

Perception of aperiodicities in synthetically generated voices

James Hillenbrand^{a)}

RIT Research Corporation and Department of Computer Science, Rochester Institute of Technology,
75 Highpower Road, Rochester, New York 14623-3435

(Received 2 December 1987; accepted for publication 15 February 1988)

The purpose of this study was to investigate univariate relationships between perceived dysphonia and variation in pitch perturbation, amplitude perturbation, and additive noise. A time-domain, pitch-synchronous synthesis technique was used to generate sustained vowels varying in each of the three acoustic dimensions. A panel of trained listeners provided direct magnitude estimates of roughness in the case of the stimuli varying in pitch and amplitude perturbation, and breathiness in the case of the stimuli varying in additive noise. Very strong relationships were found between perceived roughness and either pitch or amplitude perturbation. However, unlike results reported previously for nonspeech stimuli, the subjective quality associated with pitch perturbation was quite different from that associated with amplitude perturbation. Results also showed that perceived roughness was affected not only by the amount of perturbation, but also by the degree of correlation between adjacent pitch or amplitude values. A strong relationship was found between perceived breathiness and signal-to-noise ratio. Contrary to previous findings, there was no interaction between signal-to-noise ratio and the amount of high-frequency energy in the periodic component of the stimulus: Stimuli with similar signal-to-noise ratios received similar ratings, regardless of differences in the spectral slope of the periodic component.

PACS numbers: 43.70.Dn, 43.71.Gv, 43.72.Ja

INTRODUCTION

There is very good agreement among voice clinicians and voice scientists of the need for objective measurements that would serve to quantify both the type and severity of dysphonia that accompanies a wide range of laryngeal disorders (e.g., Aronson, 1980; Davis, 1981; Fritzell *et al.*, 1977; Hammarberg *et al.*, 1980; Hirano, 1981; Jensen, 1965). In an effort to achieve this goal, a number of investigators have made detailed acoustic measurements of a wide variety of acoustic properties associated with laryngeal disorders. Many of the acoustic techniques that have been investigated are based on the finding that dysphonic voices tend to show larger than normal deviations from perfect periodicity. As a number of investigators have pointed out, however, acoustic measurements are useful only if they can be related to specific diagnostic categories, or to meaningful perceptual dimensions. For example, Hammarberg *et al.* (1980) commented that "... acoustic measurements do not make sense on their own ... [but] must be related to perceptual characteristics in order to be clinically useful" (p. 32).

The task of relating acoustic properties to perceptual dimensions has not proven to be a simple one. The most common approach to this problem has involved the use of correlational techniques to determine relationships between the perceived *degree* and/or *type* of dysphonia and variation in specific acoustic dimensions (e.g., Kempster, 1984; Kempster and Kistler, 1984; Kojima *et al.*, 1980; Murry *et al.*, 1977; Prosek *et al.*, 1984; Smith *et al.*, 1978; Yanagihara, 1967; Yumoto *et al.*, 1982). As will be discussed in greater detail below, the interpretability of these studies has been limited by the inability to control variation in specific acous-

tic dimensions in naturally occurring voices. Of particular importance are the inability to control the range of variation on particular acoustic dimensions and the degree of intercorrelation among individual acoustic properties.

The present study represents an attempt to address this problem by studying the perceptual characteristics of synthetically generated voices. The three parameters chosen for study were pitch perturbation, amplitude perturbation, and additive noise. The long-term goal of this work is to learn how these and other acoustic parameters *combine* to affect a listener's overall impression of vocal quality. As an initial step toward this goal, the present study was designed to examine the *univariate* relationships between variation in each of these acoustic dimensions and the perception of dysphonic vocal quality.

A. Pitch perturbation

Pitch perturbation, or "vocal jitter," is defined as cycle-to-cycle variation in voice fundamental frequency (F_0). All human voices contain a certain amount of vocal jitter, and synthesis studies have shown that a minimum amount of jitter is required for a voice to sound natural (Gill, 1961; Holmes, 1962; Rozsypal and Millar, 1979; Schroeder, 1961). Jitter values in normal voices are generally less than about 1.0% (Hollien *et al.*, 1973; Horii, 1979, 1982; Jacob, 1968; Kempster, 1984; Simon, 1927). Lieberman (1963) was the first to report that dysphonic voices tend to show unusually large cycle-to-cycle variations in F_0 . This basic finding has been confirmed in several studies using a wide variety of analysis and computational procedures (Davis, 1976; Deal and Emanuel, 1978; Hecker and Krueger, 1971; Kitajima *et al.*, 1975; Koike, 1973; see Murry and Doherty, 1980, for a negative finding).

^{a)} Present address: Department of Speech Pathology and Audiology, Western Michigan University, Kalamazoo, MI 49008.

Jitter does not appear to be a factor controlling perceived roughness in naturally produced normal voices (Heiberger and Horii, 1982), but there is some evidence suggesting that jitter is correlated with perceived roughness in disordered voices. Weak- to moderate-strength correlations between pitch perturbation and perceived roughness were reported in studies of disordered speakers by Deal and Emanuel (1978), Lieberman (1963), and Takahashi and Koike (1975). Smith *et al.* (1978) reported a relatively weak ($r = 0.55$) nonsignificant correlation between jitter and perceived roughness in a group of esophageal speakers.

As mentioned previously, two important interpretive limitations of these studies using naturally produced voices concern the inability to control jitter values without affecting values on other dimensions and the inability to control the range of variation on any of the acoustic dimensions. The first of these problems is especially important because several acoustic measurement studies of disordered voices have reported significant intercorrelations among individual acoustic parameters such as pitch perturbation, amplitude perturbation, and additive noise (Davis, 1976; Deal and Emanuel, 1978; Heiberger and Horii, 1982; Horii, 1980; Kempster, 1984; Kempster and Kistler, 1984; Yumoto *et al.*, 1984). Further, a recent methodological study suggests that the acoustic analysis techniques that have been used to measure perturbation are not always able to discriminate among various sources of aperiodicity (Hillenbrand, 1987).

Because of the difficulty in interpreting the results of perception studies using naturally produced voices, several studies have examined the perception of synthetically generated signals. Wendahl (1963, 1966a,b) synthesized sawtooth waves varying in jitter and mean F_0 . Results of paired comparison listening tests showed that roughness judgments correlated strongly with pitch perturbation. Wendahl's results also showed that, for a given jitter size, a signal with a low F_0 tended to sound more rough than a signal with a high F_0 (see, also, Coleman, 1969), a finding which has also been reported for naturally produced voices (Deal and Emanuel, 1978; Heiberger and Horii, 1982). Heiberger and Horii (1982) also reported a strong correlation between jitter and perceived roughness using synthesized triangular waves. To date, no study has examined the relationship between jitter and perceived roughness in synthetically generated voice signals.

B. Amplitude perturbation

Amplitude perturbation, or "vocal shimmer," is defined as cycle-to-cycle variation in voice amplitude. Shimmer values in normal voices are generally less than about 0.7 dB (Horii, 1980, 1982; Kempster, 1984; Robbins, 1981). Using a calculation method based on successive differences from a three-point moving average, Kitajima and Gould (1976) reported that amplitude perturbation values from a group of dysphonic subjects were significantly larger than those for a nondisordered control group. Similar findings were reported by Davis (1981) using slightly different calculation methods.

