 - Line of Best Fit for SSACR Method Data

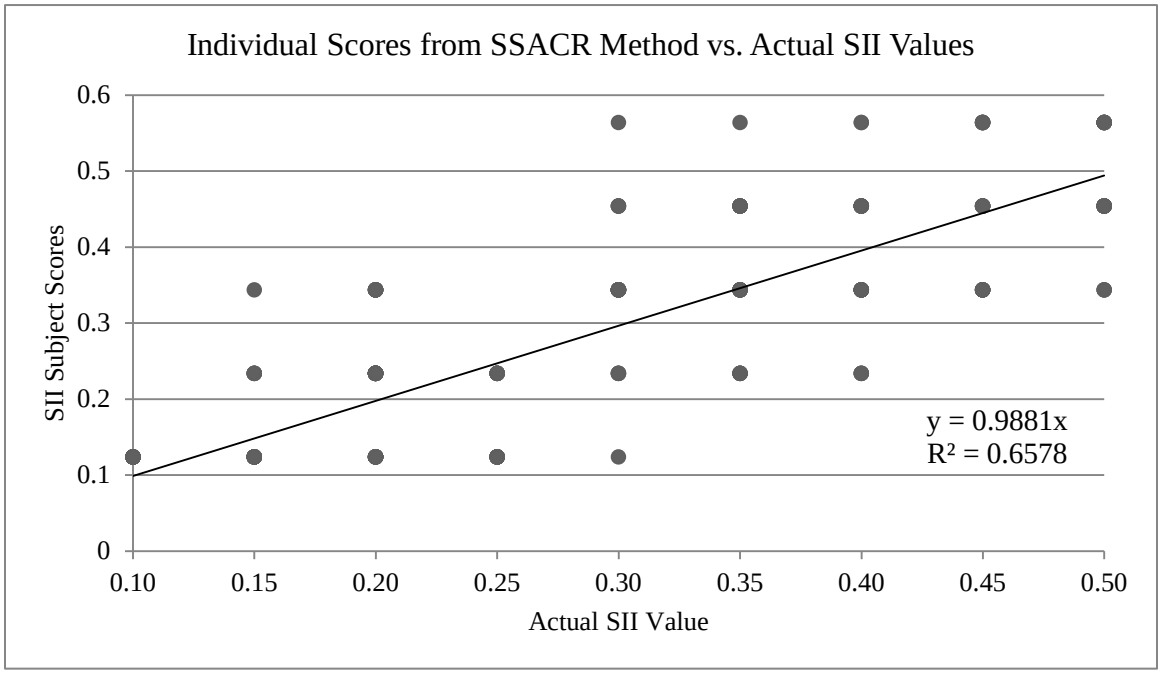


Figure 4.3 - Line of best fit for individual scores from SSACR method showing high variance (R2 = 0.6578) among subjects in the population.

In most of the test recordings that were not rejected by the t-test, the actual SII value falls in the middle of the range specified by the confidence interval. This suggests accuracy amongst subjects when using this method. This can be seen by the overlap of the confidence limits and the actual SII values in Figure 4.1. By combining the results of the t-tests, the overall low variance computed for data obtained from the SSACR method, and the results from the linear regression analysis, this method proves to be accurate, precise, reliable, and repeatable.

Using this data, the overall hypothesis concerning the validity and repeatability of the SSACR method can be discussed. When presenting the subjects with the SSACR method, subjects felt comfortable completing the experiment. While the method was easy to implement and quick to complete, the ambiguity occurring near the critical values of the SII scale and the limitations set by the number of options available on the subjective scale can attribute to flawed results. By incorporating more subjective scoring options, the variability in the data obtained from the SSACR method is believed to ultimately decrease.

4.2 IR Method

It was found that the scaling factor, β, for subjects performing the IR method was nearly equal to one; thus, a linear transformation was not needed in order to compare the IR subject data to the actual SII values. Table 4.15 displays the average mean, standard deviation, and variance for each test recording. Alongside this data are the actual SII values for each test recording.

Table 4.15 – Average Subject Data from IR Method with Outliers Removed

IR Subject Data with Outliers Removed				
Test Rec	**Actual SII**	**Mean of Scores**	**Standard Deviation**	**Variance**
1	0.10	0.03348	0.04942	0.00244
2	0.15	0.12955	0.09727	0.00946
3	0.20	0.17435	0.07372	0.00543
4	0.20	0.17957	0.06738	0.00454
5	0.25	0.13818	0.07129	0.00508
6	0.30	0.25957	0.09364	0.00877
7	0.35	0.30591	0.05068	0.00257
8	0.40	0.33957	0.07980	0.00637
9	0.45	0.38522	0.17990	0.03236
10	0.50	0.44409	0.10423	0.01086

Averaging over all variances, an overall average standard deviation of 0.0867 and variance of 0.0088 was computed showing high precision for the population. Comparing mean scores to the corresponding actual SII value, an overall linear correlation is observed (excluding Test Recording 5) between the two values. It is important to note for all test recordings, subjects on average scored the test recordings lower than the actual SII values. Accordingly, the IR method proves to slightly underestimate the SII value for each test recording, on average, which is shown in Figure 4.4. With both the computed standard deviation and variance being such low values, the IR method proves to be a precise method capable of reproducible results.

A one-sample t-test was performed for each test recording to test a null hypothesis, with the results shown in Table 4.16. The null hypothesis is shown in equation 4.3 and states that the mean of the subject scores for Test Recording *n* is equal to the actual SII value for Test Recording *n*.

$$H_0: \mu_i = SII_i \qquad 4.3$$

The mean, standard deviation, t-statistic, and 95% confidence interval are shown, as labeled in Table 4.16, for each test recording. The final column in the table states whether the null hypothesis, H_o, was rejected or not. The null hypothesis is rejected for a test recording if the 95% confidence interval does not include the actual SII value for that test recording. Figure 4.4 displays an overall clearer representation of the results derived from the t-test.

