A pitch-synchronous analysis of hoarseness in running speech

Hiroshi Muta and Thomas Baer
Haskins Laboratories, 270 Crown Street, New Haven, Connecticut 06511

Kikuju Wagatsuma and Teruo Muraoka
Central R&D Center, Research and Development Division, Victor Company of Japan, Ltd., 58-7 Shinmei-cho, Yokosuka, Kanagawa, 239 Japan

Hiroyuki Fukuda
Department of Otolaryngology, Keio University School of Medicine, 35 Shinanomachi, Shinjuku-ku, Tokyo, 160 Japan

(Received 7 December 1987; accepted for publication 13 June 1988)

A method of pitch-synchronous acoustic analysis of hoarseness requiring a voice sample of only four fundamental periods is presented. This method calculates a noise-to-signal (N/S) ratio, which indicates the depth of valleys between harmonic peaks in the power spectrum. The spectrum is calculated pitch synchronously from a Fourier transform of the signal, windowed through a continuously variable Hanning window spanning exactly four fundamental periods. A two-stage procedure is used to determine the exact duration of the four fundamental periods. An initial estimate is obtained using autocorrelation in the time domain. A more precise estimate is obtained in the frequency domain by minimizing the errors between the preliminary calculated power spectrum and the predicted spectrum spread of a windowed harmonic signal. Analysis of synthesized voices showed that the N/S ratio is sensitive to additive noise, jitter, and shimmer, and is insensitive to slow (8 Hz) modulation in fundamental frequency and amplitude. An analysis of pre- and postoperative voices of six patients with benign laryngeal disease showed that the N/S ratio for vowel /u/ in running speech consistently improved after surgery for all subjects, in agreement with their successful therapeutic results.

PACS numbers: 43.70.Jt, 43.70.Dn

INTRODUCTION

A degradation in voice quality, generally called hoarseness, is one of the major symptoms of such benign laryngeal diseases as vocal cord polyps or nodules, and is often the first symptom of neoplastic diseases such as laryngeal cancer, as well. Quantitative measures of the acoustic characteristics associated with laryngeal pathology have focused on two different kinds of parameters, which are compatible with the source-filter model of voice production (Isshiki *et al.*, 1966): the parameters defined by cycle-to-cycle variation of the glottal source signal, and those defined for individual cycles of the source signal, such as the signal-to-noise ratio and the relative intensity of higher harmonics.

Measures of the glottal source periodicity in a sustained vowel, such as cycle-to-cycle perturbation of pitch period (Lieberman, 1961) and amplitude (Koike, 1969), have provided objective indicants of the degree of hoarseness either directly from the audio signal or from the glottal source signal calculated by the inverse filtering (Davis, 1976). However, while these measures may change in advanced laryngeal cancer, they do not always show significant glottal source perturbation in a hoarse voice associated with a benign disease or an early cancer (Ludlow *et al.*, 1987).

Sound spectrographic analysis of sustained vowels shows less conspicuous harmonic structure in hoarse voices than in normal voices (Yanagihara, 1967). This phenomenon, low intensity of the harmonic component relative to the background, has been explained either as a decrease of higher harmonics in the source spectrum (Isshiki *et al.*, 1966), or as an increase of additive noise in the source signal (Kasuya *et al.*, 1986). The modulation effect of cycle-to-cycle perturbation of the glottal source may also contribute to the apparent decay of harmonic structure.

Several methods for quantitative documentation of the spectrographic phenomenon have been reported, using calculations either in the frequency domain (Kojima *et al.*, 1980; Kitajima, 1981; Hiraoka *et al.*, 1984; Kasuya *et al.*, 1986) or in the time domain (Yumoto *et al.*, 1982). All of them showed differences between normal and pathological subjects, as well as correlations with subjective ratings of hoarseness severity. However, such methods require a long sustained vowel for analysis, and thus are sensitive to fluctuations of pitch, intensity, or articulation, as well as intentional vibrato. Any of these factors would contribute to an apparent reduction of the harmonic structure of the voice. Reliability of these methods thus depends on the subjects' ability to produce a long sustained vowel at constant pitch and intensity.

An additional problem with previous methods for quantifying harmonic content and spectral noise is their limited ability to resolve individual glottal cycles for analysis. A fractional error in fundamental period extraction or in pitch synchronization causes additional spectrum leakage of the

 0001-4966/88/101292-10$00.80

original harmonics, causing further deterioration of the harmonic structure. Possibly as a result of all these problems, previous quantification methods have yet to demonstrate their clinical usefulness in the evaluation of mild to moderate hoarseness, such as objective measurement of the therapeutic effects of phonosurgery.

We have developed a method of pitch-synchronous analysis that requires a very short voice sample, consisting of only four fundamental periods. The four-cycle sample can be extracted not only from sustained vowels, but also from vowels in running speech. A precise pitch-synchronous spectrum is calculated from a Fourier transform of the windowed signal, through a continuously variable Hanning window spanning exactly four fundamental periods. A two-stage procedure is used to determine the exact duration of the four fundamental periods: one in the time domain, and one in the frequency domain. A noise-to-signal (N/S) ratio, which indicates the depth of valleys between harmonic peaks in the power spectrum, is calculated from the pitch-synchronous spectrum. This acoustic analysis will be useful in assessing mild or moderate hoarseness, because the examinees do not have the difficult task of producing a constant long sustained vowel for analysis.

I. ANALYSIS PROCESS

A. Pitch extraction

1. Estimation of the fundamental period in the time domain

The continuous-time waveform of the speech signal is denoted by $s(t)$. Then, the discrete-time sequence $s^*(n)$ is given by

$$s^*(n) = s(n\Delta t), \tag{1}$$

where Δt is the sampling period.

