The estimation of signal-to-noise ratio in continuous speech for disordered voices

Yingyong Qi
Department of Speech and Hearing Sciences and Department of Electrical and Computer Engineering, University of Arizona, Tucson, Arizona 85721

Robert E. Hillman
Voice and Speech Laboratory, Massachusetts Eye and Ear Infirmary, Boston, Massachusetts, Department of Otology and Laryngology, Harvard Medical School, Communication Science and Disorders, MGH-Institute of Health Professions, and Research Laboratory of Electronics, Massachusetts Institute of Technology

Claudio Milstein
Voice and Speech Laboratory, Massachusetts Eye and Ear Infirmary, Boston, Massachusetts and Department of Speech and Hearing Sciences, University of Arizona, Tucson, Arizona 85721

(Received 3 November 1998; revised 29 December 1998; accepted 11 January 1999)

Presented is a method of estimating the signal-to-noise ratio (SNR) of continuous utterances for patients with various types of voice disorders that ranged in severity of dysphonia from mild to severe. The SNR is estimated based on the residual that is left after systematically removing the short- and long-term correlations that exist in the speech signal. Results indicate that the SNR is consistent with human perceptual judgments, particularly those that consistently differentiate close-to-normal versus highly disphonic voices. © *1999 Acoustical Society of America.* [S0001-4966(99)01604-5]

PACS numbers: 43.70.Gr [AL]

INTRODUCTION

In current research and clinical practice, acoustic analyses of speech signals often rely on vowel phonations that are sustained for several seconds. A number of methods have been developed to analyze a selected segments from such samples of sustained vowel phonation. These include perturbations of fundamental frequency and amplitude, and harmonics-to-noise ratio (Horii, 1980; Yumoto *et al.*, 1982; Qi and Hillman, 1997)

Sustained vowel phonation, however, is not necessarily a valid representation of an individual's vocal function during continuous speech. For example, real running speech involves constant and rapid adjustments of vocal mechanisms (e.g., rapid initiation and termination of voicing) that are not present during sustained phonation of a vowel. Thus, it would be desirable, and potentially more valid, to obtain estimates from continuous speech of acoustic parameters that are associated with abnormalities in voice quality. Toward this goal, we here introduce a method for acoustically estimating the signal-to-noise ratio (SNR) of continuous utterances, which we hope will be useful both for clinical and research-related evaluations of voice production. This method is evaluated by comparing its results with human perceptual evaluations.

I. METHODS

In the analysis proposed here, speech signals are decomposed into two components: a correlated/predictable component (signal) and an uncorrelated/unpredictable component (noise). The SNR, thus, defines the strength of the correlated component of a speech signal relative to the uncorrelated/unpredictable, noise component. The decomposition of the speech signal into correlated and uncorrelated components is accomplished by systematically removing existing correlations from the signal until the residual signal appears to be a random Gaussian (normally distributed) sequence (Schroeder and Atal, 1985). This approach is similar to the statistical procedures used in analysis of variance, where known variables are successively factored out until the remaining variations appear to be random with a normal distribution.

According to the acoustic theory of speech production, there are two known types of correlations present in a speech signal: a short-term correlation and a long-term correlation (Schroeder and Atal, 1985). Short-term correlation refers to the correlation/predictability of a signal on a sample-by-sample basis. Such a correlation is primarily associated with the resonances of the vocal tract. For example, the pattern of oscillation (resonance) within each fundamental period is predictable, i.e., the magnitude of the current sample could be predicted from the samples that immediately precede the current sample when the formants of vocal tract are known (Fant, 1981). This short-term predictability would be disrupted by the glottal input for the next cycle and/or by any random variations.

Long-term correlation refers to the correlation/predictability of the signal based on samples that do not immediately precede the current sample. Such a correlation is primarily associated with the quasi-periodical nature of voice production (Ramachandran and Kabal, 1989). For example, the signal characteristics around the beginning of each cycle would be predictable, to a certain extent, based on information from around the beginning of previous cycles. This pre-

diction would be disrupted by the onset/offset of voicing and/or by any random variations.

The decomposition of speech signals into short- and long-term correlations plus Gaussian noise has been successfully applied in telecommunication systems (Schroeder and Atal, 1985). In the code-excited linear prediction (CELP) based speech coders of some cellular phone systems, for example, only parameters related to the short- and long-term correlations are transmitted. Speech signals are reconstructed at the receiver by adding (filtering) random Gaussian noise using the transmitted short- and long-term correlation coefficients. Although it is necessary to synthesize a random noise in the receiver that has similar variance and temporal distribution as that in the transmitter, the unpredictable, noise component of speech is **not** transmitted in the cellular system. The adequacy of decomposing speech signals into short- and long-term correlated components plus a noise component is demonstrated by the fact that cellular phones provide adequate speech quality for normal communication.

