Perceptually-Motivated Modeling of Noise in Pathological Voices

Brian Gabelman*, Jody Kreiman*, Bruce R. Gerratt*, Norma Antonanzas-Barroso*, and Abeer Alwan**

*Division of Head and Neck Surgery and **Dept. of Electrical Engineering, UCLA, Los Angeles, CA 90095 USA

Abstract: To determine the perceptual importance of differences in noise characteristics, a sample of pathological voices was modeled using analysis by synthesis techniques. Spectral characteristics of noise were varied to create different synthetic versions of each voice sample. Expert listeners compared each synthetic stimulus to the original voice sample. The perceptual importance of differences in how vocal noise is synthesized will be discussed, as will the relative contributions of jitter, shimmer, and aspiration noise in modeling pathological phonation.

INTRODUCTION

Research investigating the correlation of acoustic measures of noise and the perception of pathological voice quality has consistently demonstrated a moderate association. However, this correlational approach cannot address basic questions concerning their cause and effect relationship, such as how carefully noise must be modeled to preserve the perceived pathological vocal quality, and what particular aspects of the noise are necessary to preserve natural vocal quality. To address these questions, an analysis by synthesis approach was applied to a sample of natural pathological vowels varying in severity of pathology.

STIMULI AND RECORDING PROCEDURES

The voices of randomly-selected speakers complaining of voice disorders were recorded as part of a phonatory function analysis. Speakers were recorded using a condenser microphone (Bruel & Kjaer) held a constant distance from the lips. They were asked to sustain the vowel /a/ for as long as possible. Voice samples were low-pass filtered at 8 kHz and digitized at 20 kHz. They were down-sampled to 10 kHz prior to analysis.

ANALYSIS AND SYNTHESIS TECHNIQUES

For each voice sample, source characteristics and vocal tract resonances were first analyzed and parameterized as described in (1). Noise characteristics were next determined from the original signal as follows (cf. 2). A short-time cepstral analysis was performed. The cepstral peak was identified in the quefrency domain, and a comb lifter was applied to remove the periodic component of the signal. The vocal tract response component was removed by a high-pass lifter, and an FFT was then performed on the residual signal. This FFT resulted in an estimate of the noise spectrum. Note that this technique separates periodic from aperiodic components of the signal, but does not necessarily separate aspiration noise from noise contributed by jitter and shimmer. Best estimates of jitter and shimmer were also made directly from the acoustic waveform when possible, using interactive software (3). It is unclear whether noise due to irregular vocal fold vibration can be separated from aspiration noise in practice (4, 5; but see also 6).

Estimated spectral noise was next modeled in various ways, ranging from a rather coarse approach (which mimicked the overall spectral slope of the noise, but neglected all fine details) to rather precise modeling (where both the precise spectral shape and temporal variability in noise levels were matched).

Modeled noise was then combined with the noise-free synthetic vocal signal. Noise amplitude was modulated by the glottal pulse. Several synthetic versions of each voice were created. The success of our noise-modeling efforts was assessed by recombining the unsmoothed noise spectrum with the source and vocal tract functions ("reconstructed" version). In order to assess the precision with which noise must be modeled, test

signals with a range of noise amplitude levels above and below the measured value were created for each signal. Spectral and statistical characteristics of the added noise were varied in a similar fashion. Finally, voices were synthesized by adding jitter and shimmer in addition to the noise. These procedures generated synthesized voices which matched the statistical characteristics of the original voice sample, but did not always provide precise models of the time-varying characteristics of the token under study.

PERCEPTUAL EVALUATION

Ten expert listeners participated in the perceptual experiments. They were asked to assess the similarity of each synthetic stimulus to the original voice sample. Similarity ratings were made on visual analog scales. Analyses will assess the perceptual importance of differences in how vocal noise is synthesized.

ACKNOWLEDGMENT

This research was supported by NIDCD grant DC01797.

REFERENCES

1. Gerratt, B.R., Kreiman, J., Antonanzas-Barroso, N., Gabelman, B., and Alwan, A., "Source modeling of severely pathological voices," *Proceedings of the ICA/ASA*, Seattle, WA, 1998.
2. de Krom, G., *J. Speech Hear. Res.* **36**, 254-266 (1993).
3. Bielamowicz, S., Kreiman, J., Gerratt, B.R., Dauer, M.S., and Berke, G.S., *J. Speech Hear. Res.* **39**, 126-134 (1996).
4. Hillenbrand, J., *J. Speech Hear. Res.* **30**, 448-461 (1987).
5. Hillenbrand, J., *J. Acoust. Soc. Am.* **83**, 2361-2371 (1988).
6. Michaelis, D., Gramss, T., and Strube, H.W., *Acustica* **83**, 700-706 (1997).