VOQUAL'03, Geneva, August 27-29, 2003

Jitter, Shimmer, and Noise in Pathological Voice Quality Perception

Jody Kreiman and Bruce R. Gerratt

Division of Head and Neck Surgery, School of Medicine
UCLA, Los Angeles
jkreiman@ucla.edu

Abstract

Although jitter, shimmer, and turbulent noise characterize all voice signals, their perceptual importance has not been established psychoacoustically. To determine which of these acoustic attributes is important in listeners' perceptions of pathologic voices, listeners used a speech synthesizer to adjust levels of jitter, shimmer, and noise so that synthetic voices matched natural pathological voices. Results show that listeners perceived spectrally-shaped, additive noise as the critical inharmonic acoustic element contributing to pathologic voice quality. Listeners proved highly insensitive to the amounts of jitter and shimmer present in a voice sample. These results suggest that jitter and shimmer are not useful indices of voice quality.

1. Introduction

Jitter, shimmer, and the ratio of harmonic to inharmonic energy in a voice (the signal to noise ratio) are the cornerstones of acoustic voice measurement. Hundreds of papers have described dozens of algorithms for calculating these measures, which can be generated by every existing commercial voice analysis system. However, the perceptual importance of these measures has never been demonstrated experimentally in natural contexts, and remains unknown. Previous studies have generally been limited to examining correlations between acoustic perturbation measures and listeners' ratings of qualities like "rough" and "breathy." Such studies have produced highly variable results. The present study used high-quality synthesis in a method of adjustment task to directly assess the extent to which listeners are perceptually sensitive to the amounts of jitter, shimmer, and noise present in a voice signal.

2. Experiment 1

2.1. Method

Twenty pathological voice samples (10 male and 10 female speakers; /a/, 1 s in duration), ranging from mildly to severely dysphonic, were selected from a library of samples. Speakers represented a variety of primary diagnoses, including mass lesions (7 cases), glottal incompetence (6 cases), chronic laryngitis (4 cases), adductory spasmodic dysphonia (2 cases), and Parkinson disease (1 case).

Vowel formant frequencies and bandwidths were estimated via LPC. Characteristics of the voicing source were estimated as follows. A single glottal pulse was inverse filtered, and the output of the inverse filter was least-squares fit with a modified LF source model (as described in [1]). F0 was tracked pulse by pulse on the time domain waveform to estimate the low-frequency (<12 Hz) amplitude and frequency contours. Cepstral comb liftering was applied to remove the periodic components and estimate the shape of the noise spectrum.

Synthesis was accomplished on a custom formant synthesizer implemented in Matlab. Formant frequencies, bandwidths, and LF parameters were adjusted to provide good perceptual matches to the target voices, and then held constant across experimental conditions. Tremor was modeled by incorporating the low frequency amplitude and frequency tracks from the original voices. Jitter was modeled using a time-warping algorithm. Period lengths were randomly sampled from a normal distribution of values with σ determined by the desired level of jitter. Shimmer was modeled by randomly altering the power of each cycle of phonation, with power values sampled at random from a normal distribution of values whose σ corresponded to the desired amount of shimmer. Noise was synthesized by passing white noise through a 100 tap finite impulse response filter that modeled the measured noise spectrum. The jittered and shimmered LF pulse train was added to this noise time series to create a complete source time series. Finally, the source was filtered through the vocal tract model.

Seventy non-expert listeners participated in this experiment. All reported normal hearing. Listeners initially heard the natural voice sample paired with a copy synthesized without jitter, shimmer, or noise. They adjusted jitter, shimmer, and/or noise levels by manipulated sliding cursors until the features of the synthetic voice perceptually matched those of the original voice. Listeners were allowed to make as many adjustments as necessary to achieve a satisfactory match to the target voice.

Each listener participated in 20 trials, one for each voice. For each trial, they were asked to match one of the following: jitter only, shimmer only, noise only, jitter + shimmer, jitter + noise, shimmer + noise, or jitter + shimmer + noise. Which task a listener performed for a given voice was assigned at random, with the constraint that 10 listeners performed each task for each voice. Voices were presented to each listener in a different random order. Prior to the experiment, the synthesizer was demonstrated and two practice items were presented, so that listeners could become familiar with the task and with the sound of jitter, shimmer, and noise. Practice lasted about 15 minutes. Total testing lasted 1.5 - 2 hours.