A number of predictions could be made about the decomposition of speech signals. The residual signal, for example, should approximate a Gaussian process. This has been well demonstrated in previous publications (Schroeder and Atal, 1985). In this work, the proposed SNR was evaluated by comparing it to human perceptual ratings of a relatively large set of speech samples.

A. Subjects and recordings

Speech samples were recorded at the Voice and Speech Laboratory of Massachusetts Eye and Ear Infirmary. Eighty-seven subjects (40 men and 47 women) diagnosed with a wide variety of laryngeal voice pathologies provided the speech samples. Each subject was asked to read the Rainbow Passage at comfortable fundamental frequency and intensity levels. Audio recordings were made using a condenser microphone (Sennheiser) and a digital tape recorder (Tascom, DA-30) in a sound treated booth. The microphone was suspended a constant distance of 15 cm from the lips of each subject using a head-mounted device. All recordings were low-pass filtered ($f_c = 7.5$ kHz) and redigitized into a computer at a sampling rate of 16 kHz and a 16-bit A/D resolution. The first two sentences of the Rainbow Passage were used for subsequent acoustic analysis and perceptual evaluation.

B. Acoustic analysis

Linear prediction (LP) was used to determine both short- and long-term correlations (Markel and Gray, 1976; Ramachandran and Kabal, 1989). For short-term correlation, LP analysis was made on a window-by-window (no overlap) basis. The LP filter was obtained using a Hamming window with window length of 20 ms. The order of the LP filter was 14 (Markel and Gray, 1976). To remove short-term correlation, the original signal was inverse filtered by the LP filter using overlap save to ensure continuity during filter update. The residual signal of this LP inverse filtering was the short-term, decorrelated signal which was then further processed for long-term decorrelation.

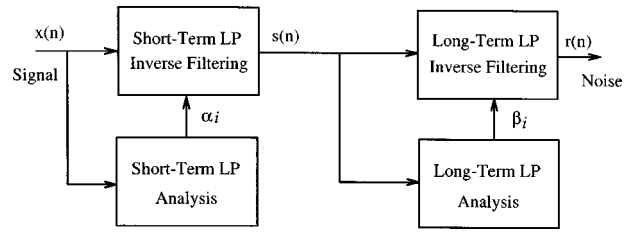

FIG. 1. Flow chart of the short- and long-term decorrelation process.

During long-term decorrelation, linear prediction was made based on samples that were not immediately preceding the current sample of the short-term decorrelated, residue signal. The window length for minimizing prediction error was 2.5 ms, which is long enough to include the pulselike peaks of short-term LP residual signals that often occur during voiced segments of speech. Because the exact location of the next residual peak varies somewhat from cycle to cycle, the closest sample used for making prediction was between 1.25 to 17.5 ms prior to the first sample to be predicted. Thus, this included the fundamental frequency range from 60–800 Hz in the predictive analysis. The LP filter that produced the minimal prediction error over this sliding range was chosen as the final long-term LP filter. The order of the filter was 3 (Ramachandran and Kabal, 1989). To remove the long-term correlation, the short-term decorrelated residual signal was inverse filtered by the long-term LP filter. Overlap save was used again to ensure continuity during filter update. The output of this second stage of inverse filtering was considered to be the final (short- and long-term decorrelated) noise component of the speech signal. A flow chart of the short- and long-term decorrelation processes is shown in Fig. 1. Example signals are shown in Fig. 2.

The final SNR was computed as the ratio of average rms amplitude between the original signal and its corresponding short- and long-term decorrelated signal. This ratio was reduced by one before converting it to dB scale because the original signal represents signal plus noise.

C. Perceptual evaluations

The recorded voice samples (87 in total) were perceptually rated by the same group of listeners using two different types of scales: a categorical scale and a continuous scale. The two ratings were made about three months apart to minimize any potential learning effects. Judges consisted of five speech pathologists with normal hearing (screened at 25 dB for speech frequencies) and extensive training and experience in the diagnosis and treatment of voice disorders. All ratings were accomplished using an interactive graphical user interface on a computer with stimuli presented over headphone. Stimuli consisted of the first two sentences of the Rainbow Passage.