Relatively little is known about the relationship between amplitude perturbation and the perception of dysphonic vo-

cal quality in naturally produced voices. Takahashi and Koike (1975) reported that "breathiness" ratings correlated with amplitude perturbation and "roughness" ratings correlated both with pitch perturbation ($r = 0.55$) and amplitude perturbation ($r = 0.72$). Deal and Emanuel (1978) made nonsequential measures of pitch and amplitude perturbation from a group of normal speakers who were asked to simulate rough voice quality and from a group of speakers with various laryngeal pathologies. Results showed significant correlations between listener ratings of roughness and measures of both pitch and amplitude perturbation. On the basis of multiple regression analyses, Deal and Emanuel concluded that, "... cyclic peak amplitude variability may provide a better index of perceived roughness than cyclic period variation ..." (p. 250). A reanalysis of these results by Nichols (1979) reached the same conclusion using partial correlation techniques.

Studies by Wendahl (1966a,b) and Heiberger and Horii (1982) reported strong relationships between amplitude perturbation and perceived roughness in synthetically generated nonspeech signals. Wendahl reported that the roughness percept resulting from the introduction of amplitude perturbation in sawtooth waveforms was very similar to that associated with pitch perturbation:

"... it is interesting to note that the roughness generated by these different procedures [i.e., pitch and amplitude perturbation] results in such similar auditory experiences. Some highly trained listeners were able to distinguish between the two types of stimuli, but the writer, who has had years of listening experience with such stimuli, is able to discriminate between the program types only at the extremes of the continuum ..." (Wendahl, 1966b, p. 106).

Heiberger and Horii studied perceptual trading relations between pitch and amplitude perturbation by synthesizing triangular waves varying in jitter, shimmer, or both jitter and shimmer. The results suggested that the perceptual effects of jitter and shimmer are, in some sense, equivalent. For example, a stimulus with 2.0% jitter was judged to be approximately equivalent in roughness to a 1.0-dB shimmer stimulus. The results also suggested that the effects of jitter and shimmer are additive; for example, a stimulus containing both 2.0% jitter and 1.5-dB shimmer sounded more rough than either a 2.0% jitter stimulus or a 1.5-dB shimmer stimulus.

C. Additive noise

The term additive noise is generally used to refer to the acoustic by-product of turbulence generated at the glottis. A number of studies have reported that noise levels in dysphonic voices tend to be higher than those in normal voices and that noise measurements correlate with subjective ratings of dysphonia (Deal and Emanuel, 1978; Emanuel and Sansone, 1969; Kojima *et al.*, 1980; Lively and Emanuel, 1970; Sansone and Emanuel, 1970; Yanagihara, 1967; Yumoto *et al.*, 1982, 1984). A straightforward interpretation of the perceptual effects of additive noise is complicated by the presence of relatively strong intercorrelations among measures of perturbation and additive noise (Deal and Emanuel,

1978; Kempster, 1984; Kempster and Kistler, 1984; Yumoto *et al.*, 1984) and by the presence of very strong measurement interactions among these variables (Hillenbrand, 1987).

Very little work has been done on the synthesis and perception of stimuli varying in additive noise. In a study that is described very briefly, Yanagihara (1967) mixed filtered and unfiltered naturally produced sustained vowels with various types of bandpass filtered noise. The signal-to-noise ratios were held constant and the primary purpose of the study was to determine the relationship between perceived dysphonia and the spectral properties of the periodic and aperiodic components of the stimuli. Yanagihara reported a strong relationship between the loss of high-frequency harmonics and perceived dysphonia: "Even if the relative intensity of the noise components and the harmonic components remain unchanged, the loss of high-frequency harmonics results in an increase of the degree of perceived dysphonia" (p. 538). Yanagihara's results are consistent with Froekjaer-Jensen and Prytz (1976), who reported an increase in high-frequency energy in long-term average spectrum measurements following treatment for voice disorders (see, also, Hammarberg *et al.*, 1980; Fritzell *et al.*, 1977; Gauffin and Sundberg, 1977).

The present study was designed to extend the work of Wendahl (1963, 1966a,b) and Heiberger and Horii (1982) in studying the relations between perturbation and perceived roughness in synthetically generated signals and to extend the work of Yanagihara (1967) in studying the perceptual effects of additive noise. The experiments on the perception of perturbation were designed primarily to address two limitations of previous research on jitter and shimmer synthesis. First, the present study was designed to examine roughness-perturbation relations in synthetically generated voices, rather than the nonspeech waveforms used in the studies by Wendahl and Heiberger and Horii. Second, unlike the previous perturbation synthesis studies, the present study used synthesis techniques that attempted to model the sequential properties of cycle-to-cycle pitch and amplitude change in naturally produced voices. The purpose of the additive noise experiments was to examine the relationship between noise level and perceived breathiness over a wide range of signal-to-noise ratios and to test for possible interactions between signal-to-noise ratio and energy levels in high-frequency harmonics over a broad range of signal-to-noise ratios.

I. EXPERIMENT 1

A. Methods

1. Synthesis technique

Stimuli for all the experiments described in this article were generated with a pitch-synchronous synthesis program called VSYN (Wilde and Martens, 1985; modeled after Wilde *et al.*, 1986). The program was designed to generate sustained vowels differing in jitter, shimmer, additive noise, and mean F_0 . The first step in the synthesis process involved using Klatt's (1980) formant synthesizer to generate a *single* pitch pulse with formant frequency characteristics appropriate for whatever vowel quality is desired. Stimuli for the present study used a formant frequency pattern appropriate

for the vowel [a] ($F_1 = 720$, $F_2 = 1240$, $F_3 = 2400$, $F_4 = 3300$, $F_5 = 3700$). The pitch pulse was generated with a 40-kHz sample frequency, 12 bits of amplitude resolution, and consisted of 512 data points (12.8 ms). As shown in Fig. 1, sustained vowels were synthesized by stringing together the individual damped oscillations produced by the Klatt synthesizer. The F_0 was controlled by adjusting the interval between the onset of one damped oscillation and the onset of the next oscillation. For fundamental periods that are less than 12.8 ms, the end of one damped oscillation will overlap with the beginning of the next. This effect was accounted for simply by adding the tail end of one damped oscillation into the beginning of the next. In Fig. 1, the onset-to-onset interval was fixed at 8 ms, producing a vowel with a constant F_0 of 125 Hz. Pitch perturbation was controlled by introducing specific amounts of variability in the onset-to-onset intervals.

Although not used in experiment 1, VSYN controls amplitude perturbation by scaling each pitch pulse individually to achieve the desired amount of amplitude variability. Additive noise can be controlled by the appropriate scaling and point-for-point addition of a separate noise signal.

Using a method such as this to synthesize stimuli that differ only in pitch perturbation is problematic since, as discussed in detail in a related article (Hillenbrand, 1987), amplitude perturbation is produced as a side effect of pitch perturbation. For the stimuli that were intended to differ only in pitch perturbation, this artifact was removed by a separate program that measured the intensity of individual pitch pulses and scaled all pitch pulses to the same rms value.

2. Random-number generation

A random-number generator of some type is needed to produce the sequence of F_0 and/or pitch-pulse amplitude values that control the synthesizer. The random-number

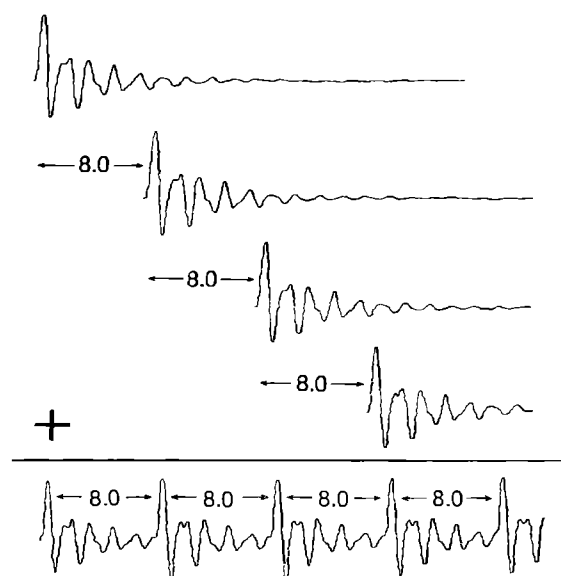


FIG. 1. Pitch-synchronous, time-domain synthesis technique used by VSYN.