A rough estimate of the fundamental period $K_0\Delta t$ is first obtained, using any available method. The size for the four fundamental periods M is defined by

$$M = 4K_0. \tag{2}$$

The Hanning window function for this analysis frame is defined as

$$w(t) = 0.5(1 - \cos 2\pi t/T) \quad (0 \leqslant t \leqslant T), \tag{3}$$

where $T = M\Delta t$. The continuous-time waveform of the windowed speech signal $s_w(t)$ is defined by

$$s_w(t) = w(t)s(t) \quad (0 \leqslant t \leqslant T). \tag{4}$$

The discrete autocorrelation function $R(n)$ for this frame is defined as

$$R(n) = \sum_{i=0}^{M-n-1} s_w^*(i)s_w^*(i+n), \tag{5}$$

where $s_w^*(n)$ is the discrete-time sequence of $s_w(t)$.

The fundamental period size K is obtained from the function $R(K)$. If K is not equal to K_0, K_0 is set to K, and steps (2) to (5) are repeated until the frame size M consists of four fundamental periods. The fundamental frequency f_0 is given by

$$f_0 = 1/K\Delta t. \tag{6}$$

2. Calculation of the precise fundamental frequency in the frequency domain

The amplitude spectrum $|X(k)|$ is derived by computing the discrete Fourier transform $X(k)$ of the windowed signal:

$$X(k) = \sum_{n=0}^{M-1} s_w^*(n)e^{-j2\pi kn/M}. \tag{7}$$

The analysis frame consists of four fundamental periods, so there is one harmonic peak of $|X(k)|$ for every four steps of k.

Hanning windowing causes the line spectrum of a harmonic signal to spread. If there is a small error in the estimated fundamental frequency, this spread will not be centered around the harmonic peaks of $X(k)$. We define a function $F_h(f,x)$ that describes the spectrum spread of the h th harmonic, as a function of the error in fundamental frequency x, given the measured amplitude of the h th harmonic $|X(4h)|$.

$$F_h(f,x) = [|X(4h)|/|W(-hx)|]W[f - h \times (f_0 + x)], \tag{8}$$

where $W(f)$ is the Fourier transform of the window function $w(t)$:

$$\begin{aligned} W(f) &= \int_0^T w(t)e^{-j2\pi ft}\,dt \\ &= 0.5T\left[\frac{\sin \pi fT}{\pi fT} + 0.5\left(\frac{\sin \pi(fT-1)}{\pi(fT-1)} + \frac{\sin \pi(fT+1)}{\pi(fT+1)}\right)\right]e^{-j\pi fT}. \end{aligned} \tag{9}$$

A better estimate of the fundamental frequency is obtained by searching for the value of x for which the difference between $|F_h(f,x)|^2$ and the measured power spectrum $|X(k)|^2$ on both sides of each harmonic peak is minimized. The estimation errors for the lower and higher spectrum spread of the h th harmonic $E_{Lh}(x)$ and $E_{Hh}(x)$ are defined as

$$\begin{aligned} E_{Lh}(x) &= |X(4h-1)|^2 - |F_h(hf_0 - 1/T,x)|^2 \\ &= |X(4h-1)^2 - \frac{|X(4h)|^2}{|W(hx)|^2}\left|W\left(hx + \frac{1}{T}\right)\right|^2, \end{aligned} \tag{10}$$

$$\begin{aligned} E_{Hh}(x) &= |X(4h+1)|^2 - |F_h(hf_0 + 1/T,x)|^2 \\ &= |X(4h+1)|^2 - \frac{|X(4h)|^2}{|W(hx)|^2}\left|W\left(hx - \frac{1}{T}\right)\right|^2. \end{aligned} \tag{11}$$

The total square error $G(x)$ from the first to the L th harmonic is

$$G(x) = \sum_{h=1}^{L} E_{Lh}^2(x) + \sum_{h=1}^{L} E_{Hh}^2(x). \tag{12}$$

In this study, the square errors are calculated up to the 16th harmonic peak, which is lower than the Nyquist frequency for all subjects.

The minimum of $G(x)$ is found from its derivative $G'(x)$:

$$G'(x) = 0. \tag{13}$$

This equation is solved using Newton's method, starting with an initial guess of $x = 0$. Thus the precise fundamental frequency f_R is given by

$$f_R = f_0 + x. \tag{14}$$

B. Pitch-synchronous spectrum analysis

The Hanning window is redefined in order to cover four pitch cycles more precisely according to the new estimate of the fundamental frequency f_R. The window size T_R is defined as

$$T_R = 4/f_R. \tag{15}$$

The Hanning window function is defined as

$$w_R(t) = \begin{cases} 0.5(1 - \cos 2\pi t/T_R), & 0 \leqslant t \leqslant T_R, \\ 0, & \text{otherwise.} \end{cases} \tag{16}$$

The continuous-time waveform of the windowed speech signal $s_R(t)$ is defined by

$$s_R(t) = w_R(t)s(t) \quad (-\infty \leqslant t \leqslant \infty), \tag{17}$$

and the corresponding discrete-time sequence $s_R^*(n)$ is, therefore,

$$s_R^*(n) = \begin{cases} w_R(n\Delta t)\ s^*(n), & n = 0,1,2,...,M_R, \\ 0, & \text{all other } n, \end{cases} \tag{18}$$

where M_R is the largest integer that is smaller than $T_R/\Delta t$.