In the categorical rating task, judges were asked to classify each voice sample as (1) normal, (2) mild, (3) mild-to-moderate, (4) moderate, (5) moderate to severe, (6) severe, or (7) aphonic. The judges were allowed to listen to each voice sample as many times as they wished before entering their response. For assessing intrajudge reliability, each

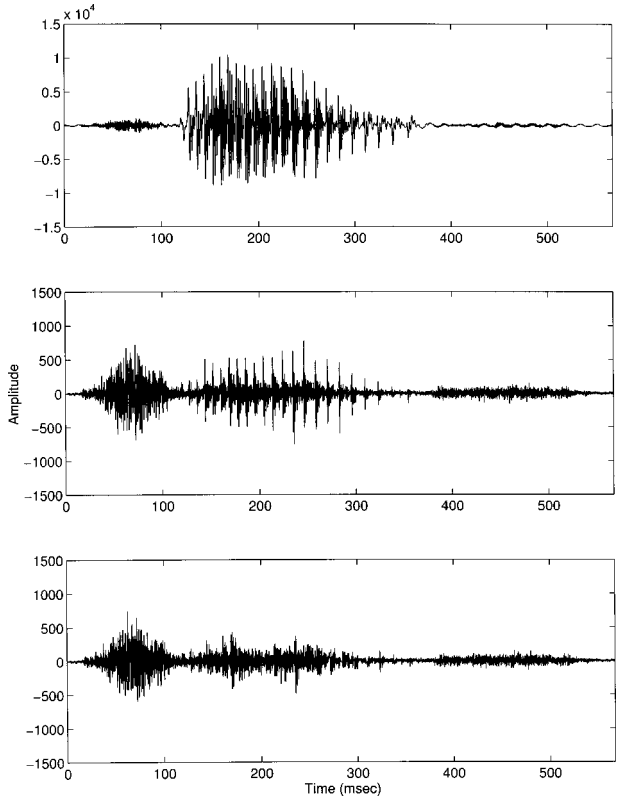

FIG. 2. Example of original signal (top), short-term decorrelated signal (middle), and short- and long-term decorrelated signal (bottom). The original signal was recorded from a male talker saying the word ''choice.''

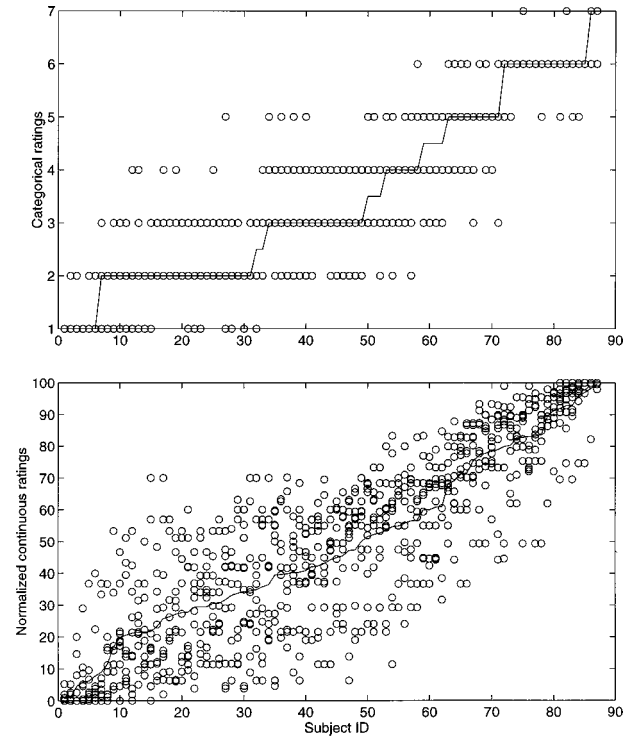

FIG. 3. Categorical ratings for all judges and voice samples with associated median values (top) and normalized, continuous ratings for all judges and voice samples with associated mean values (bottom).

judge repeated the entire rating session twice for a different random ordering of stimulus presentation with a period of at least 24 h in between rating sessions.

In the continuous rating task, judges were asked to use a number to describe the degree of perceived dysphonia relative to a standard voice sample. The standard voice sample was assigned a number of 100. Voices perceived to have more dysphonia than the standard, for example, would be given a rating of more than 100, and voices with less dysphonia, would be assigned a number less than 100. Raters were free to assign any value, as high or as low as they considered necessary. They also had access to the reference sample at all times, and were allowed to listen to each voice sample as many times as they wished before entering their response. As was the case for the categorical task, intrajudge reliability was assessed by repeating the entire rating session (stimuli in different random order) on a different day.