(i.e., the standard deviation of the distribution cycle-to-cycle differences in fundamental period with the sign retained).

Equating the stimuli for either fundamental period standard deviation or perturbation factor increased rather than decreased differences in roughness magnitude between the correlated and uncorrelated continua. When the stimuli were equated for fundamental period standard deviation, the average difference in roughness magnitude between correlated and uncorrelated signals was 20.4%; for the perturbation factor, the difference was 29.4%. Koike's (1973) "relative average perturbation," which uses a three-point moving average, produced results that were very similar to the mean jitter data shown in Fig. 3. It is also interesting to note that the uncorrelated signals had substantially higher values of directional jitter than the correlated signals (72.7% vs 40.6%). The fact that the correlated signals sounded more rough than the uncorrelated signals would seem to indicate that directional jitter does not play a role in roughness perception.

In general, our preliminary conclusion from the comparison between the correlated and uncorrelated signals is that the standard deviation of signed jitter, or Davis' (1981) PPQ, shows a stronger relationship to perceived roughness than other methods of representing pitch perturbation. However, none of the calculation methods that were used eliminated the difference in perceived roughness between the correlated and uncorrelated signals.

II. EXPERIMENT 2: PERCEPTION OF AMPLITUDE PERTURBATION

A. Methods

VSYN was used to synthesize two 22-member shimmer continua using methods that were analogous to those used to create the jitter continuum in experiment 1. As in the pitch perturbation experiment, one continuum was created using the modified $1/f$ random-number generator and the other was created using a standard white-noise generator. The stimuli along each continuum varied from 0.0–2.6 dB and were spaced at 0.1-dB increments from 0.0–1.0 dB and at 0.2-dB increments from 1.0–2.6 dB. The decision to restrict the continuum to the range below 2.6 dB was somewhat arbitrary and was based on the increasingly unnatural perceptual quality of the synthesized signals as shimmer values approached about 2.0 dB. All stimuli were 1.0 s in duration and were synthesized at 40 kHz, with a constant F_0 of 130 Hz. As in the previous experiment, the stimuli were gated on and off with a 20-ms cosine function and all stimuli on the continuum were equated for overall rms intensity.

Subjects consisted of the same ten listeners who participated in experiment 1. As in the previous experiments, subjects were asked to rate the stimuli on the degree of perceived roughness. Each of the 44 stimuli was presented 16 times in pseudorandom order. The first 132 trials were considered to be practice and these data were not included in the analysis. Methods used for stimulus presentation were identical to experiment 1.

B. Results and discussion

The function relating shimmer to perceived roughness is shown in Fig. 5. The smooth curve is a second-order polynomial in the case of the correlated continuum and a fourth-order polynomial in the case of the uncorrelated continuum. As was true for the pitch-perturbation data, the signals that were produced from correlated sequences sounded more rough than signals with the same mean perturbation values that were produced from uncorrelated sequences. For the data in Fig. 5, the average difference in roughness magnitude between correlated and uncorrelated signals with the same mean shimmer value was 27.1%. This value is nearly three times larger than the difference that was observed between correlated and uncorrelated stimuli for the pitch-perturbation continua.

Unlike the pitch-perturbation data, this discrepancy between the correlated and uncorrelated signals does not seem to be related to the choice of perturbation calculation methods. In general, stimuli that were matched for mean shimmer tended to show very similar ratings when perturbation was measured using other calculation methods, such as pitch-pulse amplitude standard deviation, amplitude-perturbation quotient (the amplitude analog of PPQ), standard deviation of signed shimmer, and Koike's (1973) relative average perturbation.

One important finding of this experiment that cannot be observed in Fig. 5 concerns the subjective quality of the stimuli varying in amplitude perturbation. Recall that previous synthesis research with nonspeech signals suggested that amplitude perturbation produced a sensation of roughness that was virtually indistinguishable from that produced by pitch perturbation (Wendahl, 1966b). Although subjects in the present experiment were asked to rate the stimuli on

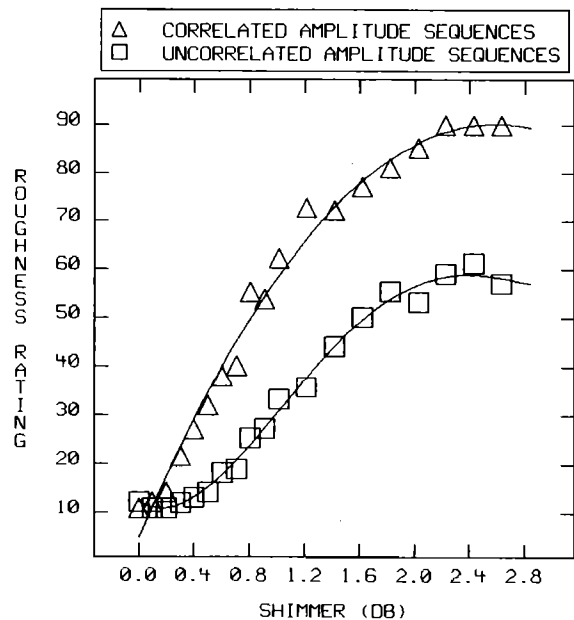


FIG. 5. Perceived roughness as a function of shimmer for correlated and uncorrelated pitch-pulse amplitude sequences.

themselves to be experienced in the evaluation and treatment of voice disorders. These same subjects participated in experiments 2–4 as well. The order of presentation of the four experiments was counterbalanced across subjects, with one exception: All subjects participated in experiment 3 (comparison of pitch and amplitude perturbation) following their participation in experiment 1 (perception of pitch perturbation) and experiment 2 (perception of amplitude perturbation).

B. Results and discussion

Results are shown in Fig. 3, which plots normalized roughness magnitude as a function of percent jitter, pooled across all ten listeners. Direct magnitude estimates were rescaled separately for each subject so that the numbers ranged from 10–90. The smooth curves are third-order polynomials that were fit to the data. Ignoring for the moment the difference between the correlated and uncorrelated stimuli, it can be seen that there is a very strong relationship between pitch perturbation and perceived roughness. The compression of the function at the high end of the jitter continuum is consistent with Heiberger and Horii (1982), who reported that, "... beyond a certain point, relatively large increases in [jitter] did not result in similarly large increases in roughness level ..." (p. 321). However, in Heiberger and Horii's nonspeech data, the change in slope occurred between jitter values of 5.0% and 10.0%, much larger than the value of approximately 2.0% found in the present study. Although this discrepancy might reflect differences in the perception of triangular waves versus the more harmonically rich voice signals used in the present study, there are two other possibilities. The stimuli used by Heiberger and Horii were higher in mean F_0 (165 vs 130 Hz used in the present study) and were presented to subjects over earphones rather than a

loudspeaker. Both of these differences would be expected to make the Heiberger and Horii stimuli sound less rough than the stimuli used in the present study (Wendahl, 1963, 1966a,b; Coleman, 1969; Wilde *et al.*, 1986), which might have the effect of moving the entire roughness-perturbation function to the right.

The other obvious feature of the data in Fig. 3 is that the stimuli generated from the correlated period sequences were perceived as more rough than the stimuli generated from the uncorrelated period sequences. The differences in roughness magnitude for a given jitter value averaged 9.3% and were highly significant ($t = 26.0$, $df = 29$, $p < 0.01$). It is important to note, however, that stimuli on the correlated and uncorrelated continua were equated for mean jitter (the average absolute difference in fundamental period between adjacent pitch pulses), but were not necessarily matched in terms of other calculation methods that have been used to represent pitch perturbation. For example, stimuli from the correlated continuum generally had larger values of fundamental period standard deviation (Deal and Emanuel, 1978) and larger mean jitter values when calculations were made from either a three-point moving average (Koike, 1973) or a five-point moving average (Davis, 1981).

