FROM VOCAL QUALITY MEASUREMENT TO PERCEPTION

Rahul Shrivastav
raheel@csd.ufl.edu
Department of Communication Science & Disorders, University of Florida, Gainesville

Abstract: Quantification of vocal quality of a speech signal is essential in a number of applications. However, existing measures for this purpose are often characterized by poor sensitivity and specificity to perceptual judgments. These shortcomings may have arisen because (1) these measures are often validated against “noisy” perceptual data and (2) the non-linear and multidimensional relationships between the physical signal and perceptual judgments have often been ignored. This paper describes the psychometric principles underlying quantification of subjective judgments and the use of an auditory processing model as a signal processing front-end when measuring “breathy” voice quality. Preliminary data for quantification of “roughness” is also discussed.

INTRODUCTION
The speech signal is rich in information and conveys a large amount of information to a listener. For example, apart from the meaning contained in the utterance, the speech signal conveys such information as emotions, speaker identity, age and gender. A number of applications require accurate quantification of such perceptual attributes of a speech signal. In the clinical domain, one may need to quantify aspects such as speech intelligibility, voice quality, nasality, etc. These attributes are important because these are often affected by disease and are frequently the target of surgical, pharmacological or behavioral treatment. Precise quantification of these attributes can enhance the assessment and rehabilitation procedures by providing a baseline against which any change can be measured.

Various techniques to quantify these perceptual attributes have been developed over the years. Some of these require listeners to make a subjective decision using a particular rating scale (for example, the CAPE-V developed by the American Speech-Language and Hearing Association). Others use a variety of signal processing techniques to quantify certain aspects of the speech acoustic signal [1-4]. Unfortunately, all of these methods have been compromised by a variety of problems. The accuracy of subjective judgments has been measured by calculating the reliability and agreement within and across listeners. Reliability measures the degree to which one listener’s ratings on a set of stimuli follow the same trends as that of another listener. Agreement, on the other hand is a measure of the probability that two listeners give the same stimulus the same exact rating. Unfortunately, both reliability and agreement have been found to be poor for subjective ratings of voice quality [5-7]. Similarly, the accuracy of automated measures to quantify perception is measured by calculating the correlation between these measures and perceptual judgments of voice quality. Unfortunately, most automated measures have been observed to show poor to moderate correlation with perceptual data. Additionally, these measures often lack consistency and show poor sensitivity and specificity to the perceptual construct that they intend to quantify [8].

When attempting to quantify attributes such as voice quality, it is important to remember that these are inherently perceptual constructs. Attributes such as voice quality, nasality or “acceptability” of speech essentially reflect a listener’s judgment about that particular construct. The speech signal itself does not possess quality; rather, it “evokes it in the listener” [8]. Therefore, any method to quantify such perceptual attributes must be validated against perceptual judgments made by listeners. These perceptual judgments serve as the gold standard for any other method to quantify perceptual attributes of speech. Unfortunately, perceptual judgments made by a listener are highly variable and are affected by a number of factors [9, 10]. While some of these factors are related to the stimulus characteristics, others are related to extraneous variables such as listener experience and training, instructions given to the listeners, nature of the scaling task and the experimental design. These extraneous variables introduce “noise” in the perceptual data, thereby, making it difficult to interpret the true perceptual magnitude of a given stimulus. However, these errors can be minimized through the use of appropriate experimental designs to obtain perceptual judgments [11]. These procedural modifications include multiple presentations of each stimulus to each listener, randomizing the order of stimulus presentation, modifying the instructions given to the listeners, etc.