Table 4.16 – Results from one-sample t-test performed for IR Method

T-Test (α=0.05) for IR Subject Data with Outliers Removed						
Test Rec	**Actual SII**	**Mean**	**Std. Dev.**	**T**	**95% CI**	**Reject or Fail to Reject H_o**
1	0.10	0.0335	0.049	-6.456	**(0.012,0.055)**	*Reject*
2	0.15	0.1295	0.097	-0.986	(0.086,0.173)	Fail to Reject
3	0.20	0.1743	0.074	-1.669	(0.142,0.206)	Fail to Reject
4	0.20	0.1796	0.067	-1.454	(0.150,0.209)	Fail to Reject
5	0.25	0.1382	0.071	-7.357	**(0.107,0.170)**	*Reject*
6	0.30	0.2596	0.094	-2.071	(0.219,0.300)	Fail to Reject
7	0.35	0.3059	0.051	-4.081	**(0.283,0.328)**	*Reject*
8	0.40	0.3396	0.080	-3.632	**(0.305,0.374)**	*Reject*
9	0.45	0.3852	0.180	-1.727	(0.307,0.463)	Fail to Reject
10	0.50	0.4441	0.104	-2.516	**(0.398,0.490)**	*Reject*

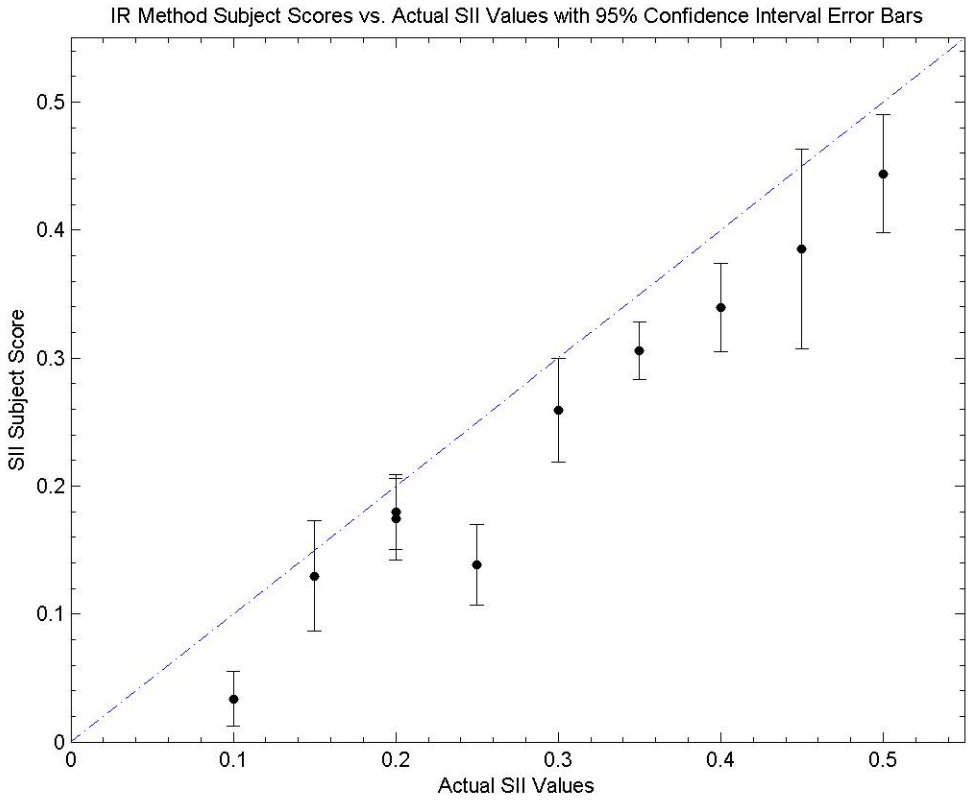


Figure 4.4 – Mean values from IR Method with Error Bars showing 95% Confidence Limit

In Figure 4.4, the x-axis corresponds to the actual SII value of the test recording and the y-axis represents the mean score computed from all subject responses. The population means for each test recording are denoted by the filled in circles (*Note: There are two test recordings located at an actual SII value of 0.2*). The error bars for each mean value show the 95% confidence limit derived from the t-test. The dotted line across the plot illustrates perfect subject score responses, having a slope equal to one. The error bars rarely overlap the dotted line, suggesting the method is not consistently accurate. However, all population means fall below the actual SII value implying that the IR method underestimates the actual SII value. It is also noted that the confidence limits

corresponding to the two recordings possessing actual SII values of 0.2 overlap very closely, further supporting the precision observed between subjects performing this method.

While a low variance was computed for the population from the IR method indicating reproducibility between subjects, the outcomes of the t-test suggest that the subjects' scores are not accurate. In five of the ten test recordings, the t-test suggests rejecting the null hypothesis which states the mean subject score is equal to the actual SII value. In these five cases, a 95% confidence interval displays the actual SII value is not contained in the interval. In each of these five test recordings, the actual SII value falls above the upper limit of the confidence interval. In the five remaining cases that the t-test failed to reject, the actual SII value falls towards the upper limit of the confidence interval. This reinforces that the IR method underestimates the actual SII value, although precisely as shown by the small variance computed for all recordings.

Using Excel, a line of best fit was applied to the mean subject scores with a y-intercept at zero as shown in Figure 4.5. The resulting R^2 value of 0.9421 indicates a high level of precision over the population. This displays a greater precision than the SSACR method. The slope of the line of best fit is equal to 0.8443, showing the accuracy of the population in this method is below that of the SSACR method. Linear regression analysis was performed on the entire data set and an R^2 value of 0.6037 was computed, indicating large variance between subjects, as shown in Figure 4.5.