Figure 4 shows the roughness-perturbation function for the correlated and uncorrelated continua, but with the stimuli equated for Davis' (1981) "pitch perturbation quotient" (PPQ), which uses a five-point moving average. It can be seen that the differences in perceived roughness between the correlated and uncorrelated signals are reduced significantly, although not eliminated entirely. For the data in Fig. 4, the average difference in roughness magnitude between the correlated and uncorrelated signals was 4.5%. Very similar results were found when the correlated and uncorrelated signals were equated for the standard deviation of signed jitter

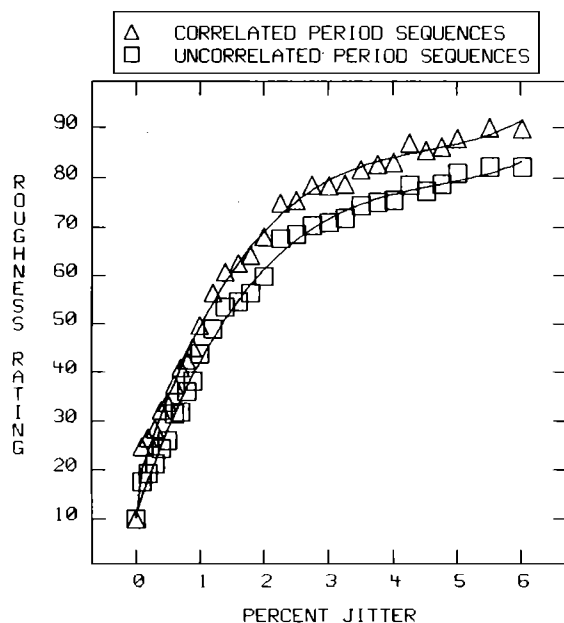


FIG. 3. Perceived roughness as a function of jitter for correlated and uncorrelated period sequences.

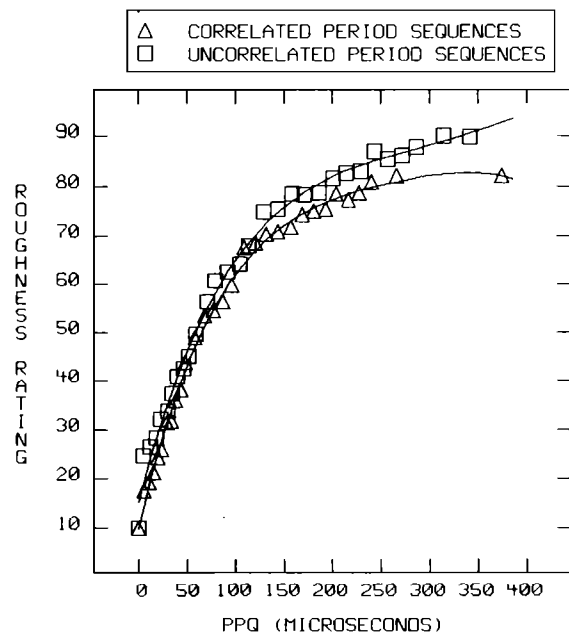


FIG. 4. Perceived roughness as a function of jitter for correlated and uncorrelated period sequences with stimuli equated for PPQ.

(i.e., the standard deviation of the distribution cycle-to-cycle differences in fundamental period with the sign retained).

Equating the stimuli for either fundamental period standard deviation or perturbation factor increased rather than decreased differences in roughness magnitude between the correlated and uncorrelated continua. When the stimuli were equated for fundamental period standard deviation, the average difference in roughness magnitude between correlated and uncorrelated signals was 20.4%; for the perturbation factor, the difference was 29.4%. Koike's (1973) "relative average perturbation," which uses a three-point moving average, produced results that were very similar to the mean jitter data shown in Fig. 3. It is also interesting to note that the uncorrelated signals had substantially higher values of directional jitter than the correlated signals (72.7% vs 40.6%). The fact that the correlated signals sounded more rough than the uncorrelated signals would seem to indicate that directional jitter does not play a role in roughness perception.

In general, our preliminary conclusion from the comparison between the correlated and uncorrelated signals is that the standard deviation of signed jitter, or Davis' (1981) PPQ, shows a stronger relationship to perceived roughness than other methods of representing pitch perturbation. However, none of the calculation methods that were used eliminated the difference in perceived roughness between the correlated and uncorrelated signals.

II. EXPERIMENT 2: PERCEPTION OF AMPLITUDE PERTURBATION

A. Methods

VSYN was used to synthesize two 22-member shimmer continua using methods that were analogous to those used to create the jitter continuum in experiment 1. As in the pitch perturbation experiment, one continuum was created using the modified $1/f$ random-number generator and the other was created using a standard white-noise generator. The stimuli along each continuum varied from 0.0–2.6 dB and were spaced at 0.1-dB increments from 0.0–1.0 dB and at 0.2-dB increments from 1.0–2.6 dB. The decision to restrict the continuum to the range below 2.6 dB was somewhat arbitrary and was based on the increasingly unnatural perceptual quality of the synthesized signals as shimmer values approached about 2.0 dB. All stimuli were 1.0 s in duration and were synthesized at 40 kHz, with a constant F_0 of 130 Hz. As in the previous experiment, the stimuli were gated on and off with a 20-ms cosine function and all stimuli on the continuum were equated for overall rms intensity.

Subjects consisted of the same ten listeners who participated in experiment 1. As in the previous experiments, subjects were asked to rate the stimuli on the degree of perceived roughness. Each of the 44 stimuli was presented 16 times in pseudorandom order. The first 132 trials were considered to be practice and these data were not included in the analysis. Methods used for stimulus presentation were identical to experiment 1.

B. Results and discussion

The function relating shimmer to perceived roughness is shown in Fig. 5. The smooth curve is a second-order polynomial in the case of the correlated continuum and a fourth-order polynomial in the case of the uncorrelated continuum. As was true for the pitch-perturbation data, the signals that were produced from correlated sequences sounded more rough than signals with the same mean perturbation values that were produced from uncorrelated sequences. For the data in Fig. 5, the average difference in roughness magnitude between correlated and uncorrelated signals with the same mean shimmer value was 27.1%. This value is nearly three times larger than the difference that was observed between correlated and uncorrelated stimuli for the pitch-perturbation continua.

Unlike the pitch-perturbation data, this discrepancy between the correlated and uncorrelated signals does not seem to be related to the choice of perturbation calculation methods. In general, stimuli that were matched for mean shimmer tended to show very similar ratings when perturbation was measured using other calculation methods, such as pitch-pulse amplitude standard deviation, amplitude-perturbation quotient (the amplitude analog of PPQ), standard deviation of signed shimmer, and Koike's (1973) relative average perturbation.