Once a good estimate of the perceptual magnitude of an attribute has been obtained for several stimuli, these judgments may be used to develop a model that predicts listener judgments based on various stimulus characteristics. Such a model can be used to generate automated measures of vocal quality or other perceptual attributes of the speech signal. The development of such a model requires attention to the sensitivity of the human auditory system and the characteristics of the
acoustic-auditory transduction process. Previous research has shown that the relationship between a physical stimulus and its perceptual consequence is often non-linear. For example, the relationship between intensity and loudness may be described with a power law [12] and that between frequency and pitch is better described using non-linear scales such as the Bark, Mel or ERB-scales [13, 14]. In a similar manner, the perception of complex attributes such as voice quality may be best characterized by a non-linear function of specific stimulus characteristics. When the goal of measurement is to quantify perception, we need to (a) determine what aspects of the speech acoustic signal are perceptually relevant, and (b) determine the nature of the relationship between these stimulus characteristics and their perceptual consequences. One method to account for some of these non-linear processes is through the use of an auditory-processing model as a signal processing front-end. The general form of such a model is shown in Figure 1. The use of such front-ends has been shown to give better estimates of the perceptual judgments of voice quality [15, 16].

Figure 1: General form of an auditory processing model.

This paper describes a series of experiments to understand the perception of dysphonic voice quality. The first section describes the psychometric principles used to obtain a good estimate of the perceptual magnitude of dysphonic voice quality. In the second section, one particular auditory processing model and its utility in the quantification of “breathy” voice quality is described. And finally, preliminary data is presented on the perception of “rough” voice quality.

**PYSCHEMETRIC THEORY TO QUANTIFY SUBJECTIVE JUDGMENTS OF VOICE QUALITY**

A variety of techniques have been used for scaling perceptual magnitude of a physical stimulus [17]. It is necessary to differentiate two aspects in this process – sensory capability and response proclivity [18]. Sensory capability refers to the resolving power of the sensory mechanism; it defines the limits of the sensory system. On the other hand, response proclivity refers to the tendency of a listener to respond in a specific manner when encountering a specific stimulus. Since proclivity is affected by several factors, many unrelated to the stimulus itself, it is necessary to take appropriate steps to minimize “noise” in perceptual judgments.

For example, inter- and intra-rater reliability (measured as the correlation between ratings made within- or across-listeners) in perceptual ratings can be minimized by averaging multiple ratings of each stimulus by each listener [11]. Such precautions can avoid errors such as those arising due to “order-effects” and frequently seen with rating scales. If measurement of “agreement” (i.e. the probability that two raters would give the same stimulus exactly the same rating) is essential, the ratings from individual listeners should be standardized using z-scores or other procedures. Figure 2 shows the improvement in reliability and agreement for perceptual ratings of breathiness when these procedures are used.

Figure 2: Improvement in reliability and agreement when multiple ratings of each stimulus from each listener are used. Each listener’s ratings were converted to corresponding z-scores.

Although these techniques help improve agreement and reliability of rating scale data, these measures may not necessarily indicate the true magnitude of the stimulus. Rather, rating scale measures may only provide the rank ordering of the stimuli tested in the experiment. Other techniques, such as magnitude estimation, magnitude production, matching or paired comparisons may be better suited to obtain an accurate estimate of perceptual “distance” between two stimuli.

**AUDITORY PROCESSING MODEL AS A SIGNAL-PROCESSING FRONT END TO QUANTIFY “BREATHY” VOICE QUALITY**

Breathiness in voices has been found to correlate with a number of acoustic measures, including aspiration noise,
frequency/intensity perturbation and spectral slope. However, the correlation between these measures and perceptual judgments of breathiness has been found to be inconsistent across different experiments. 

Auditory processing models can allow us to estimate how acoustic signals may be represented in the auditory system. Several such models have been proposed [19-21]. One such model, proposed by Moore et al. (1997) was implemented for the study of breathy voice quality [15, 16]. This model simulates the outer and middle ear as band pass filters. The cochlear filtering is simulated with a filter-bank of asymmetric rounded-exponential filters. Finally, the neural excitation is modeled as a non-linear compressive function. The total neural excitation for a given sound provides an estimate of the loudness of that sound. The neural excitation within each “channel” is called the specific loudness.