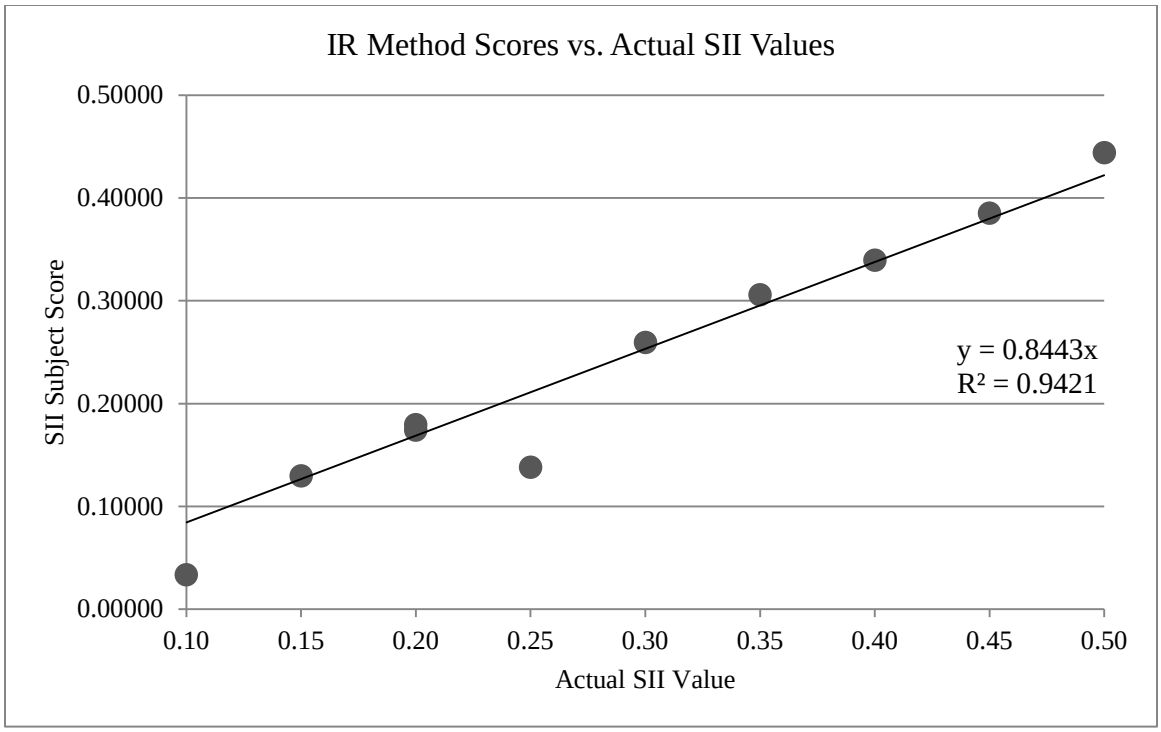


Figure 4.5 - Line of Best Fit for Data obtained from IR Method

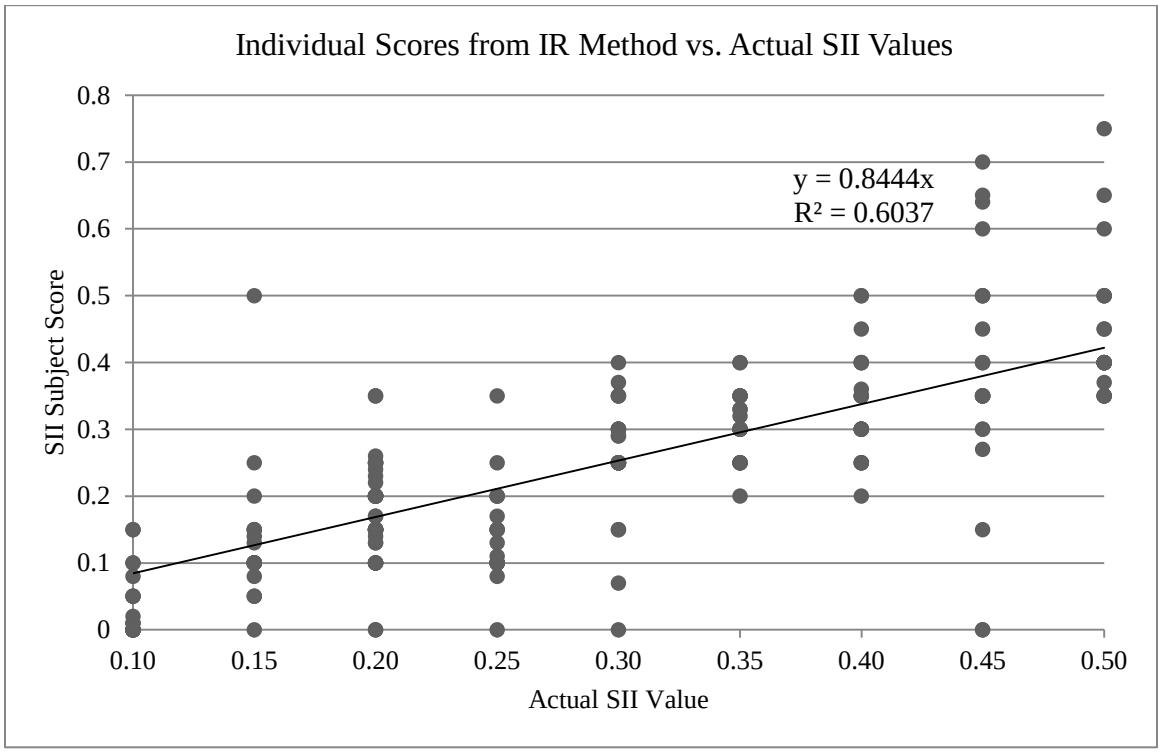


Figure 4.6 - Line of best fit for individual scores from IR method showing high variance (R2 = 0.6037) among subjects in the population.

Taking into account that all mean scores underestimate the actual SII value in the IR method, as illustrated by the means falling below the line in Figure 4.4, a bias may be present in the data. In the SSACR method, a linear transformation was performed in order to map the data onto the SII scale, as described by equation 3.2. Previously discussed was the decision to not use a linear transformation on the IR data due to the scaling factor, β, being nearly equal to 1. However, a level shift was applied to the SSACR data by means of the variables α and μ. This level shift resulted in an upward shift in the data by 0.1238 SII units; therefore, there was a bias present in the SSACR method which resulted in an underestimate of the actual SII values. Considering this same bias to be present in the IR method, an upward level shift by 0.0511 SII units was computed for IR data. Figure 4.7 shows the results of the data after applying the level shift to the IR data. A one-sample t-test was performed on the shifted data with 95% confidence limits also shown in the figure.

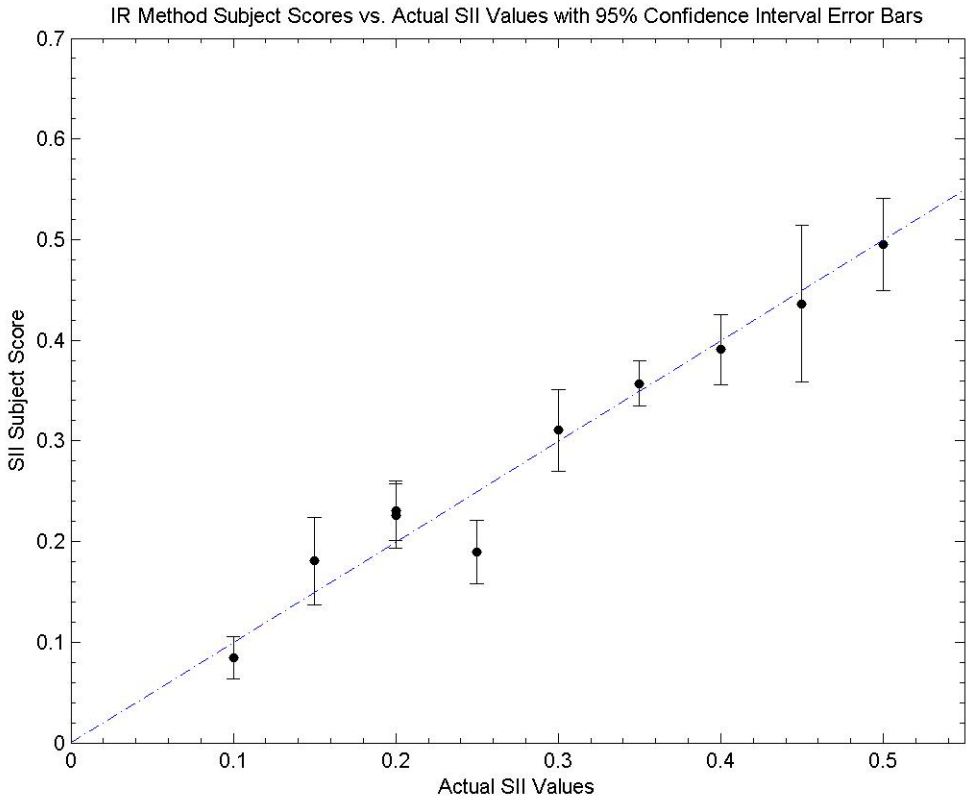


Figure 4.7 - Mean values from IR Method with Error Bars showing 95% Confidence Limit after Level Shift Applied

The results from the t-test are shown in Table 4.17, with the last column stating whether the null hypothesis, H_0, was rejected for each recording after the level shift was applied. The null hypothesis is rejected for a test recording if the 95% confidence interval does not include the actual SII value for that test recording. After applying the level shift, the number of test recordings that failed the statistical test decreased from five to two. Therefore, it was concluded that a bias was present in both methods which resulted in subject scores underestimating the actual SII values of the test recordings. By applying a level shift to the data, the data more accurately described the actual SII values with 95% confidence, as shown in Figure 4.7.