The continuous spectrum of a continuous-time signal is obtained from the Fourier transform of its discrete-time sequence provided that the signal is bandlimited within the Nyquist frequency. As long as the original signal is sufficiently bandlimited, the windowed signal is bandlimited to a good approximation. Therefore, the Fourier transform $X_R(f)$ of $s_R^*(n)$ is given by

$$X_R(f) = \sum_{n=-\infty}^{\infty} s_R^*(n)e^{-j2\pi fn\Delta t}. \tag{19}$$

The pitch-synchronous power spectrum of the windowed signal $P(k)$, which is evaluated at frequency steps of $1/T_R$, is thus calculated as

$$P(k) = |X_R(k/T_R)|^2 = \left| \sum_{n=0}^{M_R} s_R^*(n)e^{-j2\pi kn\Delta t/T_R} \right|^2. \tag{20}$$

C. Calculation of noise-to-signal ratio

Because the Hanning window covers exactly four fundamental periods, harmonic peaks and valleys appear in every four steps of k. If the signal consists of pure harmonics, the h th mainlobe consists of $P(4h-1)$, $P(4h)$, and $P(4h+1)$, and no sidelobes appear in the valley $P(4h+2)$. The shallower the valley, the higher the level of the nonharmonic components.

The smallest value of the signal power $P(k)$ over the h th harmonic peak and valley $4h - 1 \leqslant k \leqslant 4h + 2$ is taken as the power of the noise component for the h th harmonic peak P_{Nh}. Therefore, the estimated power spectrum of the noise component $P_N(k)$ is defined as

$$P_N(k) = \min_{i=-1,0,1,2} P(4h+i) = P_{Nh} \quad (4h-1 \leqslant k \leqslant 4h+2), \tag{21}$$

where $h = 1,2,3,...,L$. In this study, these spectra are calculated up to the 16th harmonic.

The noise-to-signal ratio R_{NS} is defined as

$$R_{NS} = 10 \log\left(\sum_{k=3}^{4L+2} P_N(k) \Big/ \sum_{k=3}^{4L+2} P(k) \right). \tag{22}$$

II. METHOD OF THIS STUDY

A. Analysis of synthesized voices

In order to study the sensitivity of the N/S ratio, voices synthesized by the SPEAK program (Titze, 1986) were analyzed by the present method. The source model was noninteractive with the vocal tract, and a parametrized model of the glottal flow waveform was used. Voice samples were created with varying amounts of jitter, shimmer, additive noise, amplitude modulation, and frequency modulation; the vowel /u/ was used for synthesis. Samples were synthesized at a rate of 20 000 samples per second with 6 dB/oct preemphasis.

B. Analysis of pre- and postoperative voices

Table I describes the subjects used in the present study. Three males and three females with mild or moderate hoarseness due to benign laryngeal disease were selected for study. All subjects underwent microscopic laryngeal surgery and had sufficiently good perceptual voice quality after surgery so that both surgeons and patients were satisfied with

TABLE I. Subjects for analysis.

Subject	Name	Age	Sex	Diagnosis	Perceptual result $N = 34$ (%)	H/N ratio (dB)	
						Pre	Post
1	NO	39	M	polyp	94.1	14.4	10.6
2	KI	46	M	polyp	79.4	18.2	18.2
3	FI	29	M	polyp	100	2.3	14.8
4	KI	35	F	cyst	100	19.0	21.4
5	NK	30	F	nodules	67.6	16.0	11.2
6	MU	46	F	polyp	100	19.9	17.5

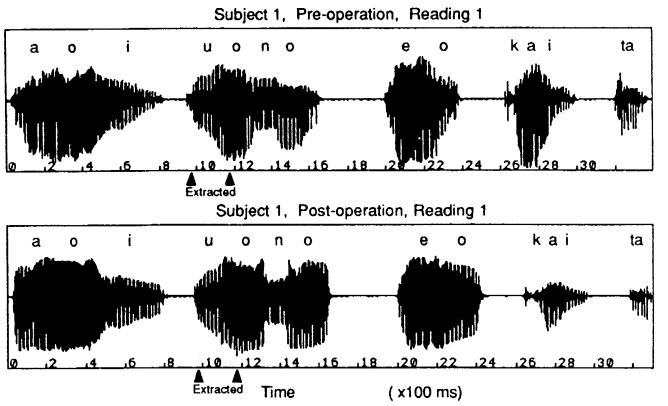

FIG. 1. Waveforms of the sentence /aoi uo no e o kaita/ for the first preoperative reading (top) and the first postoperative reading (bottom) by subject 1.

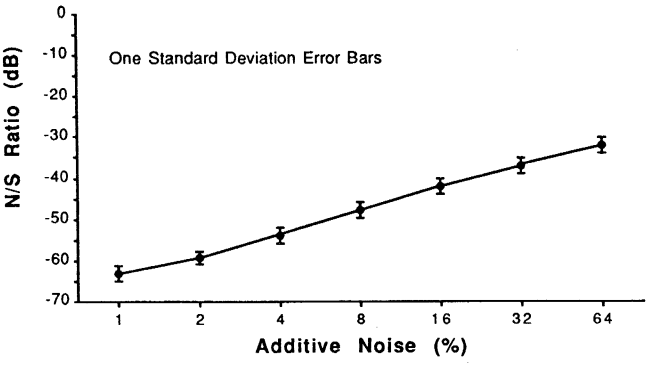

FIG. 3. The N/S ratio for synthesized voices, vowel /u/, $F0 = 220$ Hz, with varying amounts of additive noise in the glottal source.

the results. Pre- and postoperative samples of the six speakers were presented to 34 listeners, in paired comparison format. The listeners correctly selected the postoperative sample at the levels indicated in Table I. The levels are above chance ($p < 0.03$) for each speaker. However, the calculated pre- and postoperative values of the harmonics-to-noise (H/N) ratio for sustained vowel /a/ (Yumoto *et al.*, 1982) fall within the normal range of 7.4 dB or greater in all cases except the preoperative value for subject 3. These results suggest that most of the preoperative samples may be considered to be mild or moderate hoarseness, although voice quality was definitely improved after surgery for all subjects.