One important finding of this experiment that cannot be observed in Fig. 5 concerns the subjective quality of the stimuli varying in amplitude perturbation. Recall that previous synthesis research with nonspeech signals suggested that amplitude perturbation produced a sensation of roughness that was virtually indistinguishable from that produced by pitch perturbation (Wendahl, 1966b). Although subjects in the present experiment were asked to rate the stimuli on

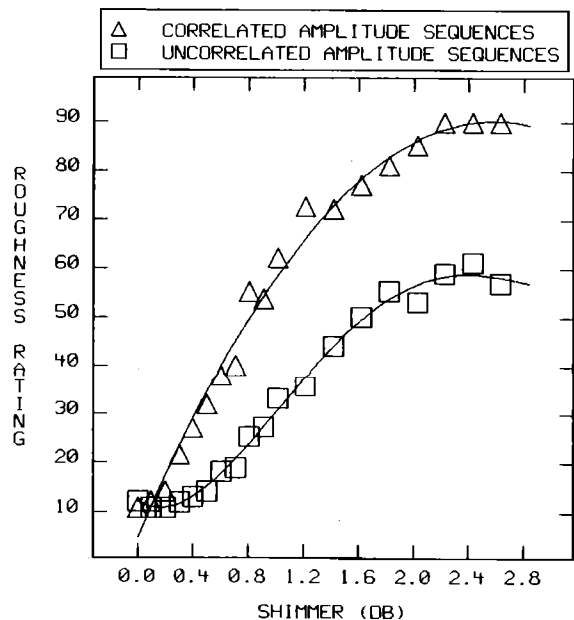


FIG. 5. Perceived roughness as a function of shimmer for correlated and uncorrelated pitch-pulse amplitude sequences.

roughness, this term is almost certainly not a good description of the perceptual quality of the stimuli varying in amplitude perturbation. Unlike the results for sawtooth waves reported by Wendahl, the perceptual quality of the stimuli varying in amplitude perturbation in the present study was quite different from that produced by pitch perturbation. When asked to provide verbal descriptions of the stimuli, subjects generally commented that the signals toward the high end of the shimmer continuum had an unnatural "popping" quality. For example, one subject commented that the stimuli sounded as though they were being played through a loudspeaker with a loose wire and another compared the signals to speech played over a radio during an electrical storm. By contrast, stimuli toward the high end of the pitch-perturbation continuum are perceived as very rough; however, with the exception of the very high jitter values from the correlated continuum, the stimuli sounded as though they could have been produced by a talker with a severely disordered voice.

III. EXPERIMENT 3: COMPARISON OF PITCH AND AMPLITUDE PERTURBATION

A. Methods

1. Stimuli

The purpose of experiment 3 was to determine whether subjects could, in fact, differentiate between the effects of pitch and amplitude perturbation. The test stimuli consisted of nine signals each from the correlated pitch-perturbation continuum, the uncorrelated pitch-perturbation continuum, the correlated amplitude-perturbation continuum, and the uncorrelated amplitude-perturbation continuum. The nine stimulus values from each continuum were chosen in such a way that the spacing between stimuli was approximately even in perceptual terms, as determined by the roughness magnitude estimates. Each series of nine stimuli began with a stimulus having a perturbation value of zero. These four stimuli should have been identical and were included as a reliability check. The 36 stimuli were equated for overall rms intensity and presented over a loudspeaker using the procedures described previously.

2. Subjects and procedures

The ten subjects who had participated in experiments 1 and 2 served as listeners. Two separate identification tasks were run in counterbalanced order. One task used the correlated signals from each continuum, and the other used the uncorrelated signals. A very brief training session preceded each identification task. The training session consisted of two randomly ordered presentations of each of the 18 stimuli. Subjects were asked to press one of two keys on a terminal keyboard to indicate whether the stimulus was drawn from the jitter continuum or the shimmer continuum. Feedback was provided on each of the 36 trials. The testing sessions were identical except that feedback was not provided, and each stimulus was presented ten times in pseudorandom order.

B. Results and discussion

The results are presented in Fig. 6, which shows percent correct identification for each stimulus. With the obvious exception of stimuli with zero-perturbation values, subjects were generally able to determine whether the stimulus represented perturbations in pitch or amplitude. Identification performance improved at higher perturbation levels and was generally better for the correlated rather than uncorrelated stimuli. These results suggest that, contrary to the findings reported by Wendahl (1966b) for sawtooth waves, the subjective qualities produced by jitter and shimmer in synthetic vowels are quite different, except at very low levels of aperiodicity.

IV. EXPERIMENT 4: PERCEPTION OF ADDITIVE NOISE

The purpose of experiment 4 was to study the relationship between additive noise and perceived dysphonia and to determine how this relationship might be affected by the slope of the spectrum in the periodic component of the stimulus. Examination of the role spectral slope was motivated by Yanagihara's (1967) finding that the loss of energy in high-frequency harmonics results in an increase in perceived dysphonia even when stimuli are matched for signal-to-noise ratio.

A. Methods

1. Stimuli

All stimuli for experiment 4 were synthesized with VSYN, which controls signal-to-noise ratio by the appropriate scaling and point-for-point addition of separate periodic and aperiodic components. A single aperiodic signal was generated with the Klatt (1980) synthesis program by passing the aspiration source through formant resonators that were set appropriate for [a] ($F1 = 720$, $F2 = 1240$, $F3 = 2400$, $F4 = 3300$, $F5 = 3700$). Except for differences in amplitude, the noise waveform was identical for all stimuli.

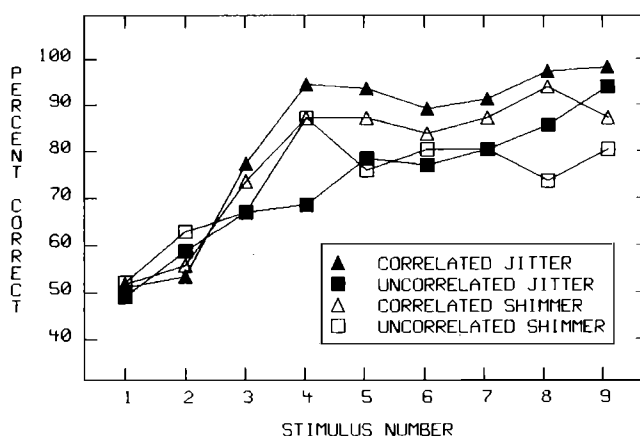


FIG. 6. Percent correct identification of stimuli from the correlated and uncorrelated jitter continua and from the correlated and uncorrelated shimmer continua. The subjects' task was to judge whether the stimulus was drawn from one of the jitter continua or one of the shimmer continua.

Several versions of the periodic component were generated using the Klatt synthesizer. Two methods were used to control spectral slope. In method 1, spectral slope was controlled at the glottal level by varying the "bandwidth of glottal resonance" (BGR) parameter in the Klatt synthesizer. This parameter controls the cutoff frequency of a low-pass filter that is used to shape the voice impulse train. The BGR parameter was set at 75, 150, and 300 Hz, producing the glottal source functions shown in Fig. 7. The glottal waveforms were passed through formant resonators appropriate for [a] and then mixed with the scaled noise described above. Three 13-step continua were synthesized that varied in signal-to-noise ratio in 3-dB steps from -10 to 26 dB (39 stimuli). All stimuli were 1.0 s in duration and were synthesized with a constant 130-Hz F_0 and a 40-kHz sample frequency.

Method 2, which was more nearly analogous to Yanagihara's (1967) technique, used the same glottal waveform for all stimuli and controlled the spectral slope by adjusting for-

mant amplitudes. With the synthesizer set in parallel mode, the resonator gains associated with F_1 - F_6 were spaced at -3-, -5-, or -10-dB increments. For example, for the -10-dB signal, F_1 gain was set to 66 dB, F_2 gain was set to 56 dB, F_3 gain was set to 46 dB, etc. The spectral characteristics of these stimuli are shown in Fig. 8. The periodic components that were generated with this method were mixed appropriately with the noise signal described above to produce three additional 13-step continua varying in signal-to-noise ratio from -10 to 26 dB.

2. Subjects and procedures

Listeners consisted of nine of the ten speech pathologists who participated in the other experiments. An additional subject meeting the same criteria was recruited to replace one speech pathologist who was not available at the time the experiment was run. Using the magnitude estimation task described above, subjects were asked to rate the stimuli on the degree of perceived breathiness. Subjects were run in two blocks of 624 trials in counterbalanced order. One block consisted of 16 pseudorandomly ordered presentations of the 39 stimuli created using method 1 to control spectral slope; a second block consisted of 624 presentations of the 39 stimuli created using method 2. For both blocks of trials, the first 117 stimulus presentations were considered to be practice trials and were not included in the data analysis.