Table 4.17 - Results from one-sample t-test performed for IR Method after level shift was applied

T-Test (α=0.05) for IR Subject Data with Outliers Removed						
Test Rec	**Actual SII**	**Mean**	**Std. Dev.**	**T**	**95% CI**	**Reject or Fail to Reject H_o**
1	0.10	0.0335	0.049	-6.456	(0.063,0.106)	Fail to Reject
2	0.15	0.1295	0.097	-0.986	(0.138,0.224)	Fail to Reject
3	0.20	0.1743	0.074	-1.669	(0.194,0.254)	Fail to Reject
4	0.20	0.1796	0.067	-1.454	**(0.202,0.260)**	Reject
5	0.25	0.1382	0.071	-7.357	**(0.158,0.221)**	*Reject*
6	0.30	0.2596	0.094	-2.071	(0.270,0.351)	Fail to Reject
7	0.35	0.3059	0.051	-4.081	(0.335,0.379)	Fail to Reject
8	0.40	0.3396	0.080	-3.632	(0.356,0.425)	Fail to Reject
9	0.45	0.3852	0.180	-1.727	(0.359,0.514)	Fail to Reject
10	0.50	0.4441	0.104	-2.516	(0.449,0.541)	Fail to Reject

After applying the level shift, a line of best fit was applied to the data using Excel as shown in Figure 4.8. The resulting R^2 value increased to 0.9545 from 0.9421, indicating increased precision over the population. The slope of the line of best fit is equal to 0.9923, illustrating an increase in the accuracy over the population. Comparing this value to the slope obtained from the SSACR method (0.9878), the IR method proved to produce more accurate results over the population, as long as the level shift was applied.

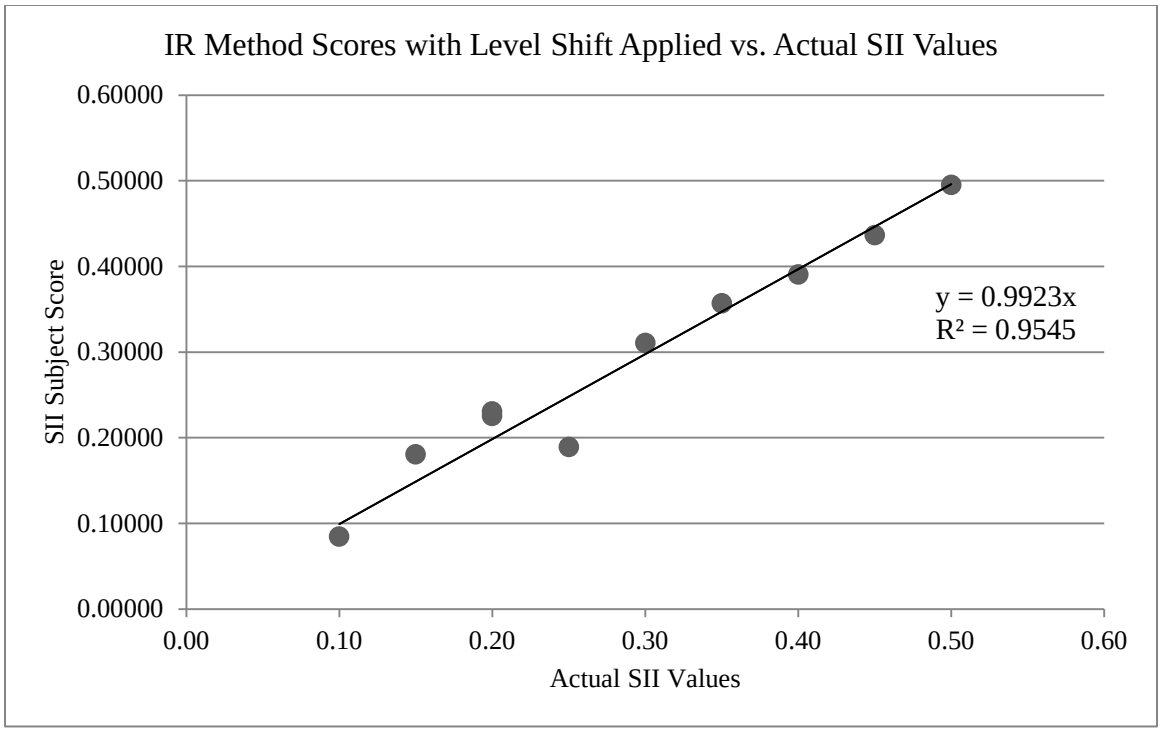


Figure 4.8 - Line of Best Fit for Data obtained from IR Method after applying level shift

Using the results from statistical tests implemented earlier, it is concluded that the decreased accuracy present in the IR method contributes most to the rejection of the null hypothesis when comparing the means of the two methods. The accuracy can be increased by applying an upward level shift to the data, resulting in a decrease in the number of test recordings which fail the statistical test. However, the source of the bias is unknown; therefore, the application of the level shift to the data is not justified in examining the accuracy of the IR method.

A paired-sample t-test was performed to compare the means and standard deviations between both methods. In each case, the null hypothesis is shown in equation 4.4 and states that there is no difference between the subject scores from each method.

$$H_0: \mu_{SSACR} - \mu_{IR} = 0 \qquad 4.4$$

Performing the paired-sample t-test over the population means of the methods resulted in a failure to reject the null hypothesis at a confidence level of 95%, as long as level shifts were applied to both data. When the level shift is not applied to the IR data, the null hypothesis is rejected by the t-test, implying a difference is present between the methods. Failure to reject the null hypothesis (95% confidence level) resulted after performing the paired-sample t-test over the population standard deviations of the methods. The results describe high accuracy in both methods, as long as a level shift is applied to the data, while also portraying the precision of the populations in both methods. Therefore, the results from the paired-sample t-test reinforce the results discussed previously regarding the accuracy and precision over the population.

Using this data, the overall hypothesis concerning the validity and repeatability of the IR method can be discussed. When presenting the subjects with the IR method, many subjects had a difficult time understanding how to rate the test recordings. Due to the confusion of subjects and monotony of the test, the results of the IR method may be skewed. This lack of confidence while completing the experiment can explain the average scores underestimating the SII value of the SOI.

