The Effect of Experience on Perceptual Spaces When Judging Synthesized Voice Quality: A Multidimensional Scaling Study

*Jessica Sofranko Kisenwether and †Robert A. Prosek, *Albany, New York, and †State College, Pennsylvania

Summary: Objectives/Hypothesis. The purpose of this study was to determine the effect of experience on the perceptual space of listeners when judging voice quality.
Study Design. This was a within-subjects group design.
Method. Speech-language pathologists, singing voice teachers, speech-language pathology graduate students with and without experience with a voice client, graduate students who have completed a voice pedagogy course, and inexperienced served as listeners. Each participant rated the similarity of pairs of synthesized stimuli with systematically altered measurements of jitter, shimmer, and noise-to-harmonics ratio on a visual analog scale ranging from no similarity to extremely similar.
Results. Results showed that participants with different levels and types of experience used different perceptual spaces (of additive noise and perturbation measures) when judging the similarity of stimulus pairs.
Conclusion. The conclusion was that perceptual spaces differ among individuals with different levels and types of experience when judging the similarity of pairs of stimuli with systematically altered acoustical measurements.
Key Words: Voice perception–Experienced listener–Multidimensional scaling–Acoustical measures–Synthesized stimuli.

INTRODUCTION

A recent study showed a difference in judgment of voice quality among individuals with different types of experience.[1] Individuals with a singing background, individuals with a speech-language pathology background, and inexperienced individuals were asked to judge a synthesized sustained vowel on a visual analog scale (VAS) ranging from mild to severe for overall severity, roughness, breathiness, strain, and pitch. Overall, individuals with a singing voice background rated signals more severely than individuals with a speech-language pathology background. Inexperienced listeners (IEs) did not follow a consistent pattern. The authors contributed these results to a possible effect of type of experience on listener judgment; however, due to moderate agreement levels during this task, the authors concluded that the use of specific voice quality terms may have biased the scale of measurement which forced the listeners to make a unidimensional judgment of voice quality.

Frequent questions of reliability with voice quality perceptions may be due to the multidimensional nature of voice stimuli. Listeners' judging/rating voice qualities are often using more than one parameter throughout their classification/rating tasks.[2–5] For instance, common perceptions of breathiness and roughness are used during voice quality rating.[6] Also, speech-language pathology graduate students' ratings of breathiness, hoarseness, and nasality encompassed many di-

mensions of the signal including: airflow, glottal periodicity, noise, and second formant frequency rise/fall, accounting for 48% of the variance during rating tasks.[5] These multiple dimensions can include but are not limited to intensity, noise-to-harmonics ratio (NHR), fundamental frequency, jitter, shimmer, and so forth.[2–5]

Because listeners will focus on multiple aspects of each signal to make a perceptual judgment of voice quality, some researchers feel that continuous scales are more suitable for rating.[7–10] Equal-appearing interval scales force a listener to make a unidimensional judgment on a multidimensional signal,[10] impacting listener agreement. In fact, even a training session did not help listeners obtain an agreement greater than a 0.80 when using a seven-point equal-appearing interval (EAI) scale for a variety of vocal qualities.[2] Through the use of multidimensional scaling (MDS), listeners are only asked to rate the similarity between a pair of stimuli, minimizing individual bias.[11] This allows the researcher to explore the dimensions within the acoustical signal used to make judgments.[6,11] The difficulty in choosing an appropriate rating scale is eliminated as dimensions are determined by the stimuli and not the scale. INDSCAL,[12] or individual differences scaling, is often used because it extracts the dimensions which represent the underlying judgments made by each participant as well as the judgments made by each group of listeners.[5,13] The results reveal the perceptual space being used for judgment of voice quality by each group of listeners. This space is a visual representation of the differences in domain and range for those perceptions.

Although continuous scales have been found to have better agreement, researchers continue to use EAI scales during the MDS task.[5,6,10,13–16] In turn, ratings may be skewed or participants may not be using the entire scale during perceptual tasks. In summary, speech-language pathology graduate students, or groups of listeners with mixed levels and types

of experience, are often asked to make unidimensional judgments on a multidimensional signal. Research has shown that experience can affect judgments of voice quality.[1,17] Results are then correlated to obtain acoustical measures to determine possible relationships among subjective and objective measures of voice quality. As a result, not only do results vary in regard to perceptual judgments of voice but research regarding objective measures related to these perceptions also remains contradictory.