II. RESULTS AND CONCLUSIONS

The categorical ratings and their medians are shown in Fig. 3 (top) for each voice sample. Spearman correlations were computed for all ten judgments (5 judges×2 sessions). Results indicated that all correlations were significant ($p<0.001$). Intrajudge correlations ranged from 0.87 to 0.93 and interjudge correlations ranged from 0.82 to 0.91.

The continuous ratings and their means are shown in Fig. 3 (bottom) for each voice sample. Here, all scores were normalized to the range of 0–100 based on the maximum and minimum scores of a rating session. Pearson correlations were computed for all ten judgments (5 judges×2 sessions). Results indicate that all correlations were significant ($p<0.001$). Intrajudge correlations ranged from 0.93 to 0.96. Interjudge reliability was calculated using Cronbach's α (Cronbach, 1970). This statistics entails measuring the correlation between each individual listener's mean rating for each stimulus with the group mean of all the other listeners. Cronbach's α was 0.97, indicating adequate reliability among listeners in the continuous scaling task.

As shown, despite significantly high intra- and interjudge correlations, both categorical and continuous ratings extend (overlap) over a relatively large range for most voice samples. For example, a voice sample in the middle of the perceptual scale could have a categorical rating ranging from mild to severe or a continuous rating ranging from 25 to 75. This overlap, however, is minimal between samples that are rated as close-to-normal (normal and mild, 31 samples) and those that are rated as highly disphonic (moderate to severe, severe, and aphonic, 25 samples). A histogram of the categorical scores for these two subgroups (31+25=57 total samples) is shown in Fig. 4. As expected, the judges appeared inconsistent when viewing their performance across the entire set of voice samples, but they were able to differentiate between close-to-normal and highly disphonic voice samples quite well.

The median categorical ratings and mean continuous ratings are shown in Fig. 5 as a function of the computed SNRs for all subjects. The Spearman correlation between the categorical ratings and SNRs ($r=-0.76$, $p<0.001$) and the Pearson correlation between the continuous ratings and SNRs ($r=-0.78$, $p<0.001$) are both statistically significant, but relatively low in terms of the amount of variation

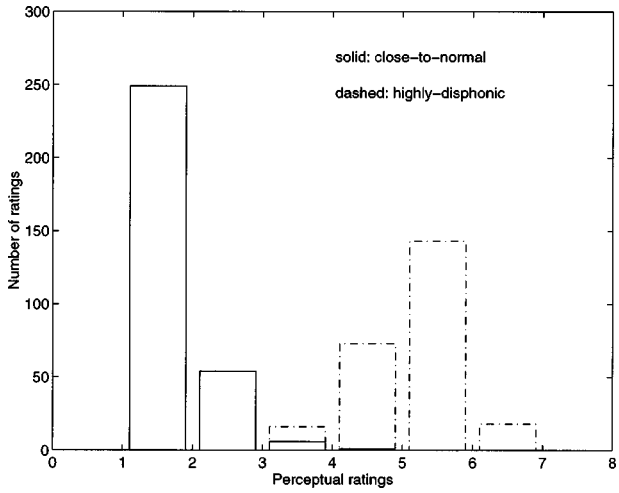

FIG. 4. Histogram for voice samples whose perceptual ratings fell in either the close-to-normal or highly disphonic range.

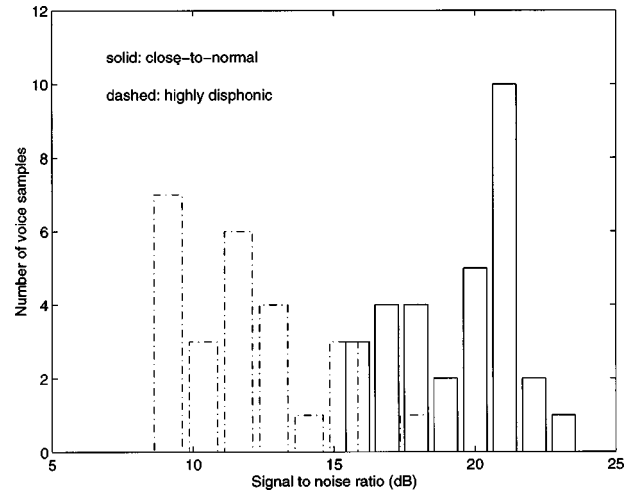

FIG. 6. Histogram of SNR values for voice samples whose perceptual ratings are either in the close-to-normal or highly disphonic range.

actually accounted for ($r^2 < 0.61$). By way of comparison, the SNRs for samples that are rated as close-to-normal or highly disphonic are shown in Fig. 6. There is a clear separation (54/57=95%) in SNR between these two groups, indicating that the computed SNRs are in agreement with the perceptual ratings when the ratings are made consistently.