The subjects were requested to read the Japanese sentence /aoi uo no e o kaita/ ("I drew a picture of a blue fish"). The sentence was read twice in a session, and recordings were made both preoperatively and postoperatively, 3–8 weeks after the surgery. Recording was made using a high-fidelity electret condenser microphone (Sony ECM-23F) and a cassette tape recorder (Sony TC-2890SE) in a lightly sound-treated booth at Keio University Hospital. The noise level in the booth was approximately 45 dB (SPL). Figure 1 shows the waveform for the preoperative and postoperative utterances of subject 1. The sentence was read rather slowly and distinctly, as can be seen in the figure.

The recorded voice was digitized with 12-bit precision at a sampling rate of 10 000 samples per second without preemphasis. The cutoff frequency for the antialiasing low-pass filter was 4.8 kHz. Voice samples of 200-ms duration, which covered the vowel /u/ in /uo no/, were extracted for the analysis. We chose this vowel because the phrase /uo no/ has a flat accent pattern and is located in the middle of the sentence. The extracted regions are indicated by arrows in Fig. 1.

III. RESULTS OF ANALYSIS

A. Results of synthesized voice analysis

Results of the synthesized voice analysis demonstrate the sensitivity of the N/S ratio. Figure 2 shows the waveform and the power spectrum for an analysis frame of a synthe-

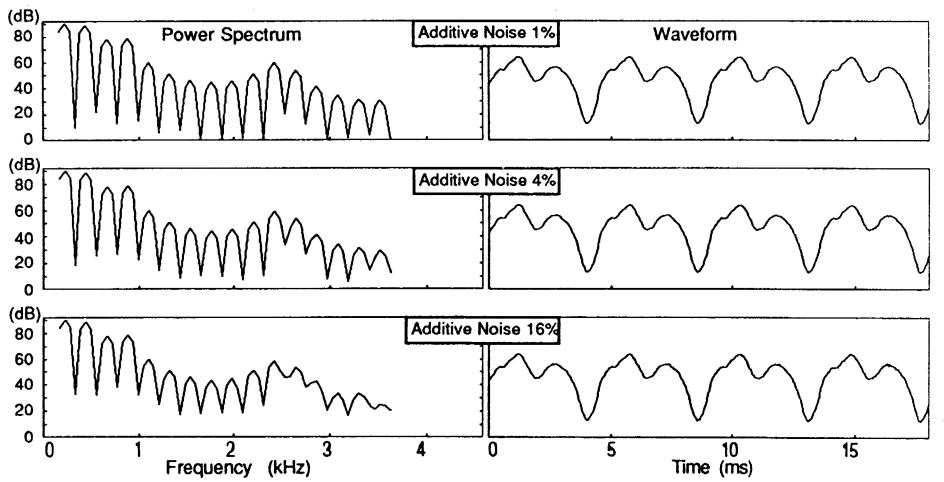

FIG. 2. Waveforms and power spectra for an analysis frame of the synthesized voices, vowel /u/, $F0 = 220$ Hz, with 1%, 4%, and 16% additive noise in the glottal source.

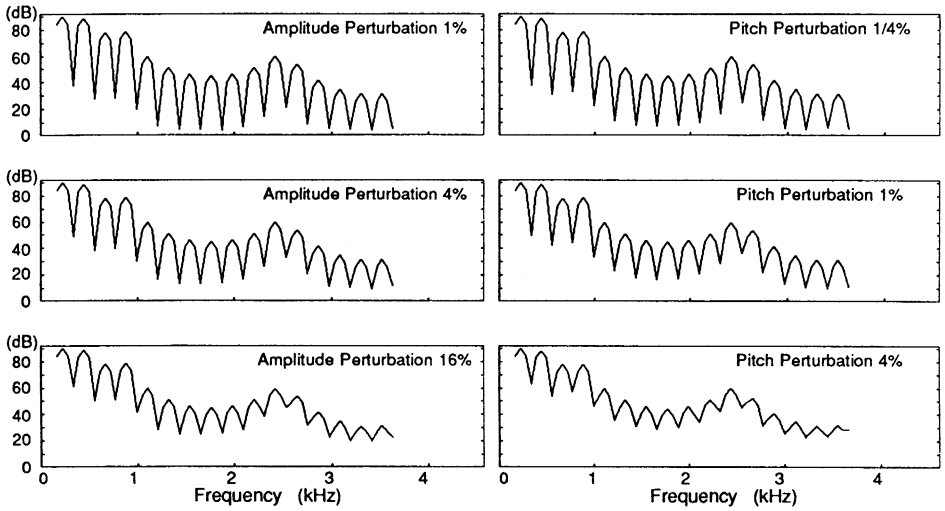

FIG. 4. Averaged power spectra for the synthesized voices, vowel /u/, $F0 = 220$ Hz, with 1%, 4%, and 16% amplitude perturbation and $\frac{1}{4}$%, 1%, and 4% pitch perturbation of the glottal source.