2.2. Results

To measure response variability, we calculated the coefficient of variation (the standard deviation divided by the mean) for the jitter, shimmer, and noise responses for each voice. Figure 1 shows the range of variability coefficients for each measure. Listener responses were significantly more variable overall for jitter and shimmer than they were for noise [$F(2, 297) = 162.12$, $p < .01$; Bonferroni post-hoc comparisons $p < .01$; Figure 1).

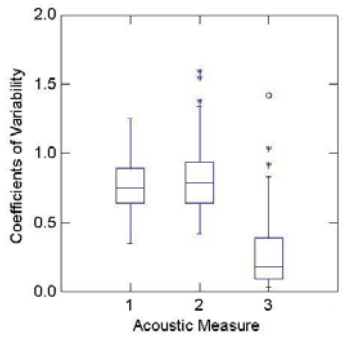

Figure 1: Overall variability in jitter (1), shimmer (2), and noise (3) responses.

Jitter and shimmer responses varied significantly with the particular listening task. Listeners used significantly less jitter and shimmer when asked to adjust all three parameters at once than they did when matching jitter or shimmer alone (jitter: $F(3, 796) = 4.32$, $p < .01$; shimmer: $F(3,796) = 4.14$, $p < .01$; Bonferroni post-hoc analyses $p < .01$). In contrast, listeners always adjusted noise to similar levels, even when they were also adding jitter and/or shimmer to the voice, so no significant effect of response condition on noise measures was observed ($F(3, 796) = 1.01$, n.s.). This contrast contributed to the large amounts of response variability for jitter and shimmer relative to noise.

Variability in noise responses *decreased* with increasing severity of pathology. In other words, the worse the voice, the easier it was to estimate how much noise was present ($r = -.64$, $p < .01$). No relationship between severity of pathology and response variability was observed for jitter or shimmer (jitter: $r = -.23$; shimmer $r = -.155$).

For noise, response variability increased as H1-H2 decreased ($r = -.54$, $p < .01$). (H1-H2 was calculated only for the harmonic voice source, and measures its spectral slope.) Listeners agreed best in their noise responses when the harmonic source was near-sinusoidal, as predicted by theory and previous research. No such relationship was observed between harmonic source characteristics and agreement levels for jitter or shimmer.

2.3. Discussion

The present results indicate that the correspondence between jitter and shimmer and perceived vocal quality is highly imprecise at best. Listeners produced values of jitter and shimmer ranging from 0 to extreme for each of the 20 test voices. Listeners' jitter and shimmer responses varied significantly with experimental condition, and response variability for jitter and shimmer could not be predicted by any other acoustic attributes of the voices. This pattern of substantial variability combined with task dependency suggests that jitter and shimmer cannot be considered reliable or valid as measures of perceived vocal quality.

In contrast, noise appears to be a highly salient perceptual attribute of voices. Listeners' noise responses varied much less than their jitter and shimmer responses (about 4 times less on average, and as much as 10 times less). The extent of listener disagreements in perceived noise levels could be well predicted by other characteristics of the voices (the shape of the harmonic part of the voice source and the amount of noise present). Finally, listeners' noise responses were consistent across experimental conditions.

Two explanations are possible for the large amounts of variability observed in jitter and shimmer responses. Listeners may be highly insensitive to differences in amounts of jitter and shimmer in a voice, so that their responses were in fact perceptually equivalent. Alternatively, listeners may be able to perceive differences in jitter and shimmer levels, but may have difficulty determining which level is the correct response because they cannot separate jitter and shimmer perceptually from the composite noise component. The following experiment examined the perceptibility of differences in listeners' jitter and shimmer responses in Experiment 1, to determine the source of the variability in those responses.

3. Experiment 2

3.1. Method

Eight voices were selected from the set of 20 studied in Experiment 1, based on speaker sex, spectral slope of the harmonic part of the source (H1-H2 large, H1-H2 small), and harmonics-to-noise ratio (very noisy, not very noisy). For each voice, 5 series of 5 stimuli each were synthesized: 2 series in which the amount of jitter varied in steps (1 series synthesized with no noise other than jitter; 1 series synthesized with a constant amount of noise in addition to the jitter); 2 series in which the amount of shimmer varied in steps (1 series synthesized with no noise other than shimmer; 1 series synthesized with a constant amount of noise in addition to the shimmer); and 1 series in which the amount of noise varied in steps. Endpoints of the series represented the largest and smallest values for that parameter and that specific voice observed in Experiment 1, with intermediate points evenly spaced (in acoustic units) between these extremes. Step sizes for jitter, shimmer, and noise thus varied by voice. For jitter, step sizes ranged from 0.19-0.65%; for shimmer, they ranged from 0.10-0.45 dB; and for noise they ranged from 0.87-4.57 dB. Constant levels for noise (series 2 and 4) were set to the mean value of listeners' selected noise levels in our previous experiment. All other synthesis parameters were held constant for each voice at levels used in Experiment 1. Stimuli were 1 sec in duration, and were synthesized at a sample rate of 10 kHz.