These results seem to indicate that the proposed SNR, similar to the harmonics-to-noise ratio for vowels, has a moderate degree of correlation with perceptual ratings of human listeners (Yumoto *et al.*, 1982). Obviously, experiments undertaken here merely provide some preliminary support for the proposed SNR measurement. Further experiments are necessary to more rigorously establish the relationship between SNR and specific aspects of pathological voice/speech production/perception. To date, comprehensive understanding and agreement on methods for evaluating pathological voice production/perception is lacking. The SNR measurement described here is developed largely based on estimations of short- and long-term correlations that have been successfully applied to modern telecommunication systems. It represents a first attempt to directly quantify acoustic properties of continuous utterance for disordered voices. It is our hope that the proposed SNR measure could be developed into a useful tool for clinical and research-related voice assessment.

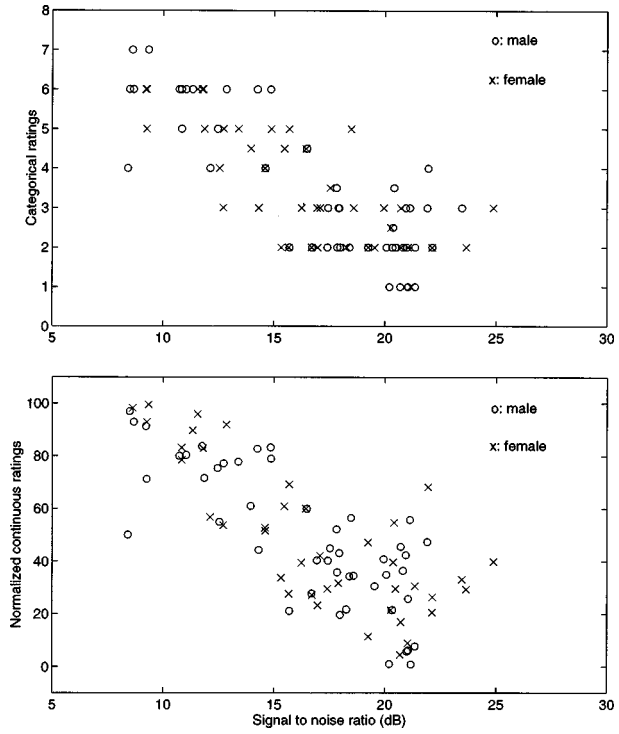

FIG. 5. Categorical (top) and continuous (bottom) ratings as a function of the computed SNR.

ACKNOWLEDGMENT

This work was supported, in part, by grants from the National Institute of Deafness and Other Communication Disorders: DC00266, Objective Assessment of Vocal Hyperfunction.

Cronbach, L. (**1970**). *Essentials of Psychological Testing* (Harper and Row, New York).

Fant, G. (**1981**). ''The source filter concept in voice production,'' STL-QPSR **1**, 21–37.

Horii, Y. (**1980**). ''Vocal shimmer in sustained phonation,'' J. Speech Hear. Res. **23**, 202–209.

Markel, J., and Gray, A. (**1976**). *Linear Prediction of Speech* (Springer-Verlag, Berlin).

Qi, Y., and Hillman, R. (**1997**). ''Temporal and spectral estimations of harmonics-to-noise ratio in human voice signals,'' J. Acoust. Soc. Am. **102**, 537–543.

Ramachandran, R. P., and Kabal, P. (**1989**). ''Pitch prediction filters in speech coding,'' IEEE Trans. Acoust., Speech, Signal Process. **37**, 467–478.

Schroeder, M., and Atal, B. (**1985**). ''Code excited linear prediction (CELP): High quality speech at very low bit rates,'' in Proc. ICASSP, pp. 937–940.

Yumoto, E., Gould, W., and Baer, T. (**1982**). ''Harmonics-to-noise ratio as an index of the degree of hoarseness,'' J. Acoust. Soc. Am. **71**, 1544–1550.