This auditory processing model can also be used to simulate masking. Masking refers to the phenomenon where the loudness of a sound is reduced if it is presented along with a background noise. The loudness of a specific component, when it is presented simultaneously with an auditory masker is called the partial loudness.

The utility of this auditory processing model in predicting perceptual judgments of breathiness was tested in separate experiments [15, 16]. One experiment studied 13 voice stimuli, and compared the results to perceptual judgments obtained using a multidimensional scaling design. The other studied 27 stimuli and compared the results to perceptual judgments made on a 5-point rating scale. In both these experiments, the voice stimuli were first separated into a periodic component representing the complex wave produced by vocal fold vibration and an aperiodic component representing the aspiration noise. The auditory processing model was the used to estimate the partial loudness of the complex wave, while treating the aspiration noise as an auditory masker. In both these experiments, the partial loudness of the complex wave was found to correlate highly with the perceptual judgments of breathiness. This measure accounted for greater variance in the perceptual ratings of breathiness than any other acoustic measure of breathiness. Figure 3 shows the relationship between partial loudness of the complex wave and perceptual judgments of breathiness.

The use of an auditory-processing model accounts for multiple factors that may affect breathiness – the overall intensity of the complex wave and aspiration noise, the spectral shape of these components as well as the non-linear interaction between the two. This, presumably, accounts for greater variance in the perceptual data than using conventional acoustic measures such as measures of noise or spectral slope. The change in partial loudness patterns for different voices is shown in Figure 4.

**Figure 3:** Linear regression predicting ratings of breathiness using partial loudness of the complex wave (and assuming that aspiration noise acts as an auditory masker).

\[ y = -3.3583x + 19.997 \]

\[ R^2 = 0.7938 \]

**Figure 4:** Partial loudness patterns for voices identified as normal, mild-, moderate- and severely- breathy.

ACOUSTIC CORRELATES FOR “ROUGHNESS” IN VOICES – PRELIMINARY DATA

Many voices frequently observed in voice clinics are described as “rough.” A number of acoustic correlates for roughness have been proposed. These include, frequency/intensity perturbation, estimates of noise in the signal and the presence of subharmonics. However, as with breathy voices, these findings lack sensitivity and specificity.

From a psychoacoustic perspective, the perception of roughness is related to the amplitude and frequency modulation of a carrier wave. More specifically, the roughness of a sound is related to the amplitude modulation within a given critical-band [22]. The perception of roughness of a carrier wave is most sensitive to specific modulation frequencies.

A much simplified implementation of this model for roughness is obtained by: (1) determining the modulation frequencies in the vowel, (2) selecting a subset of these modulating frequencies, (3) calculating the “modulation amplitude” for these frequencies, and (4) determining the average modulation due to these frequencies.
The average modulation amplitude thus obtained was first normalized to the fundamental frequency of each stimulus, and was then used to predict perceptual judgments of roughness for 34 vowel samples. An exponential fit was found to account for 80.2% of the variance in the perceptual data. These data are shown in Figure 5. Measures such as shimmer, jitter and signal-to-noise ratio accounted for considerably less variance in the same perceptual data.

Figure 5: Perceptual ratings of roughness predicted by the normalized modulation amplitude of selected modulation frequencies.

\[
y = 0.0321e^{0.8058x}
\]

\[R^2 = 0.802\]

CONCLUSIONS
Voice quality is essentially a perceptual construct. Any method to quantify perception must be validated against perceptual judgments. However, since perceptual judgments are highly variable, one needs to devise experiments that minimize response variability associated with non-stimulus factors. Additionally, methods to quantify perception are more likely to be successful if they simulate the mechanisms involved in the auditory-perceptual process. One way to achieve this is through the use of auditory-processing models as a signal-processing front end. The success of this approach is shown for quantification of “breathiness” and “roughness” in dysphonic voices.

ACKNOWLEDGEMENTS
Research funded by a grant from NIH/NIDCD (R21DC006690). Part of this work is being done in collaboration with David Eddins.

REFERENCES