While the IR method is tedious, the data indicates that the method produces precise, repeatable results amongst subjects; however, on average, the IR method underestimated the actual SII value of each test recording used in the study. By applying a level shift to the data, the accuracy of the population increases. However, without the shift the accuracy of the IR method is below that of the SSACR method as shown by the computed slopes from the line of best fit. A slope value closer to one computed from the

SSACR data shows higher accuracy among subjects identifying the intelligibility of the SOI. With a scaling factor, β, nearly equal to one and a level shift near zero (0.0511), the IR method proves to be a close estimate of the actual SII value when testing over a population. The large variance present between subjects (R^2=0.6037) suggests that a large population is needed, rather than a few people, in order to obtain an SII score which subjectively describes the SOI in a cocktail party recording.

CHAPTER 5: CONCLUSION

Much research has been done in the field of speech intelligibility, especially in cases of SOIs in cocktail party scenarios. There are existing methods, subjective and objective, which aim at quantifying the intelligibility of a SOI as it pertains to the listener's perception of intelligibility. Due to limitations which require knowledge of the clean speech sample separate from the noise sample, these methods of analysis are not applicable in actual applications.

By introducing the concept of an Audio Intelligibility Ruler, derived from the Image Quality Ruler, an experimental methodology was created which aimed at producing results which were repeatable between subjects. Adapted from the widely accepted MOS method, the SSACR method was also used to compare against the results of the IR method in order to validate the findings. The IR was developed using cocktail party recordings consisting of known SII values ranging from 0 to 1. The ruler provided the subject sample reference recordings with pre-determined SII values, which the subject used to score the intelligibility of SOIs in test recordings of unknown SII value. The SSACR method provided no reference samples, but required the subject to score the intelligibility of the SOI in the test recording using the provided subjective scale.

A least squares linear transformation had to be applied to the data obtained from the SSACR method in order to compare scores to the SII scale. Overall average variances computed suggest that both methods were precise and there was small variation over the population. An average standard deviation of 0.07260 for the SSACR method versus an average standard deviation of 0.0867 for the IR method, along with the R^2 values computed from the line of best fit, suggests high precision in both methods.

Performing one-sample t-tests for each test recording , as well as applying a line of best fit to data obtained from each method indicates more accuracy and repeatability in the SSACR method compared to the IR method without a level shift applied.

For the IR method, five of the ten recordings were rejected by the t-test (α=0.05), due to the 95% confidence interval not containing the actual SII value. In all cases of the IR method, the mean SII score was lower than the actual SII value proposing that the IR method underestimates the actual SII value of a SOI in a recording. These flaws were attributed to subject confusion pertaining to the notion of quality versus intelligibility. It was believed that subjects rated the quality of the signal rather than the intelligibility of the SOI in the signal. The tediousness of the method itself was expected to contribute to bias in the experiment, as well as factor into the variance observed over the population. Assuming bias in the experiment contributed to underestimated scores, a level shift was applied to the IR data resulting in increased accuracy and precision over the population, as well as a decrease in the number of recordings rejected by the t-test.

For the SSACR method, three of the ten recordings were rejected by the t- test (α=0.05), in which two cases involved recordings with SII values in the critical region. With the exception of these three cases, most of the mean scores sat in the middle of the range specified by the 95% confidence interval, implying accuracy over the population. Therefore, the SSACR method is believed to produce accurate SII scores for test recordings with unknown SOI SII values and has a high level of repeatability over a large population.

While subject reliability was not evaluated since only one trial of each method was required, the resulting data could be somewhat biased; however, the focus of the experiment was to look at the consistency across the population firstly. Examining the variability among subjects illustrated low R^2 values for both the SSACR and IR method (R^2=0.6578 and R^2=0.6037, respectively), suggesting that a large population is needed to subjectively rate a SOI in a cocktail party scenario. Future modifications may allow for evaluation of subject reliability, resulting in a decrease of possible biased results.

The focus of the experiment involved examining the intelligibility of a SOI in cocktail party scenarios; therefore background noise used in the test recordings consisted of only multiple people talking simultaneously and the noise sources were mixtures of men and women. Future adjustments could allow for different types of background noise, including music included and all noise sources being of the same gender. Finally, the arrangement of the IR could be altered in such a way that allows the subject to continuously alter the SII of the SOI in the reference samples. By using a slider to control the SII level of the SOI, the subject may be able to better distinguish the level of intelligibility that best matches the SOI in the test recording. While this may result in making the IR method more tedious than it presently is, better quality SII scores may result.

With the completion of this research, a method of analyzing subjective perception was modified from the image realm and applied to the topic of SOI intelligibility in the audio world. By using the IR, developed using the SII scale, to compare and score a SOI in recordings mimicking cocktail party scenarios, the IR method proves to be precise and repeatable amongst subjects; however, it does present itself as tedious and can be

confusing. The outcome of this experiment revealed the ability of subjects to precisely match various levels of speech intelligibility from different sources. By using the IR, subjects were able compare and distinguish SII values between a sample and a series of references consistently, as illustrated by the low standard deviations computed over the population means.

Further insight in using a modified MOS method for analysis in quantifying the intelligibility of a SOI suggests more reliable, accurate results in comparison to the IR method. It is also less tedious and has room for improvement which may result in less biased results. Research in the area of quantification of subjective perception will ultimately help lead to ways to better understand a subject's perception of intelligibility as it relates to a SOI in cocktail party scenarios. This allows for better advancements in many areas, such as covert surveillance, business related applications, and speech recognition.

APPENDIX A: LIST OF ABBREVIATIONS

AI	Articulation Index
ANOVA	Analysis of Variance
CD	Cepstrum Distance
DCR	Degradation Category Rating
DRT	Diagnostic Rhyme Test
DSCQS	Double Stimulus Continuous Quality Scale
DSIS	Double Stimulus Impairment Scale
DSP	Digital Signal Processing
IR	Intelligibility Ruler
IS	Itakura-Saito
ITU-T	International Telecommunications Union – Telecommunications
JND	Just Noticeable Difference
LLR	Log-Likelihood Ratio
LP	Linear Prediction
LPC	Linear Prediction Coefficients
MOS	Mean Opinion Score
MRT	Modified Rhyme Test
PC	Paired Comparison
PESQ	Perceptual Evaluation of Speech Quality
QR	Quality Ruler
RMS	Root Mean Square
SII	Speech Intelligibility Index
SNR	Signal-to-Noise Ratio
SOI	Speaker of Interest
SSACR	Single Stimulus Absolute Category Rating
WSS	Weighted Spectral Slope

APPENDIX B: CONSENT FORM

Consent to Participate in a Research Study

Quantifying Subjective Perception of Intelligibility of Speaker of Interest in Cocktail Party Scenarios

WHY ARE YOU BEING INVITED TO TAKE PART IN THIS RESEARCH?
You are being invited to take part in a research study about quantifying and understanding a listener's perception of quality as it pertains to the intelligibility of a Speaker of Interest (SOI) in a cocktail party audio recording. If you volunteer to take part in this study, you will be one of about 30 people to do so at the University of Kentucky.