B. Results and discussion

Functions relating signal-to-noise ratios to normalized breathiness ratings are shown in Fig. 9 for method 1 and Fig. 10 for method 2. The smooth curves are third-order polynomials. Not surprisingly, there is a very strong relationship between signal-to-noise ratio and listeners' perception of breathiness. However, contrary to the results reported by Yanagihara (1967), the amount of high-frequency energy in the periodic component did not appear to play a role in controlling the degree of perceived dysphonia. In general, stimuli with similar signal-to-noise ratios tended to receive very

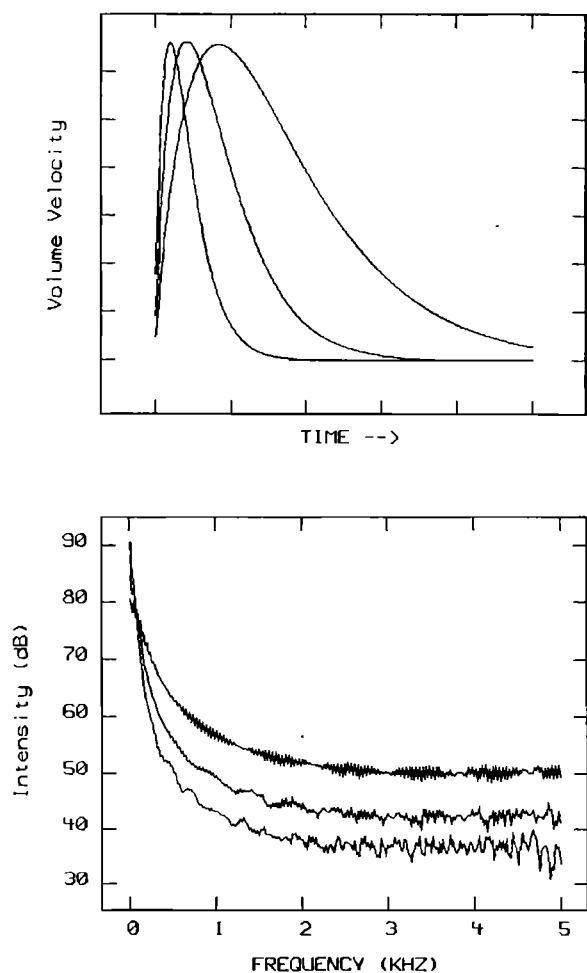


FIG. 7. Time-domain (top) and frequency-domain (bottom) representations of glottal source functions varying in spectral slope. The function showing the highest rate of change in the time domain and the greatest amount of high-frequency energy was produced with a BGR value of 300 Hz; the function with the most gradual rate of change and the least amount of high-frequency energy was produced with a BGR of 75 Hz. The middle function in both panels was produced with a BGR of 150 Hz.

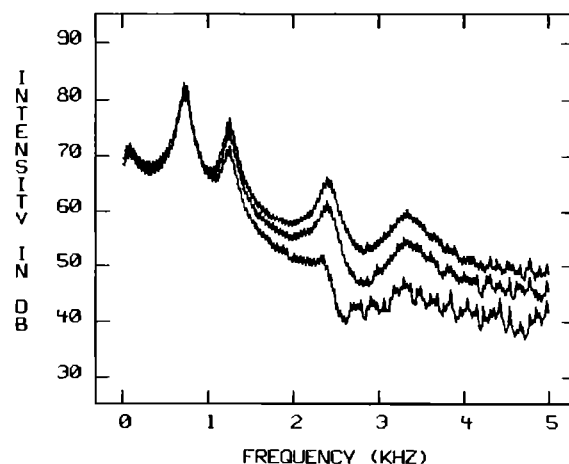


FIG. 8. Fourier spectra (1024 points) of the periodic components produced by controlling formant amplitudes with the synthesizer in parallel mode.

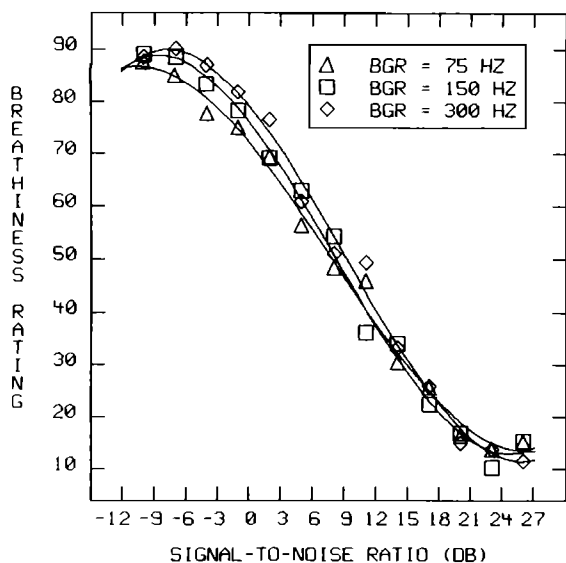


FIG. 9. Perceived breathiness as a function of signal-to-noise ratio. The parameter is the spectral slope of the periodic component, which was controlled by varying the cutoff frequency of a low-pass filter that shapes the glottal source function. This is the BGR parameter in the Klatt (1980) synthesis program.

similar breathiness ratings, regardless of differences in the spectral slope of the periodic component.

It is not clear why the present findings do not agree with those of Yanagihara (1967). Comparing the synthesis procedures used in the two studies is difficult since the methods used by Yanagihara are not described in great detail. One possibility that was considered is that the discrepancy is related to the fact that Yanagihara's subjects rated the stimuli on hoarseness, while subjects in the present study were asked to make breathiness judgments. To test this possibility, the experiment was rerun with four additional speech patholo-

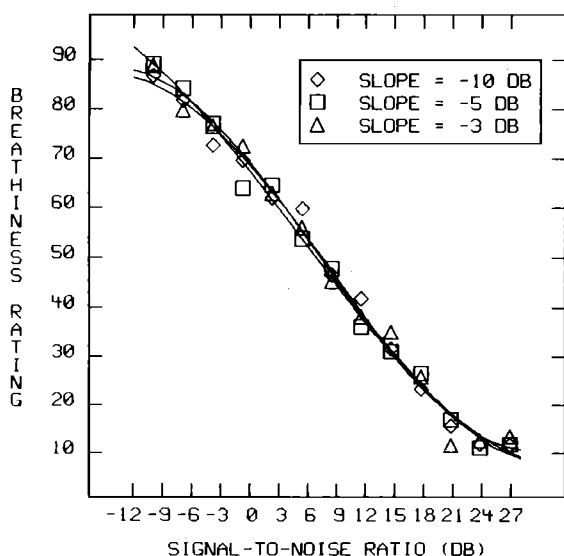


FIG. 10. Perceived breathiness as a function of signal-to-noise ratio. The parameter is the spectral slope of the periodic component, which was controlled by varying parallel formant amplitudes. Parallel resonator gains for F1-F6 were spaced at intervals of -3, -5, or -10 dB.

gists who were asked to rate the same set of stimuli on hoarseness. The results of that test were virtually identical to the data shown in Figs. 9 and 10.