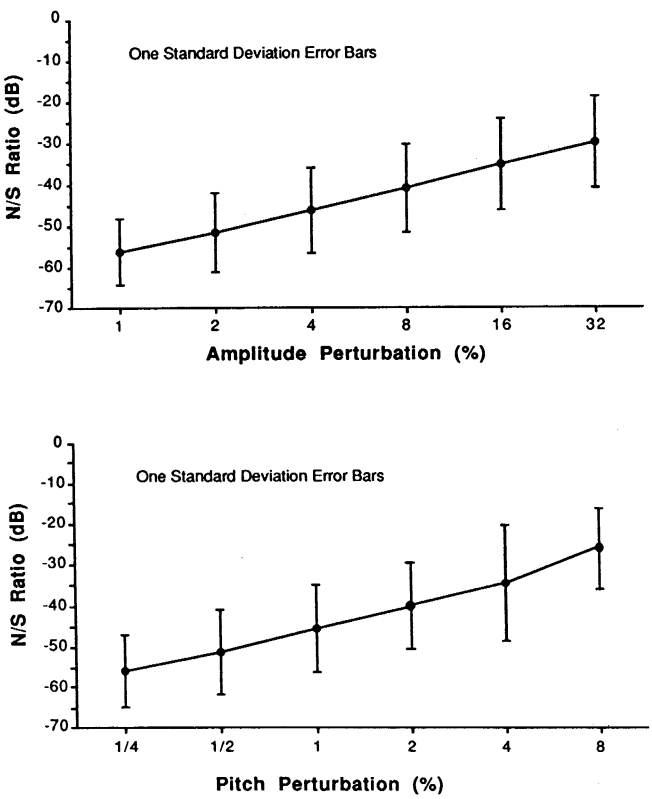

FIG. 5. The N/S ratio for synthesized voices, vowel /u/, $F0 = 220$ Hz, with varying amounts of amplitude perturbation (top) and pitch perturbation (bottom) of the glottal source.

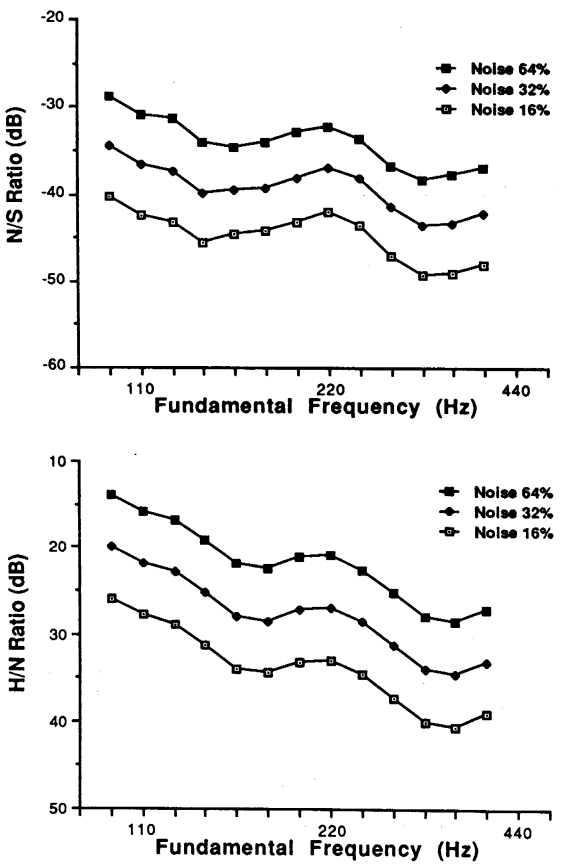

FIG. 6. The N/S ratio (top) and the H/N ratio (bottom) for the synthesized voices of varying fundamental frequency with 16%, 32%, and 64% additive noise.

sized voice, vowel /u/, $F0 = 220$ Hz, with 1%, 4%, and 16% additive noise in the glottal source. As expected, the greater the noise, the shallower the valleys in the power spectrum. Figure 3 shows the N/S ratio for synthesized voices with varying amounts of additive noise. Each result is the average of 25 frames, shifted 6.4 ms each. Standard deviations are indicated by error bars. The N/S ratio varies with the amount of additive noise in the glottal source signal. The same result was obtained from voice samples with $F0 = 110$ Hz.

Figure 4 shows the averaged power spectrum of 25 frames, shifted 6.4 ms each, for synthesized voices, vowel /u/, $F0 = 220$ Hz, with 1%, 4%, and 16% amplitude perturbation and $\frac{1}{4}$%, 1%, and 4% pitch perturbation of the glottal source. Again, the greater the perturbation, the shallower the valleys in the power spectrum. Figure 5 shows the N/S ratio for the synthesized voices with varying amounts of amplitude perturbation and pitch perturbation. The N/S ratio varies with the amount of amplitude or pitch perturbation of the glottal source, and again the same result was obtained from the voice samples with $F0 = 110$ Hz. It may be noted that the N/S ratios for pitch and amplitude perturbation show greater variance than those for additive noise. This appears to be a statistical artifact. A synthesized voice with source perturbation contains only one random factor for each glottal cycle, while there is a random component in each sample for the additive noise case.

Figure 6 shows the N/S ratio and the H/N ratio (Yumoto *et al.*, 1982) for synthesized voices of varying fundamental frequency with 16%, 32%, and 64% additive noise. The fundamental frequency was varied from 98–392 Hz at six logarithmic steps per octave. Both indices showed the same pattern of fluctuation, which appeared to be an artifact created by the synthesized program. While both N/S ratio and H/N ratio were fairly insensitive to fundamental frequency over the normal speech range, the N/S ratio was somewhat less sensitive.

Figure 7 shows time domain results for modulated synthesized voices with 16% additive noise. The glottal source was modulated at 8 Hz with 32% sinusoidal amplitude modulation or with 4% sinusoidal frequency modulation. One hundred frames with 1.6-ms frame shift were analyzed for each of the two conditions. The top panels indicate the voice waveform. Upper markings show the center of each frame. The middle panels show the fundamental frequency for each frame. The bottom panels show the N/S ratio smoothed by a moving average of three successive frames.