Research suggests that listener agreement for rating tasks may be weak due to listeners making single-dimension judgments for voice quality, such as classification tasks on a complex signal that contains multiple dimensions.[18] Speech-pathology graduate students used three dimensions consistently when judging voice quality: fundamental frequency, intensity, and perturbation.[3] However, researchers stated that listeners may also attend to other properties of the acoustic signal resulting in varied limited agreement. For example, listeners' correlations have been found to vary from 0.33 to 0.78 when rating breathy voice quality and varied from 0.17 to 0.73 when rating hoarseness.[19] Also, interjudge agreement for ratings of hoarseness was 0.51 and interjudge agreement for ratings of breathiness was 0.55. Judges were said to have experience in voice disorders; however, speech-language pathology graduate students also participated in rating voice qualities indicating a significant difference in experience among judges. In addition, listeners' correlations, when judging roughness, also had a large variation across groups of listeners with and without experience in voice and/or voice disorders, resulting in a lack of a significant difference between groups.[20] Experienced listeners included speech-language pathologists (SLPs) with at least ≥2 years of postgraduate experience in the area of voice and four otolaryngologists, which again demonstrates a difference in experience. As discussed earlier and shown in the literature,[1] IEs have been shown to demonstrate a difference in judgment of voice quality as compared with experienced listeners. In summary, rating scale and group selection should be carefully selected for perceptual voice studies.

Although there have been many studies in the area of perceptions of voice quality, variables such as the use of anchors, type of rating scale, type and length of stimuli, and level and type of experience can affect perceptual judgments of voice quality and then, in turn, affect correlations with acoustical measurements of voice. Despite our knowledge of these factors affecting perceptions of voice quality, there are very few studies controlling for all the above variables simultaneously. Most importantly, there are very few studies that address the differences between experienced and IEs for perceptions of voice quality. In addition, generalizations are made for appropriate rating scales and correlations among acoustical measurements and perceptions through the common use of speech-language pathology graduate students as judges.

One must determine the differences between experienced and IEs, if this factor has been found to affect internal standards, before generalizing results. A consistent lack of control for variables affecting perceptions may be the reason for frequent disagreement among authors in the literature. In turn, careful group selection may yield different results for correlations among

acoustical components of the signal and perceptions using a MDS task to eliminate bias. The purpose of this study was to determine the perceptual space being used across groups with different levels and types of experience when judging synthesized sustained vowels using a MDS task so as to remove listener bias.

METHODS
Stimuli
The same stimuli used in Sofranko and Prosek[1] were used for this study. One sample of sustained vowel /ɑ/ with normal voice quality obtained from a female, aged 23 years, was synthesized using the UCLA synthesizer.[21] This sample, originally recorded at the University of Utah, was chosen because of its widespread use in other studies as an anchor to control for internal standards. Also, the sample was judged to be "normal" by SLPs who have experience in the area of voice and voice disorders on the basis of quality, pitch, and loudness.[22–24]

Using the UCLA Voice synthesizer,[21] this voice sample was synthesized with a duration of 1 second and a constant fundamental frequency and amplitude. The newly synthesized file was systematically altered by changing measurements of jitter, shimmer, and NHR. Jitter was altered in increments of 0.75 microseconds (0–3 microseconds) and shimmer was altered in increments of 0.5 dB (0–2 dB) for a total of 25 variations. NHR was altered in evenly spaced intervals of 12.5 dB resulting in five stimuli (−50 to 0 dB). This resulted in 435 pairs of stimuli to be presented during the study.

Listeners
The same listeners used in Sofranko and Prosek[1] were used for this study. There were six groups with 10 listeners in each group (n = 60). Groups consisted of SLPs, singing voice teachers (SVTs), speech-language pathology graduate students who had completed a voice disorders course and had not had a voice client (SLPGRADs), speech-language pathology graduate students who had completed a voice disorders course and had treated one or more voice clients (SLPGRADVs), graduate students in the music department who had completed a voice pedagogy course (SVTGRADs), and IEs.

Group 1 consisted of seven females and three males who were American Speech Language Hearing Association certified and state licensed SLPs. Ages ranged from 29 to 67 years (M = 45.7, standard deviation [SD] = 12.92). They had a range of 5–35 years of experience in voice disorders (M = 19, SD = 11.01) and spent 10–40 hours per week treating voice disorders (M = 23.4, SD = 12.21). All participants in group 1 reported no history of a hearing loss, a language disorder, a speech impairment, and/or a neurologic disorder.

Group 2 consisted of eight females and two males, ages ranging from 48 to 69 years (M = 59.6, SD = 6) who were tenured singing voice faculty and full members of the National Association of Teachers of Singing (NATS). Individuals holding a full membership in NATS, with either a Master's Degree or Doctor of Musical Arts, teach an average of six or more singing voice students weekly, and have ≥2 years of experience.[25] The criterion of tenure implies at least 6 years of full-