V. GENERAL DISCUSSION

To summarize briefly, four experiments were run that examined univariate relationships between perceived dysphonia and variation in pitch perturbation, amplitude perturbation, and additive noise in synthetically generated sustained vowels. Among the results were: (1) Strong relationships were found between perceived roughness and variation in either pitch or amplitude perturbation; (2) stimuli that were generated from correlated pitch or amplitude sequences sounded more rough than stimuli generated from uncorrelated sequences, especially for stimuli varying in amplitude perturbation; (3) unlike findings reported for sawtooth waves varying in pitch and amplitude perturbation (Wendahl, 1966b), the percept associated with pitch perturbation was noticeably different from that associated with amplitude perturbation; (4) a strong relationship was found between additive noise and ratings of breathiness; and (5) contrary to Yanagihara's (1967) report, ratings of either breathiness or hoarseness were unaffected by the spectral slope of the periodic component.

Results of the listening tests using stimuli varying in pitch and amplitude perturbation are in general agreement with those reported for nonspeech waveforms (Heiberger and Horii, 1982; Wendahl, 1963, 1966a,b). One very important discrepancy concerns Wendahl's (1966b) report that the roughness percepts associated with pitch and amplitude perturbation were very similar to one another. Results from the present study suggest that these two percepts are quite easy to differentiate. Further, informal interviews with the speech pathologists who served as listeners suggested that jitter seems to more closely approximate the kind of roughness that is heard in naturally occurring disordered voices. Most listeners agreed that stimuli toward the high end of the shimmer continua had an unnatural quality. It might be noted that the unusual quality associated with the shimmer stimuli does not seem to be restricted to stimuli created with the time-domain synthesis method used by VSYN. Pilot work using stimuli generated with the Klatt (1980) synthesis program produced stimuli that sounded very similar to those generated by VSYN.

The unnatural quality of the stimuli varying in amplitude perturbation was not expected since previous correlational research with naturally produced voices suggested that amplitude perturbation was more strongly associated with perceived roughness than pitch perturbation (Deal and Emanuel, 1978; Nichols, 1979). It is important to note, however, that the amplitude-perturbation measures reported in the natural speech studies almost certainly reflected several different types of aperiodicity. Because of limitations in measurement techniques, measured values of parameters such as amplitude perturbation can reflect unknown combinations of amplitude perturbation, pitch perturbation, additive noise, and perhaps other sources of aperiodicity (Hillenbrand, 1987). The presence of these measurement artifacts makes it difficult to interpret perceptual results in terms of

specific underlying acoustic events.

Another possibility that should be considered is that there is some important aspect of amplitude perturbation in naturally produced voices that was not modeled accurately by the synthesis techniques used in the present study. It is also possible that the amplitude-perturbation signals sound unnatural because amplitude variability is occurring against a background of perfect periodicity in other dimensions. More natural-sounding signals might be produced if amplitude perturbation were combined with other types of aperiodicity. Experiments that are just underway in our laboratory are designed to study the perceptual properties of sustained vowels that combine several sources of aperiodicity.

The experiments with stimuli varying in pitch and amplitude perturbation also showed that perceived roughness was affected not only by the amount of perturbation, but also by the degree of correlation among adjacent pitch or amplitude values. In general, stimuli that were generated from correlated sequences tended to sound significantly more rough than stimuli that were generated from uncorrelated sequences. This was especially true for the amplitude-perturbation stimuli, where the difference in perceived roughness between correlated and uncorrelated signals was very large. For the pitch-perturbation signals, there was some evidence that this effect may have been at least partly due to the choice of mean jitter as a way to calculate pitch perturbation. The difference between correlated and uncorrelated signals was reduced significantly when stimuli were equated for either PPQ (Davis, 1981) or the standard deviation of signed jitter. However, changing the calculation method did not entirely eliminate the difference in perceived roughness between correlated and uncorrelated pitch-perturbation signals, and this effect seemed to be largely unrelated to calculation methods for the amplitude-perturbation signals.

One implication of these findings is that acoustic measurement techniques might need to account for the sequential characteristics of pitch or amplitude change in addition to the degree of perturbation. Work along these lines has been reported by Koike (1973), who studied autocorrelation functions of voice amplitude sequences from normal speakers and patients with either laryngeal neoplasms or unilateral vocal cord paralysis. Koike reported that the presence peaks in the autocorrelation function at lags of 3–12 periods could be useful in differentiating patients with neoplasms from the other two groups. The present findings suggest that the sequential characteristics of pitch and amplitude change might need to be incorporated in the development of a quantitative index of the severity of dysphonia. More research will be needed, however, to determine how the sequential properties of pitch and amplitude change in naturally produced voices can be quantified and how this information can be combined with more standard measures of perturbation.

The primary finding from the listening tests using stimuli varying in additive noise was the failure to observe an effect for the spectral slope of the periodic component. In general, stimuli with similar signal-to-noise ratios tended to receive very similar ratings of either breathiness or hoarseness. It is important to note that these findings do not indi-

cate that spectral slope plays no role in judgments of voice quality. Although it was not tested formally, most listeners indicated that they were aware of the differences in "brightness" among the stimuli. These differences, however, did not appear to influence subjects' judgments along the two quality dimensions that were tested.

As indicated previously, the long-term goal of this work is to learn something about *multivariate* rather than simply univariate relationships between perceived dysphonia and variation in underlying acoustic parameters. Experiments that are currently underway using more complex synthetic stimuli have been designed to determine how acoustic parameters such as the ones examined in the present study combine to influence listener judgments of the overall severity of dysphonia, as well as the type of dysphonia. Another important issue that will need to be addressed in future research concerns the generalizability of results based on sustained vowels to continuous speech. Sustained vowels have been studied heavily because the measurement problems are more tractable and because the psychophysical characteristics are likely to be much simpler. However, there are a number of phenomena found only in continuous speech (e.g., pitch breaks and aperiodicities at transitions between voiced and unvoiced segments) that are likely to play a strong role in the perception of dysphonia. A major challenge for future research will be to determine how these dynamic characteristics interact with the kinds of aperiodicities that were examined in the present study.

ACKNOWLEDGMENTS

I am very grateful to Dale Metz and Robert Whitehead of the National Technical Institute for the Deaf for their technical advice, comments on previous drafts, and the generous contribution of their time in listening to the test signals and suggesting improvements to the procedures. I would like to thank Raymond Colton for his helpful comments on a previous draft, Thomas Ridley for his help with data analysis, data collection, and software development, and William Martens and Martin Wilde for their help in developing techniques for the generation of correlated random numbers. This research was supported by NIH Grant No. 7-R01-NS-23703-01 to RIT Research Corporation.

- Aronson, A. (1980). *Clinical Voice Disorders: An Interdisciplinary Approach* (Thieme-Stratton, New York).
- Coleman, R. F. (1969). "Effects of median frequency levels upon the roughness jittered stimuli," *J. Speech Hear. Res.* **12**, 330–336.
- Davis, S. B. (1976). "Computer evaluation of laryngeal pathology based on inverse filtering of speech," *SCRL Monogr.* **13**, Speech Communication Research Laboratory, Santa Barbara, CA.
- Davis, S. B. (1981). "Acoustical characteristics of normal and pathological voices," *ASHA Rep.* **11**, 97–115.
- Deal, R. E., and Emanuel, F. W. (1978). "Some waveform and spectral features of vowel roughness," *J. Speech Hear. Res.* **21**, 250–264.
- Emanuel, F. W., and Sansone, F. (1969). "Some spectral features of 'normal' and 'simulated rough' vowels," *Folia Phoniatr.* **21**, 410–415.
- Fritzell, B., Hammarberg, B., and Wedin, L. (1977). "Clinical applications of acoustic voice analysis, Part I: Background and perceptual factors," *Speech Trans. Lab. Q. Prog. Stat. Rep.* **2-3**, 31–38.
- Froekjaer-Jensen, B., and Prytz, S. (1976). "Registration of voice quality," *Bruel and Kjaer Tech. Rev.* **3**, 3–17.
- Gardner, M. (1978). "White and brown music, fractal curves, and one-over-*f* fluctuations," *Sci. Am.* **238**, 16–31.