Figure 7 shows that the N/S ratio varies as a result of glottal source modulation. In order to extract the most stable parts of the modulated signals, three successive frames, whose averaged N/S ratio showed the minimum value, were taken as the representatives for these samples. These three

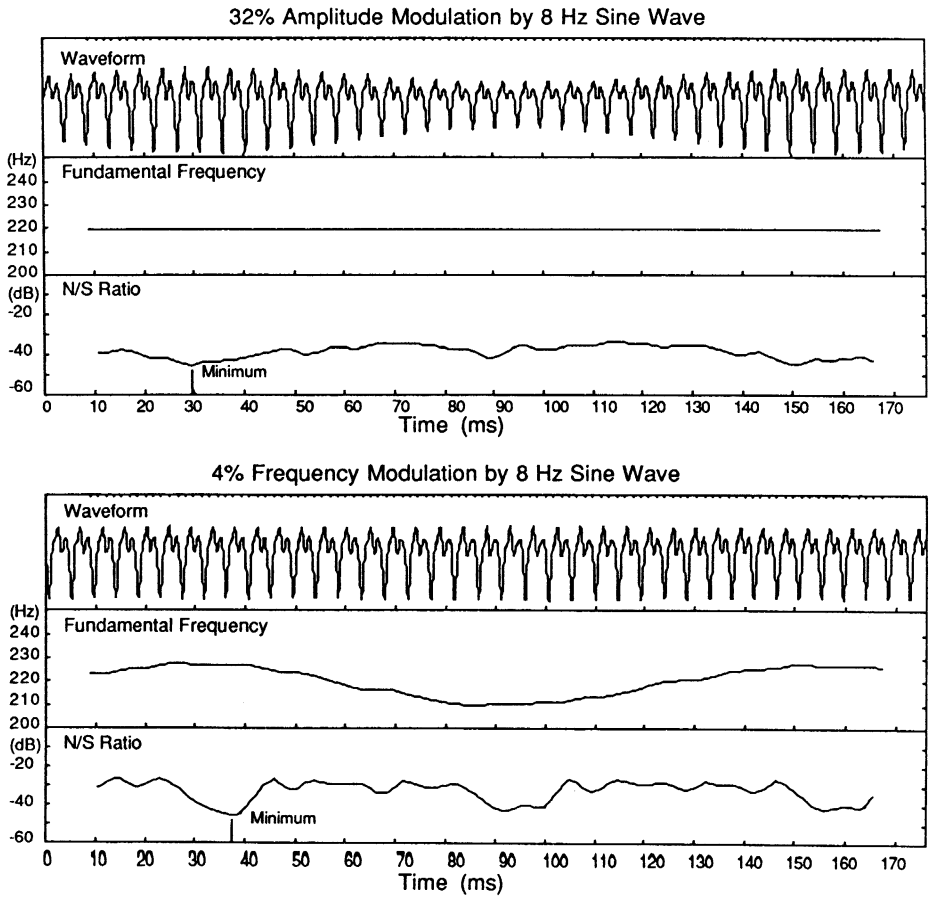

FIG. 7. Time domain results for the modulated synthesized voices, vowel /u/, $F0 = 220$ Hz, with 16% additive noise. The glottal source was modulated at 8 Hz with 32% sinusoidal amplitude modulation (top) or with 4% sinusoidal frequency modulation (bottom).

frames, whose center for each of the two conditions is indicated by the vertical bar in each bottom panel, predict the N/S ratio for this noise level without modulation. Figure 8 shows the waveforms and power spectra for the selected three frames from the modulated samples with 16% additive noise. These spectra show similar harmonic structure to those for the nonmodulated voice with the same amount of additive noise shown in Fig. 2.

Figures 9 and 10 show the N/S ratio for modulated synthesized voices with 16%, 32%, and 64% additive noise, with varying amounts of 8-Hz glottal source modulation either in amplitude (Fig. 9) or in frequency (Fig. 10). Each data point is an average of three successive frames whose N/S ratio showed the minimum value. The N/S ratio is insensitive to glottal source modulation (within 1 s.d. of the nonmodulated samples) up to 32% amplitude modulation or up to 4% frequency modulation for samples $F0 = 220$ Hz and up to 16% amplitude modulation or up to 2% frequency modulation for samples $F0 = 110$ Hz. The relatively small frame size, 18.2 ms for $F0 = 220$ Hz, compared to the period of source modulation, 125 ms for 8 Hz, is the reason for the insensitivity of the N/S ratio.

B. Results of patient voice analysis

Figure 11 shows the time domain results for the pre- and postoperative voice samples of subject 1. The N/S ratio varied during the speech sample. Three successive frames, whose averaged N/S ratio showed the minimum value, were taken as the representatives for each sample. Figure 12 shows the waveforms and power spectra for the selected three frames from the pre- and postoperative samples of this subject. The postoperative spectrum shows better harmonic structure than the preoperative spectrum.

Table II shows the analysis results for the N/S ratio and fundamental frequency for the six subjects before and after laryngeal surgery. Each result is an average of three successive frames, whose N/S ratio showed the minimum value. Figure 13 shows the averaged N/S ratio of each pair (first and second readings) of pre- and postoperative voice samples. The N/S ratio consistently improved after the surgery in all six subjects. Thus results of therapy considered to be successful by doctor and patient were indicated by the analysis.

IV. DISCUSSION

Voice quality is difficult to assess objectively. Various laryngeal diseases may cause a pathological change in voice quality, and each abnormal voice may give a different perceptual impression to different listeners. We need better understanding of the perception of voice quality as well as better understanding of pathological production in order to properly evaluate the acoustic characteristics of a deviant voice in relation to both the perceptual impression of listeners and to the pathological state of the larynx.

Classifications of listeners' impressions in multiple dimensions, such as rough, breathy, asthenic, and strained, have been proposed (Hirano, 1981), and acoustic parameters associated with different kinds of voice quality have been studied (Imaizumi, 1986a,b). For example, "roughness" may be associated with modulations over several pitch periods or, at low pitch, with factors that are the same across cycles. "Breathy" voice may be characterized by additive noise or by weakness of harmonics above the fundamental. The relative strength of harmonics also contributes to the perceptual contrast between "asthenic" and "strained" voices.