WHO IS DOING THE STUDY?
The person in charge of this study is Kirstin Brangers, a Graduate Student in the University Of Kentucky Department Of Electrical Engineering. She is being guided in this research by Dr. Kevin Donohue *(Advisor)*. There may be other people on the research team assisting at different times during the study.

WHAT IS THE PURPOSE OF THIS STUDY?
By doing this study, we hope to gain an understanding in what makes a Speaker of Interest in an audio signal intelligible. In a 'Cocktail Party' scenario, multiple people are speaking simultaneously. Existing algorithms work to focus on one individual and extract their stream of speech, but it's not flawless. By having a better understanding of how an individual perceives intelligibility, future modifications and improvements can be made to deliver more efficiency in these algorithms. The most important part to these algorithms is interpreting what the Speaker of Interest is saying; therefore, we would like to gain knowledge in this area to broaden our range of environments in which a Speaker of Interest can be extracted and still considered intelligible.

ARE THERE REASONS WHY YOU SHOULD NOT TAKE PART IN THIS STUDY?
You should not take part in this study if you are under the age of 18.

If you suffer from any auditory conditions that might impair your hearing, you may have difficulty participating in this study; however, if you are using devices that help correct the auditory condition and/or feel your impairment will not be an issue concerning your ability to provide usable data, we accept your participation in this study.

WHERE IS THE STUDY GOING TO TAKE PLACE AND HOW LONG WILL IT LAST?
The research procedures will be conducted at the Marksbury Davis Building. You will need to come to the second floor (Room 204G) one time during the study. The total amount of time this visit should take is a maximum of 30 minutes.

WHAT WILL YOU BE ASKED TO DO?
You will be required to listen to 20 audio sample test recordings during a 30 minute period and rate the samples based on intelligibility. You will participate in two sessions, each lasting approximately 15 minute.

Each recording will have a speaker of interest talking, with other people speaking simultaneously in the background. You will be given a list of possible sentences spoken by the speaker of interest in order to determine the speaker of interest in each recording.

In session 1, you will rate the quality of the speaker of interest in each test sample based on intelligibility.
In session 2, you will compare the intelligibility of the speaker of interest in the test sample to given reference recordings and select which reference recording the sample is most like.

Each session will contain 10 test samples for you to rate.

WHAT ARE THE POSSIBLE RISKS AND DISCOMFORTS?
To the best of our knowledge, the things you will be doing have no more risk of harm than you would experience in everyday life. However, you may experience a previously unknown risk or side effect. If so, please inform us of this immediately.

WILL YOU BENEFIT FROM TAKING PART IN THIS STUDY?
You will not get any personal benefit from taking part in this study.

DO YOU HAVE TO TAKE PART IN THE STUDY?
If you decide to take part in the study, it should be because you really want to volunteer. You will not lose any benefits or rights you would normally have if you choose not to volunteer. You can stop at any time during the study and still keep the benefits and rights you had before volunteering. As a student, if you decide not to take part in this study, your choice will have no effect on you academic status.

IF YOU DON'T WANT TO TAKE PART IN THE STUDY, ARE THERE OTHER CHOICES?
If you do not want to be in the study, there are no other choices except not to take part in the study.

WHAT WILL IT COST YOU TO PARTICIPATE?
There are no costs associated with taking part in the study.

WILL YOU RECEIVE ANY REWARDS FOR TAKING PART IN THIS STUDY?
You will not receive any rewards or payment for taking part in the study.

WHO WILL SEE THE INFORMATION THAT YOU GIVE?
We will keep private all research records that identify you to the extent allowed by law. However, there are some circumstances in which we may have to show your information to other people. We may be required to show information which identifies you to people who need to be sure we have done the research correctly; these would be people from such organizations as the University of Kentucky.

CAN YOUR TAKING PART IN THE STUDY END EARLY?
If you decide to take part in the study you still have the right to decide at any time that you no longer want to continue. You will not be treated differently if you decide to stop taking part in the study.

WHAT ELSE DO YOU NEED TO KNOW?
There is a possibility that the data collected from you may be shared with other investigators in the future. If that is the case the data will not contain information that can identify you unless you give your consent or the UK Institutional Review Board (IRB) approves the research. The IRB is a committee that reviews ethical issues, according to federal, state and local regulations on

research with human subjects, to make sure the study complies with these before approval of a research study is issued.

WHAT IF YOU HAVE QUESTIONS, SUGGESTIONS, CONCERNS, OR COMPLAINTS?

Before you decide whether to accept this invitation to take part in the study, please ask any questions that might come to mind now. Later, if you have questions, suggestions, concerns, or complaints about the study, you can contact the investigator, Kirstin Brangers at Kirstin.Brangers@uky.edu. If you have any questions about your rights as a volunteer in this research, contact the staff in the Office of Research Integrity at the University of Kentucky at 859-257-9428 or toll free at 1-866-400-9428. We will give you a signed copy of this consent form to take with you.

_____ _____

Signature of person agreeing to take part in the study Date

_____

Printed name of person agreeing to take part in the study

_____ _____

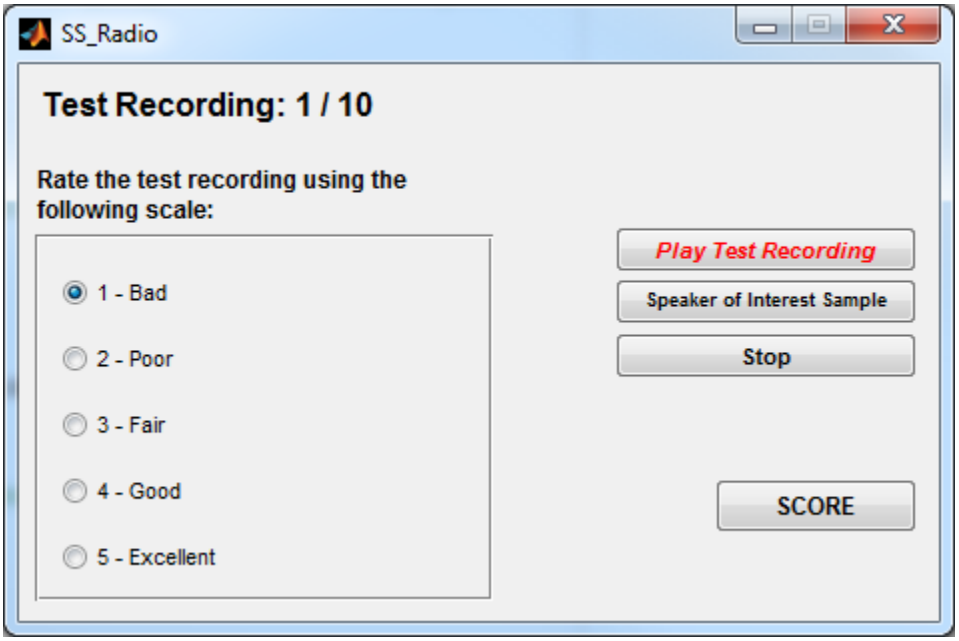
Name of (authorized) person obtaining informed consent Date

APPENDIX C: WRITTEN INSTRUCTIONS FOR BOTH METHODS

Method 1 – SSACR

You will be scoring 10 *test recordings* in this session. Each *test recording* will have a Speaker of Interest (SOI) talking, along with multiple other speakers speaking simultaneously.