Eighteen naïve listeners participated in this experiment. For each series of stimuli for each voice, listeners heard all possible pairs of the 5 synthetic tokens, plus an equal number of pairs where stimuli were the same, for a total of 800 trials/listener. For each pair, listeners determined whether the stimuli were the same or different, and rated their confidence in their response on a 5 point scale ranging from "positive" to "wild guess." Testing took place in two sessions, each lasting about 45 minutes.

3.2. Results

For each series of stimuli and each voice, "same" and "different" responses were combined with confidence ratings to create 10-point scales ranging from "positive voices are the same" to "positive voices are different." These unfolded confidence ratings were then used to construct receiver operating characteristics (ROCs) for the different experimental conditions.

Values of A_z (the area under the ROC) were calculated to determine whether listeners could hear differences in the amounts of jitter, shimmer, and noise present in the stimuli. 99% confidence intervals around these values were also calculated. Confidence intervals that did not include the chance value of 0.5 indicated that the stimuli are reliably discriminable. These ROCs were used to interpret the method of adjustment data from Experiment 1. Listener responses in Experiment 1 that fall within one confidence interval of each other are considered to be in agreement, because they are indiscriminable.

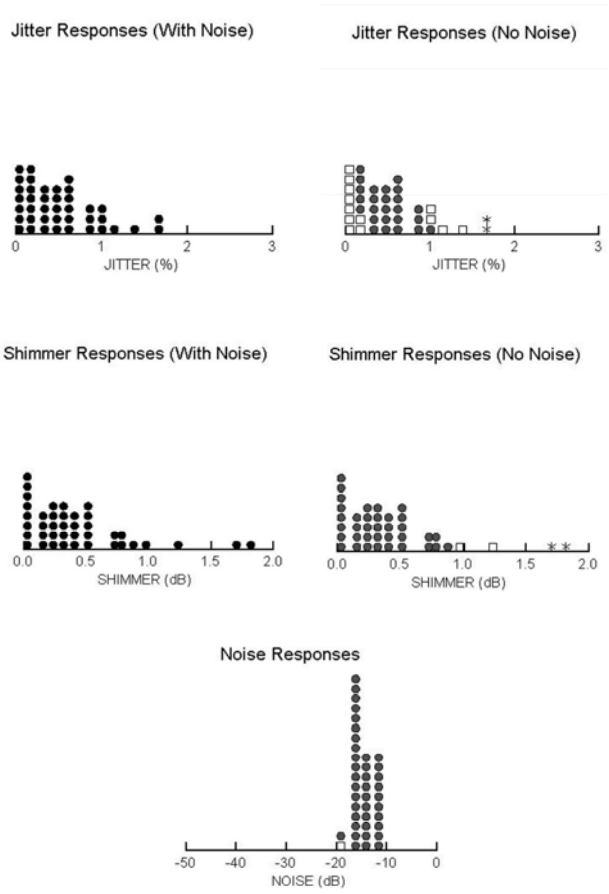

Figure 2: Distribution of responses for a single voice from method of adjustment task (Experiment 1). Filled circles represent indistinguishable responses; open squares represent responses that were minimally distinguishable from others; and stars represent perceptual outliers.

Responses for a representative voice are shown in Figure 2. In this figure, filled circles represent responses that were indiscriminable (area under the ROC < 0.5); open squares represent significantly but poorly discriminable responses (0.6 < area under the ROC < 0.7); and stars represent responses that were consistently discriminable from the majority of responses (area under the ROC > 0.8).

Listeners were unable to perceive differences in the amounts of jitter present for any voice when stimuli were synthesized with perceptually appropriate levels of noise, even when differences in jitter levels approached 3%. Some differences in jitter levels were perceptible

when stimuli were synthesized without additional noise, but across voices only 8/320 jitter responses (2.5%) were consistently distinguishable from the majority of responses. Thus, most of the variability in jitter responses observed in Experiment 1 is apparently due to listeners' insensitivity to differences in the amounts of jitter present in natural stimuli. Difference limens for jitter in realistic (noisy) voice contexts are greater than 1% for every voice studied in Experiment 1, and approach 3% for some voices.