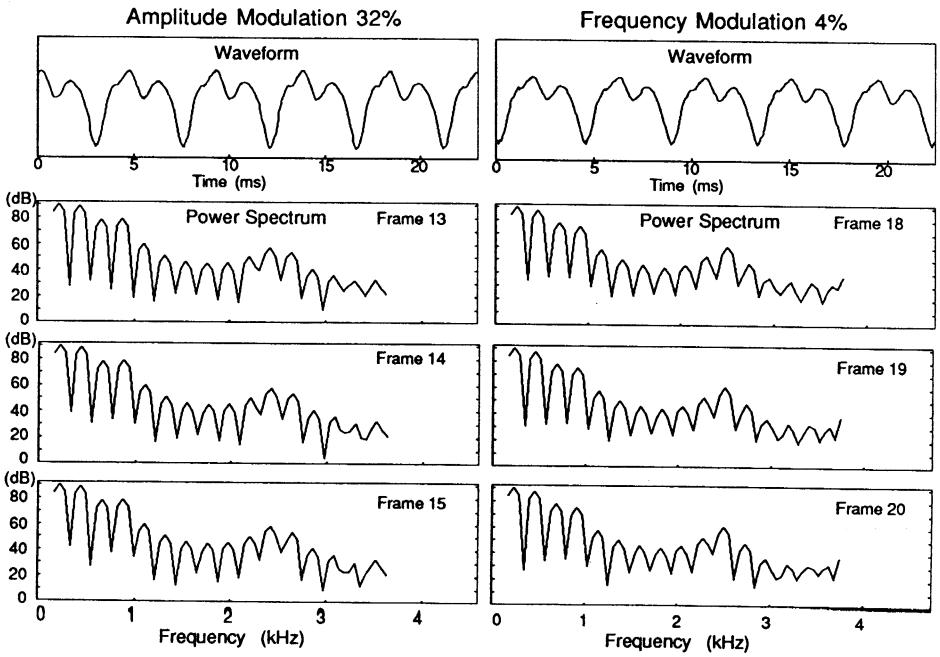

FIG. 8. Waveforms and power spectra for the three frames, with minimum N/S ratio, from the modulated synthesized voices, vowel /u/, $F0 = 220$ Hz, with 16% additive noise. The glottal source was modulated at 8 Hz with 32% sinusoidal amplitude modulation (left) or with 4% sinusoidal frequency modulation (right).

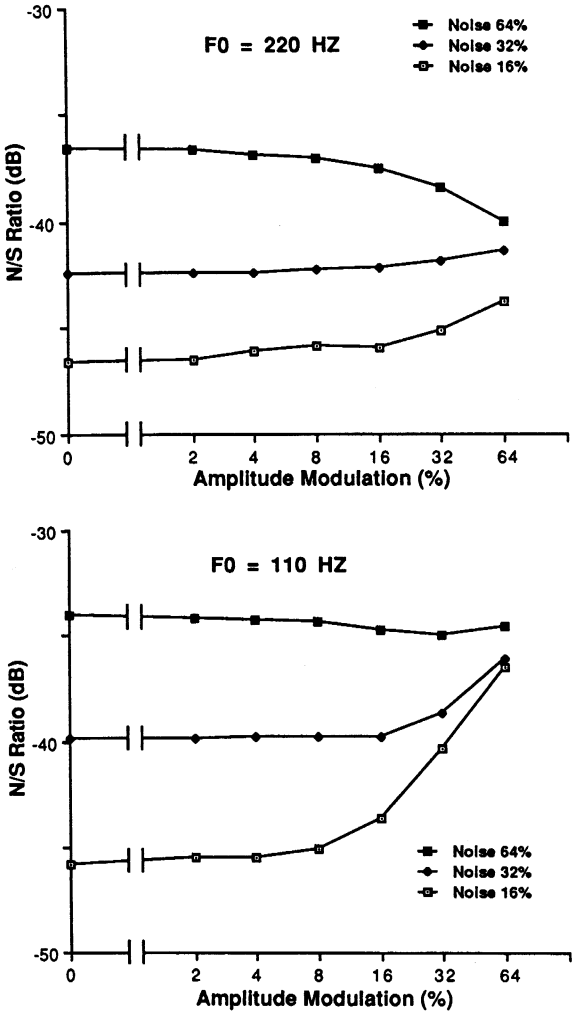

FIG. 9. The N/S ratio for the modulated synthesized voices, vowel /u/, $F0 = 220$ Hz (top) and 110 Hz (bottom), with 16%, 32%, and 64% additive noise, whose glottal source contained varying amounts of 8-Hz sinusoidal modulation in amplitude. Each data point is an average of the three frames, whose N/S ratio showed the minimum value.