The figure below shows the interface you will be using.



- *Speaker of Interest Sample* – Press to listen to the Speaker of Interest Recording with no added background speakers
- *Play Test Recording* – Press to listen to the Test Recording
- *Stop* – Press to stop any recordings currently playing

Your goal is to rate the intelligibility of the Speaker of Interest using the provided subjective scale.

1. Press the button labeled 'Speaker of Interest Sample' to listen to the Speaker of Interest recording with no background speakers interfering.

2. Press the button labeled 'Play Test Recording' to listen to the *test recording*.

3. Rate the intelligibility of the SOI in the *test recording* using the provided scale.

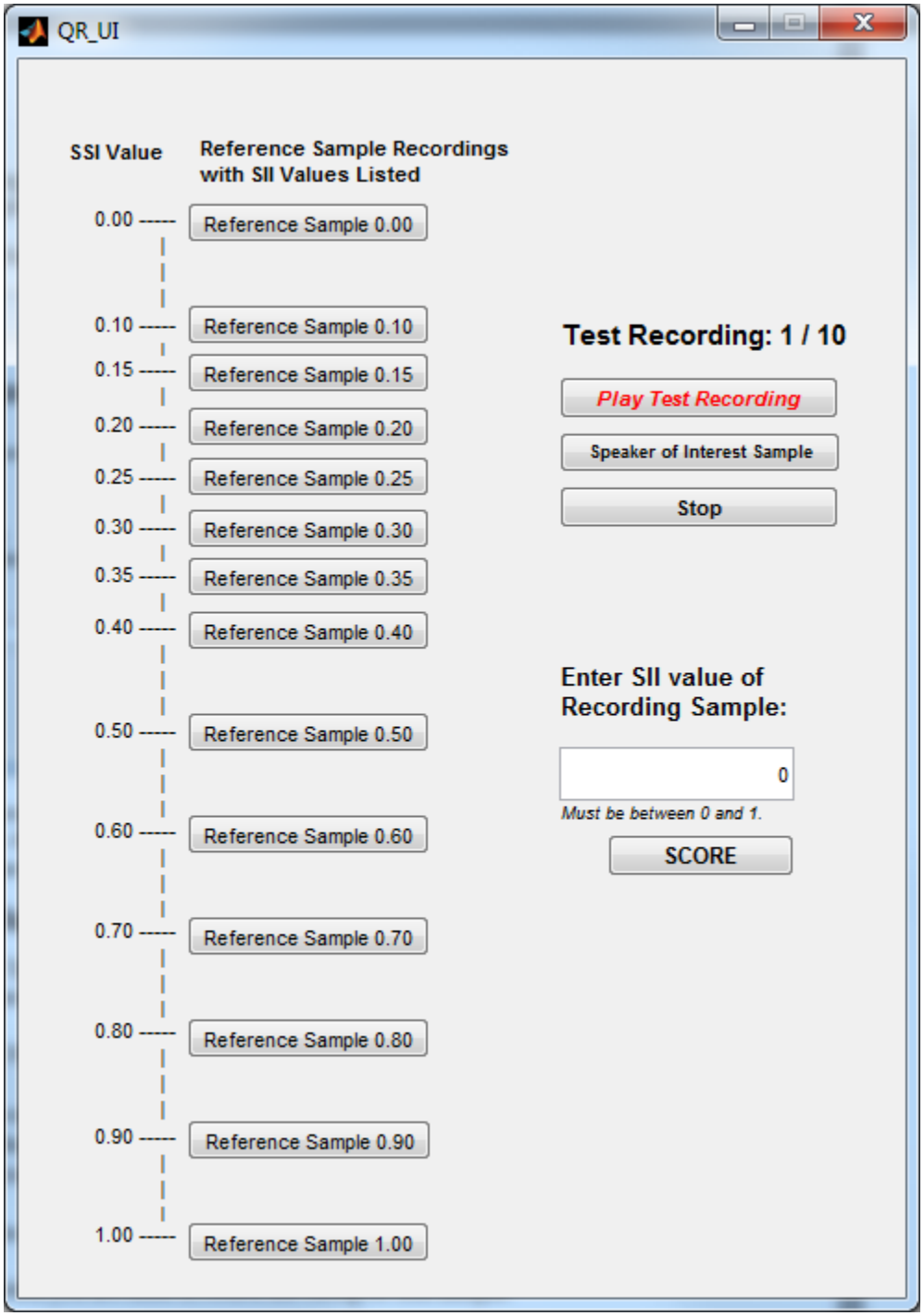
4. Press the 'Score' button when you are done to record the rating for that sample.

The next *test recording* will automatically appear. For example, the interface will show 'Test Recording: 2/10' at the top after you score the first sample.
If you have any questions, please ask the PI.

Method 2 – IR

You will be scoring 10 *test recordings* in this session. Each *test recording* will have a Speaker of Interest (SOI) talking, along with multiple other speakers speaking simultaneously.

The figure below shows the interface you will be using.



- *Speaker of Interest Sample* – Press to listen to the Speaker of Interest Recording with no added background speakers
- *Play Test Recording* – Press to listen to the Test Recording
- *Stop* – Press to stop any recordings currently playing

Your goal is to rate the intelligibility of the Speaker of Interest using the provided Intelligibility Ruler.

The Intelligibility Ruler provides reference sample recordings, as shown in the figure.

You can listen to the reference recordings by clicking the buttons labeled 'Reference Sample X.XX.' You are to listen to all reference recordings and compare the intelligibility of the *test recording* to the references given.

1. Press the button labeled 'Speaker of Interest Sample' to listen to the Speaker of Interest recording with no background speakers interfering.

2. Press the button labeled 'Play Test Recording' to listen to the Test Recording.

3. Listen to the reference samples and compare the intelligibility of the Speaker of Interest in the Test Recording to the intelligibility of the Speaker of Interest in the reference recordings.
   (Listen to 'Reference Sample 1.00' first in order to determine the Speaker of Interest in the reference recordings).
   a. The following sentence is spoken by the Speaker of Interest in the reference recordings:
      "Railroads are for catching trains. Sidewalks should be kept clean in winter."

4. Place the Test Recording at a location on the Intelligibility Ruler by choosing where the test recording is most like the reference recording in terms of Speaker of Interest intelligibility.