Listeners were similarly unable to perceive differences between amounts of shimmer when stimuli were synthesized with perceptually appropriate levels of noise, even when differences in shimmer were as great as 2 dB. In Experiment 1, this corresponds to a total of 6/320 shimmer responses (1.9%) that were significantly but minimally discriminable from the majority of responses. Difference limens for shimmer in noisy contexts averaged about 1.3 dB, with a maximum value of about 2 dB.

Finally, listeners were almost never able to perceive differences between noise responses from Experiment 1. Twelve out of 320 responses in Experiment 1 (3.75%) were perceptual outliers or errors. This compares to an error rate of 2.5% for expert listeners performing the same task in [1]. Difference limens for noise averaged about 10 dB, although this value varies significantly with the shape of the spectrum of the harmonic part of the source (range = appx. 7-35 dB; smaller value = more sinusoidal source), also consistent with [1]. The present data thus replicate our previous finding that an analysis by synthesis approach to measuring vocal quality is a reliable way of measuring listeners' perceptions.

3.3. Discussion

Listeners are remarkably insensitive to the amounts of jitter and shimmer present in a voice. When asked to match the amounts of jitter and shimmer, listeners gave responses ranging over as much as 3% jitter and 2 dB shimmer. These values are large compared to differences usually treated as clinically or scientifically important. For example, measured jitter values for stimuli in Experiment 1 ranged from 0.1% (mild vocal pathology) to 2.57% (severe vocal pathology), and shimmer values ranged from 0.28 dB to 1.23 dB. These ranges are less than a difference limen for these variables, indicating that jitter and shimmer do not perceptually distinguish mild from severe vocal pathology. In contrast, measured values of the noise-to-signal ratio ranged from –40.4 dB (mild pathology) to –6.93 dB (severe pathology). This range is far greater than the difference limen of approximately 10 dB for noise.

Experimentally, far smaller differences in amounts of jitter and shimmer have been considered meaningful. For example, Karnell et al. [2] reported that differences in % jitter from different analysis systems averaged .01%, while differences in shimmer across systems averaged .085 dB. Bielamowicz et al. [3] reported differences between systems in measured % jitter of about 0.4-0.5%. Differences between systems in measured shimmer were less than 0.1 dB.

The fact that quality is multidimensional does not necessarily imply that it can be decomposed into separable dimensions. Results of these two experiments suggest that listeners judge aperiodicity by matching the overall amount of noise in the signal, rather than by decomposing aperiodicity into different, separable aspects. Noise levels are apparently separable from the overall voice pattern (as pitch and loudness are), and changes in noise levels produce predictable changes in the perceived quality of the voice. In contrast, jitter and shimmer do not appear to be separable from the overall auditory stream, but instead simply form part of the overall background pattern of a voice. This limits their utility as measures of quality, because listeners are insensitive to them within the auditory scene.

4. Conclusions

These results demonstrate the importance of determining the relevance of different acoustic attributes of a voice signal. This relevance derives from the relationship of an acoustic feature to some physiological or perceptual attribute of the voice. In the absence of such demonstrated relationships, acoustic attributes in and of themselves are not meaningful, and are therefore useless as measures of voice.

Finally, these results confirm the usefulness of psychoacoustic approaches to the study of voice quality, and of the method of adjustment approach in particular. Such approaches appear useful even when applied simultaneously to multiple facets of complex multidimensional stimuli.

5. Acknowledgements

This research was supported by grant DC01797 from the National Institute on Deafness and Other Communication Disorders. Synthesizer software was written by Brian Gabelman. Norma Antonanzas provided significant additional programming support. We also thank Jason Mallory, who recruited and patiently tested many of the listeners.

6. References

[1] Gerratt, B.R. and Kreiman, J., "Measuring Vocal Quality with Speech Synthesis", *J. Acoust. Soc. Am.*, 110:2560-2566, 2001.

[2] Karnell, M. P., Hall, K. D., and Landahl, K. L. "Comparison of Fundamental Frequency and Perturbation Measurements Among Three Analysis Systems", *Journal of Voice,* 9:383-393, 1995.

[3] Bielamowicz, S., Kreiman, J., Gerratt, B.R., Dauer, M.S., and Berke, G.S., "A Comparison of Voice Analysis Systems for Perturbation Measurement", *J. Speech Hear. Res.,* 39:126-134, 1996.