Perceptual Distances of Breathy Voice Quality: A Comparison of Psychophysical Methods

***Sona Patel, †Rahul Shrivastav, and ‡David A. Eddins,** *†*Gainesville, Florida and* ‡*Rochester, New York*

Summary: Experiments to study voice quality have typically used rating scales or direct magnitude estimation to obtain listener judgments. Unfortunately, the data obtained using these tasks are context dependent, which makes it difficult to compare perceptual judgments of voice quality across experiments. The present experiment describes a simple matching task to quantify voice quality. The data obtained through this task were compared to perceptual judgments obtained using rating scale and direct magnitude estimation tasks to determine whether the three tasks provide equivalent perceptual distances across stimuli. Ten synthetic vowel continua that varied in terms of their aspiration noise were evaluated for breathiness using each of the three tasks. Linear and nonlinear regressions were used to compare the perceptual distances between stimuli obtained through each technique. Results show that the perceptual distances estimated from matching and direct magnitude estimation task are similar, but both differ from the rating scale task, suggesting that the matching task provides perceptual distances with ratio-level measurement properties. The matching task is advantageous for measurement of vocal quality because it provides reliable measurement with ratio-level scale properties. It allows the use of a fixed reference signal for all comparisons, thus allowing researchers to directly compare findings across different experiments.

Key Words: Voice quality–Matching–Rating scale–Direct magnitude estimation.

INTRODUCTION

Voice quality is essentially a perceptual construct and obtaining listener judgments of quality is an integral part of voice quality measurement for research and clinical purposes. As with any other psychophysical task, it is necessary to obtain sensitive and reliable judgments of voice quality to develop a model for its perception. However, the methods used to study voice quality have often failed to take advantage of a vast body of knowledge in psychophysics. In this work, we attempted to address some of the shortcomings of contemporary methods to study voice quality using techniques described for other psychophysical research.

The vast majority of experiments to study voice quality obtain listener judgments using a rating scale task. Two commonly used variations include the use of an n-point rating scale or a continuous line in a "visual analog" format. Additionally, most of these experiments use an unanchored experimental design where listeners are required to make their judgments based solely on their experiences and memory, rather than using a "standard" reference stimulus for the purpose of comparison. A very limited number of experiments have used techniques such as direct magnitude estimation[1,2] and matching[3,4] to obtain perceptual judgments of voice quality.

A major limitation in using rating scales is the high variability in listener judgments, both within and across listeners. For example, Kreiman et al[5] showed that rating scale judgments for an individual voice stimulus could span the entire range of a seven-point rating scale. The variability in rating scale estimates was greatest for stimuli with an average rating in the middle of the scale and less at the two extremes. Such variability in perceptual judgments on a rating scale task is encountered in virtually all kinds of perceptual judgments. This finding has been addressed by several researchers who have proposed different approaches to explain such observations (eg, Refs. 6–8). These approaches also allow experimenters to design perceptual tests in ways that account for the variability in perceptual judgments. For example, Shrivastav et al[9] were able to show that interlistener variability in rating scale estimates of voice quality was minimized when multiple ratings of a stimulus were averaged and standardized. Therefore, although the variability in voice quality ratings poses many challenges in everyday situations (such as in a voice clinic), the variability in listener judgments can be minimized in an experimental setup as long as the experimental procedures are well designed and controlled.

Nevertheless, psychophysical scaling data obtained using rating scales have additional limitations. One problem relates to the level of measurement obtained when listeners are asked to make perceptual judgments on a rating scale. In the common parlance of voice quality research, the use of an n-point rating scale has often been referred to as an "equal-appearing interval" (EAI) scale, suggesting that the data obtained in these experiments are made on an interval scale (ie, each unit on the scale is perceptually equidistant from its neighboring units). Such a conclusion necessitates two basic assumptions. The first assumption is that listeners are able to perform an additive operation when making subjective judgments for voice quality. In other words, it assumes that listeners are able to evaluate the voice quality of samples in terms of constant perceptual distances from neighboring stimuli. Thus, if a voice is rated as a "3" on a seven-point rating scale, it implies that this voice is equally different from voices rated as "2" or "4" on the same scale. Secondly, an EAI scale further necessitates that

Accepted for publication August 4, 2008.
Part of this research was presented at the 36th Annual Symposium of the Voice Foundation in Philadelphia, Pennsylvania.

From the *Department of Communication Sciences and Disorders, University of Florida, Gainesville, Florida; †Department of Communication Sciences and Disorders, University of Florida, Gainesville, Florida and Malcom Randall VAMC, Gainesville, Florida; and the ‡Department of Otolaryngology, University of Rochester and the International Center for Hearing and Speech Research,Rochester Institute of Technology, Rochester, New York.
Address correspondence and reprint requests to Rahul Shrivastav, Department of Communication Sciences and Disorders, University of Florida, Dauer Hall, PO Box 117420, Gainesville, FL 32611. E-mail: rahul@ufl.edu