- Gauffin, J., and Sundberg, J. (1977). "Clinical applications of acoustic voice analysis, Part II: Acoustic analysis, results, and discussion," *Speech Tran. Lab. Q. Prog. Stat. Rep.* 2-3, 39-43.
- Gill, J. S. (1961). "Automatic extraction of the excitation function of speech with particular reference to the use of correlation methods," in *Proceedings of the Third International Congress on Acoustics* (Elsevier, Amsterdam), Vol. 1, pp. 217-220.
- Hammarberg, B., Fritzell, B., Gauffin, J., Sundberg, J., and Wedin, L. (1980). "Perceptual and acoustic correlates of abnormal voice quality," *Acta Otolaryngol.* 90, 441-451.
- Hecker, M., and Kruel, E. J. (1971). "Descriptions of the speech of patients with cancer of the vocal folds, Part I: Measures of fundamental frequency," *J. Acoust. Soc. Am.* 49, 1275-1282.
- Heiberger, V. L., and Horii, Y. (1982). "Jitter and shimmer in sustained phonation," in *Speech and Language: Advances in Basic Research and Practice, Vol. 7*, edited by N. J. Lass (Academic, New York), pp. 299-332.
- Hillenbrand, J. (1987). "A methodological study of perturbation and additive noise in synthetically generated voice signals," *J. Speech Hear. Res.* 30, 448-461.
- Hirano, M. (1981). *Clinical Examination of Voice* (Springer, New York).
- Hollien, H., Michel, J., and Doherty, E. T. (1973). "A method for analyzing vocal jitter in sustained phonation," *J. Phon.* 1, 85-91.
- Holmes, J. N. (1962). "The effect of simulating natural larynx behavior on the quality of synthetic speech," *Speech Communications Seminar, Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden.*
- Horii, Y. (1979). "Fundamental frequency perturbation observed in sustained phonation," *J. Speech Hear. Res.* 22, 5-19.
- Horii, Y. (1980). "Vocal shimmer in sustained phonation," *J. Speech Hear. Res.* 23, 202-209.
- Horii, Y. (1982). "Jitter and shimmer differences among sustained vowel phonations," *J. Speech Hear. Res.* 25, 12-14.
- Jacob, L. (1968). "A normative study of laryngeal jitter," Master's thesis, University of Kansas, Lawrence, KA, unpublished.
- Jensen, J. P. (1965). "Adequacy of terminology for clinical judgment of voice quality deviation," *Eye, Ear, Nose Throat Mon.* 44, 77-82.
- Kempster, G. B. (1984). "A multidimensional analysis of dysphonia in two dysphonic groups," Ph.D. thesis, Northwestern University, Evanston, IL, unpublished.
- Kempster, G. B., and Kistler, D. J. (1984). "Perceptual dimensions of dysphonic voices," *J. Acoust. Soc. Am. Suppl.* 1 75, S8.
- Kitajima, K., and Gould, W. J. (1976). "Vocal shimmer in sustained phonation of normal and pathologic voice," *Ann. Otol. Rhinol. Laryngol.* 85, 377-381.
- Kitajima, K., Tanabe, M., and Isshiki, N. (1975). "Pitch perturbation in normal and pathologic voice," *Stud. Phonol.* 9, 25-32.
- Klatt, D. H. (1980). "Software for a cascade/parallel formant synthesizer," *J. Acoust. Soc. Am.* 67, 971-995.
- Koike, Y. (1973). "Application of some acoustic measures for the evaluation of laryngeal dysfunction," *Stud. Phonol.* 7, 17-23.
- Kojima, H., Gould, W. J., Lambaise, A., and Isshiki, N. (1980). "Computer analysis of hoarseness," *Acta Otolaryngol.* 89, 547-554.
- Lieberman, P. (1963). "Some acoustic measures of the fundamental periodicity of normal and pathologic larynges," *J. Acoust. Soc. Am.* 35, 344-353.
- Lively, M. A., and Emanuel, F. W. (1970). "Spectral noise levels and roughness severity ratings for normal and simulated rough vowels produced by adult females," *J. Speech Hear. Res.* 13, 503-517.
- Mandelbrot, B. (1983). *The Fractal Geometry of Nature* (Freeman, San Francisco).
- Murry, T., and Doherty, E. T. (1980). "Selected acoustic characteristics of pathologic and normal speakers," *J. Speech Hear. Res.* 23, 361-369.
- Murry, T., Singh, S., and Sargent, M. (1977). "Multidimensional classification of abnormal voice qualities," *J. Acoust. Soc. Am.* 61, 1630-1635.
- Nichols, A. C. (1979). "Jitter and shimmer related to vocal roughness: A comment on the Deal and Emanuel study," *J. Speech Hear. Res.* 22, 670-671.
- Prosek, R. A., Montgomery, A. A., Walden, B. E., and Hawkins, D. B. (1984). "Some relations between voice-quality judgments and derived acoustic measurements," *J. Acoust. Soc. Am. Suppl.* 1 75, S8.
- Rozsypal, A. J., and Millar, B. F. (1979). "Perception of jitter and shimmer in synthetic vowels," *J. Phon.* 7, 343-355.
- Robbins, J. (1981). "A comparative acoustic study of laryngeal speech, esophageal speech, and speech production after tracheo-esophageal puncture," Ph.D thesis, Northwestern University, Evanston, IL, unpublished.
- Sansone, F., and Emanuel, F. W. (1970). "Spectral noise levels and roughness severity ratings for normal and simulated rough vowels produced by adult males," *J. Speech Hear. Res.* 13, 489-502.
- Schroeder, M. R. (1961). "Recent progress in speech coding at Bell Telephone Laboratories," in *Proceedings of the Third International Congress on Acoustics* (Elsevier, Amsterdam), Vol. 1, pp. 201-210.
- Simon, C. (1927). "The variability of consecutive wavelengths in vocal and instrumental sounds," *Psychol. Monogr.* 36, 41-83.
- Smith, B., Weinberg, B., Feth, L., and Horii, Y. (1978). "Vocal jitter and roughness characteristics of esophageal speech," *J. Speech Hear. Res.* 21, 240-249.
- Takahashi, H., and Koike, Y. (1975). "Some perceptual dimensions and acoustical correlates of pathologic voices," *Acta Otolaryngol. Suppl.* 338, 1-24.
- Wendahl, R. (1963). "Laryngeal analog synthesis of harsh vocal quality," *Folia Phoniatr.* 15, 241-250.
- Wendahl, R. (1966a). "Some parameters of auditory roughness," *Folia Phoniatr.* 18, 26-32.
- Wendahl, R. (1966b). "Laryngeal analog synthesis of jitter and shimmer: auditory parameters of harshness," *Folia Phoniatr.* 18, 98-108.
- Wilde, M. D., and Martens, W. L. (1985). "VSYN: A computer program for synthesizing vocal signals varying in perturbation and signal-to-noise ratio," Northwestern University, Evanston, IL.
- Wilde, M. D., Martens, W. L., Hillenbrand, J., and Jones, D. R. (1986). "Externalization mediates changes in the perceived roughness of sound signals with jittered fundamental frequencies," in *Proceedings of the 1986 International Computer Music Conference*, The Hague.
- Yanagihara, N. (1967). "Significance of harmonic change and noise components in hoarseness," *J. Speech Hear. Res.* 10, 531-541.
- Yumoto, E., Gould, W. J., and Baer, T. (1982). "Harmonics-to-noise ratio as an index of the degree of hoarseness," *J. Acoust. Soc. Am.* 71, 1544-1550.
- Yumoto, E., Sasaki, Y., and Okamura, H. (1984). "Harmonics-to-noise ratio and psychophysical measurement of the degree of hoarseness," *J. Speech Hear. Res.* 27, 2-6.