5. In the box on the right side of the interface under the label 'Enter SII value of Recording Sample,' enter the score you give the test recording.
   To score the test recording:
   - Enter the score based on the reference value you selected to be most like the *test recording*.
     This is under the column labeled 'SII Value'.
   - Base the score of the *test recording* on this SII Value.
   - You are allowed to enter any value in the box (0-1)

6. Once you enter a score into the box, press the 'Score' button to record the rating for that sample.

The next test recording will automatically appear. For example, the interface will show 'Test Recording: 2/10' after you score the first sample.
If you have any questions, please ask the PI.

APPENDIX D: LINEAR TRANSFORMATION

$$\boldsymbol{SII_{new} = \beta(SII_{SSACR} - \alpha) + \mu}$$

1) Make both vectors (Actual SII Values and SII Subject Scores) zero mean

$\vec{x} = \overrightarrow{x_0} - \mu_x$ Actual SII Scores
$\vec{y} = \overrightarrow{y_0} - \mu_y$ SII Subject Scores

where $\mu = \mu_x$ and $\alpha = \mu_y$.

2) Compute and apply scaling factor, β

$$(\beta\vec{y} - \vec{x})^T(\beta\vec{y} - \vec{x}) = SquaredError$$

$$\beta^2\vec{y}^T\vec{y} - \beta\vec{x}^T\vec{y} - \beta\vec{y}^T\vec{x} + \vec{x}^T\vec{x} = SquaredError$$

Minimize Squared Error → differentiate according to β

$$2\beta\vec{y}^T\vec{y} - \vec{x}^T\vec{y} - \vec{y}^T\vec{x} = 0$$

$$\beta = \frac{\vec{x}^T\vec{y}}{\vec{y}^T\vec{y}}$$

$\beta = \frac{\frac{1}{N}\sum_i^N x_i y_i}{\frac{1}{N}\sum_i^N y_i^2}$ Scaling Factor for vectors

3) Restore mean value of Actual SII Values vector to both vectors

$$\overrightarrow{x_{scaled}} = \vec{x} + \mu_x \qquad \overrightarrow{y_{scaled}} = \vec{y} + \mu_x$$

REFERENCES

[1] B. W. Keelan, Handbook of Image Quality: Characterization and Prediction, New York: Marcel Deckker, Inc., 2002.

[2] B. W. Keelan and H. Urabe, "ISO 20462, A psychophysical image quality measurement standard," in *SPIE Proceedings*, 2003.

[3] J. Redi, H. Liu, H. Alers, R. Zunino and I. Heynderickx, "Comparing Subjective Image Quality Measurement Methods for the Creation of Public Databases," in *Image Quality and System Performance VII*, San Jose, 2010.

[4] International Telecommunication Union, "ITU-T Recommendation P.911 Subjective audiovisual quality assessment methods for multimedia applications," 1998.

[5] International Telecommunication Union, "ITU-R Recommendation BT.500-13 Methodology for the subjective assessment of the quality of television pictures," 2012.

[6] M. Pinson and S. Wolf, "Comparing subjective video quality testing methodologies," in *SPIE Video Communications and Image Processing Conference*, 2003.

[7] ANSI, *ANSI S3.5-1997. American National Standard Methods for the Calculation of the Speech Intelligibility Index,* New York: ANSI, 1997.

[8] K. S. Rhebergen and N. J. Versfeld, "A Speech Intelligibility Index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *Journal of the Acoustical Society of America,* vol. 117, no. 4, pp. 2181-2192, 2005.

[9] K. Kondo, Subjective Quality Measurement of Speech: Its Evaluation, Estimation and Applications, Springer, 2012.

[10] G. Fairbanks, "Test of Phonemic Differentiation: The Rhyme Test," *The Journal of the Acoustical Society of America*, vol. 30, no. 7, pp. 596-600, 1958.

[11] H. J. Steeneken and T. Houtgast, *Basics of the STI-measuring method,* The Netherlands, 2002.

[12] Y. Hu and P. C. Loizou, "Evaluation of Objective Measures for Speech Evaluation," *IEEE Transactions on Audio, Speech, and Language Processing* , vol. 16, no. 1, pp. 229-238, Jan. 2008.

[13] International Telecommunication Union, "ITU-T Recommendation P.862 - Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," 2001.

[14] J. G. Beerends, S. van Wijngaarden and R. van Buuren, "Extension of ITU-T Recommendation P.862 PESQ towards Measuring Speech Intelligibility with Vocoders," in *New Directions for Improving Audio Effectiveness* , Neuilly-sur-Seine, France, 2005.

[15] J. G. Beerends, E. Larsen, N. Iyer and J. M. v. Vugt, "Measurement of speech intelligibility based on the PESQ approach," MESAQIN, 2004.

[16] J. Ma, Y. Hu and P. C. Loizou, "Objective measures for predicting speech

intelligibility in noisy conditions based on new band-importance functions," *The Journal of the Acoustical Society of America*, vol. 125, no. 5, pp. 3387-3405, 2009.

[17] D. H. Klatt, "Prediction of Perceived Phonetic Distance From Critical-Band Spectra: A First Step," in *IEEE ICASSP*, 1982.

[18] B. Grundlehner, J. Lecocq, R. Balan and J. Rosca, "Performance Assessment Method for Speech Enhancement Systems," 2005.

[19] C. Pavlovic, "The speech intelligibility index standard and its relationship to the articulation index, and the speech transmission index," *Journal of the Acoustical Society of America,* vol. 119, no. 5, p. 3326, 2006.

[20] A. W. Bronkhorst and R. Plomp, "Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing," *The Journal of the Acoustical Society of America,* vol. 92, no. 6, pp. 3132-3139, 1992.

[21] SII, "SII: Speech Intelligibility Index," [Online]. Available: http://www.sii.to/html/programs.html. [Accessed 2013].

[22] O.-H. Bjor, "Measure Speech Intelligibility with a Sound Level Meter," *Sound and Vibration,* pp. 10-13, 2004.

[23] H. Unnikrishnan, K. D. Donohue and J. Hannemann, "Interference Masking for Speaker of Interest Extraction if Cocktail Party Noise," *IEEE Transactions on Audio, Speech, and Language Processing.*

VITA

Name: Kirstin Marie Brangers

Birthplace: Louisville, KY

Educational Institutions

B.S. in Physics, Area of Concentration in Electrical Engineering

Morehead State University, Morehead, KY

August 2007 - May 2011

Positions Held

Graduate Research Assistant

University of Kentucky

Department of Electrical and Computer Engineering

Lexington, KY

May 2012 – August 2013

Graduate Teaching Assistant

University of Kentucky

Department of Electrical and Computer Engineering

Lexington, KY

August 2012 – May 2013

Society Memberships

Eta Kappa Nu – Treasurer (University of Kentucky)

Epsilon Pi Tau (Morehead State University)

IEEE Student Member