Journal of Voice, Vol. 24, No. 2, pp. 168-177
0892-1997/$36.00

doi:10.1016/j.jvoice.2008.08.002

listeners are aware of the total range of variation represented by the test stimuli and that they are able to effectively divide this range into subjectively equal categories. However, there is little evidence to support either of these assumptions in voice quality research. Indeed, considerable research has shown that listeners are not very good at describing prothetic continua using an interval scale (Ref. 10; however, see also Ref. 11 for a different perspective). Hence, the utility of rating scales in the measurement of voice quality may be questionable.[2] Indeed, in much of psychophysical research, a true EAI rating scale is achieved only if successive items on the rating scale are somehow determined to be perceptually equidistant from its neighbors (eg, as reported by Thorndike[12]) However, this intermediate step has seldom been addressed in voice quality research, further questioning the "equal-appearing interval" nature of the data thus obtained. Therefore, until further evidence about the equal-interval nature of rating scale data is obtained, it is best to treat the ratings as being ordinal in nature.[9] If certain assumptions regarding the distribution of this ordinal data are met, then additional statistical computations may be used to estimate interval-level information from the same ordinal data (eg, Ref. 7 for further explanation of this computation).

The first of the two problems described above has been addressed in great detail by Stevens.[10,13] His solution to the problem was to use a direct magnitude estimation task, where listeners are asked to judge ratios of sensation (instead of intervals) and to use a virtually unlimited range of numbers, including fractions, to describe the magnitude of sensation for prothetic continua. This method has been successfully used to study many different perceptual continua, resulting in a power function between the physical and perceptual magnitude of the stimulus known as Steven's Law. Although the exponent of the power function shows considerable variability across different types of perceptual continua, Stevens[10] argues that it suggests the general form in which physical stimuli may be mapped to a psychological sensation. Because the goal of the present work was to understand how a physical signal (the voice) is related to a psychological construct (its quality), we may assume that a direct magnitude estimation also may be useful for the study of voice quality perception.

However, the direct magnitude estimation task is not without its own limitations. One problem seen in both direct magnitude estimation and rating scale tasks is that listener responses are highly dependent on the context. For example, perceptual judgments on these tasks are biased significantly by factors such as the number of stimuli tested in an experiment, the perceptual range of the attribute being studied, the frequency of occurrence of different stimuli, etc.[7,8,11] This poses a significant hurdle because the results from one experiment cannot be directly compared to that of another. Because each experiment may use a different number of stimuli, often with a different range and frequency of the attribute under study, the associated contextual variability is difficult to identify and take into account. This makes it difficult to generate an appropriate model for voice quality perception based on magnitude scaling or rating scale data, because the results from either experiment may fail to generalize to a new set of data.

Direct magnitude estimation, and Steven's Law itself, are not without other criticisms as well. Poulton[11] has described a number of factors that bias listener judgments made in a direct magnitude estimation task. These include, for example, the logarithmic response bias, centering bias, contraction bias, etc. Many of these biases result from how listeners use numbers to reflect the magnitude of sensation. However, because one cannot directly access the magnitude of a sensation, the use of numbers often cannot be avoided. Nevertheless, certain steps can be taken to minimize the effects of such bias and to obtain perceptual judgments that are less influenced by factors such as the context, range, and frequency effects. One approach to minimize such errors is to use a matching task to obtain perceptual judgments. This provides listeners with a standard against which all comparisons can be made, thereby minimizing many biases associated with rating scale and the direct magnitude estimation tasks.

In a matching task, listeners are asked to manipulate a common reference signal to match the magnitude of one attribute of a test stimulus. For example, the loudness of a test sound may be judged by manipulating the sound pressure level (SPL) of a 1 kHz tone until it is perceived to have the same loudness as the test stimulus. The SPL of the 1 kHz tone then serves as a measure of loudness (measured in units called "Phons"). Although both stimuli in this example use the same sensory modality (within-modality matching), the same comparison can be made across two different sensory modalities as well (cross-modality matching). For example, observers may judge the loudness of a sound by manipulating the intensity of a light. In both cases, the reference signal acts as a yardstick that listeners can use in making perceptual judgments of the test stimuli. Using the same yardstick to obtain perceptual judgments for different stimuli, across different listeners and even across different experiments can help minimize many of the biases that plague ratings scale or direct magnitude estimation data. For these reasons, matching tasks are often the preferred method for measuring psychophysical continua and have been successfully used to study many different perceptual phenomena.

A matching task has also been used to study voice quality. In a series of experiments published over the last decade, Kreiman and colleagues have proposed a method to study voice quality using a novel matching task.[3,4,14] In this approach, they ask listeners to manipulate one or more parameters of a specially designed speech synthesizer until the quality of the synthesized speech sample matches that of the test stimulus. The settings of the synthesizer are then assumed to quantify the magnitude of the quality being studied. Although the general approach taken by Kreiman and colleagues has many similarities with the traditional matching tasks used in psychophysics, some key differences remain. Primarily, this matching technique allows listeners to vary multiple parameters of the vowel acoustic signal until a desired perceptual match in quality is obtained. In contrast, most psychophysical research has used a reference signal that can only vary along a single physical dimension, making it significantly easier to compute perceptual distances between various test stimuli. This difference in methodology likely reflects a somewhat different goal between the two