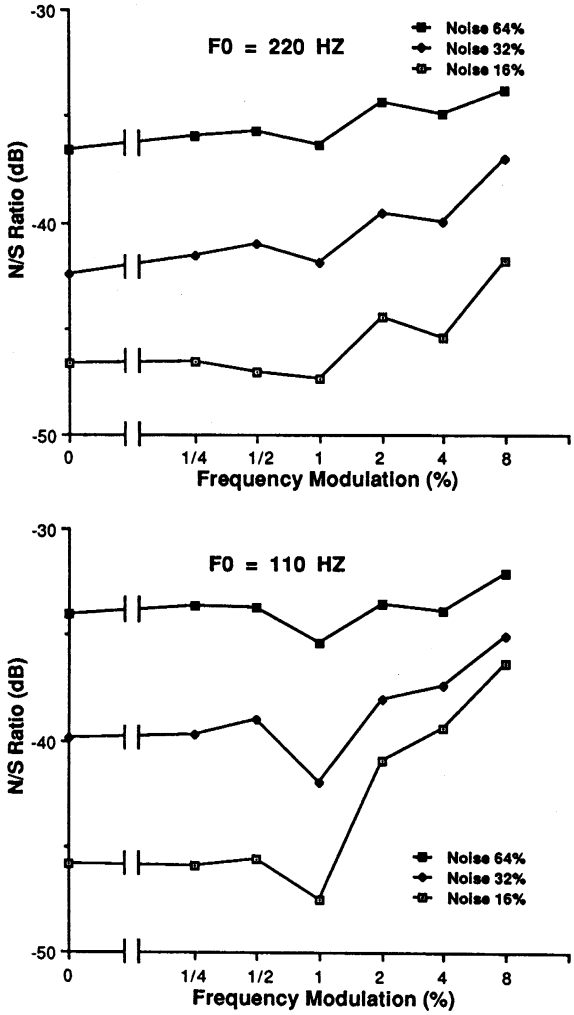

FIG. 10. The N/S ratio for modulated synthesized voices, vowel /u/, $F0 = 220$ Hz (top) and 110 Hz (bottom), with 16%, 32%, and 64% additive noise, whose glottal source contained varying amounts of 8-Hz sinusoidal modulation in frequency. Each data point is an average of three frames, whose N/S ratio showed the minimum value.

The above kinds of acoustic parameters do not bear a simple relationship to pathological modes of vocal-fold vibrations and, in addition, they interact with each other. For example, glottal source perturbations distort the harmonic structure and thus affect both noise measures and harmonic strength measures. Similarly, additive noise may contribute to acoustic measures of source perturbation. To properly evaluate each acoustic characteristic separately, it is necessary to accurately extract individual glottal cycles from the acoustic signal and to separate the glottal excitation signal from the nonspecific spectral noise in each cycle.

Inverse filtering has been proposed as a method for extracting source characteristics from the acoustic signal (Davis, 1976). However, it is doubtful whether inverse filtering provides sufficiently accurate results, especially with abnormal voices. For example, in a study applying the LPC method to hoarse voices, measured formant patterns appeared to be affected by cycle-to-cycle variations in source characteristics (Muta *et al.*, 1987).

If we are to understand fully the acoustic characteristics of hoarse voice, we will have to learn much more about the relationship between the pathological vibrations of the vocal folds and the resulting acoustic signal. In the meantime, we have adopted a simple assumption for the present analysis based on sound-spectrographic finding (Yanagihara, 1967): For whatever reason, a hoarse voice has a greater nonharmonic component and a less pure harmonic component than a normal voice.

Periodic structure in the voice signal is the prerequisite for pitch-synchronous spectrum analysis. Therefore, the present method can be applied only to a case of mild or moderate hoarseness. In such cases, the fundamental period can

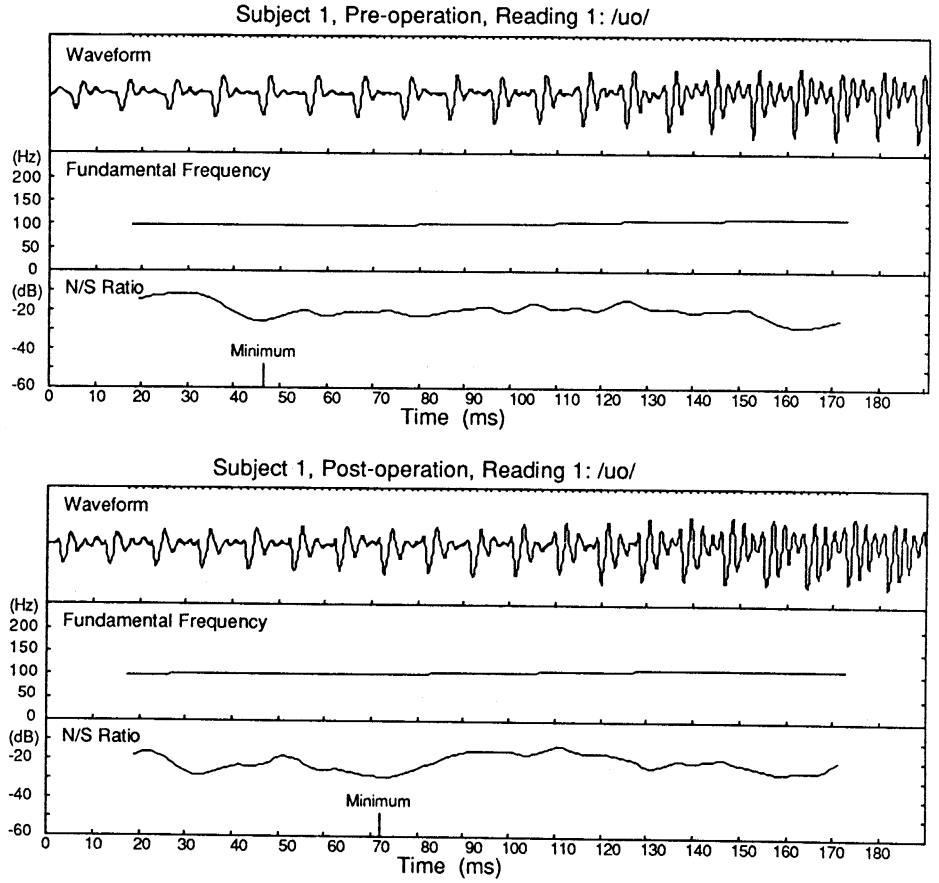

FIG. 11. Time domain results for the first preoperative reading (top) and the first postoperative reading (bottom) by subject 1. The top panels indicate the waveforms of the voice for the vowel /u/ in /uo no/. One hundred frames with 1.6-ms frame shift were analyzed for each of the two conditions. Upper markings show the center of each frame. The middle panels show the fundamental frequency for each frame. The bottom panels show the N/S ratio smoothed by the moving average of three successive frames. The vertical bar in each bottom panel, which shows the minimum of the smoothed N/S ratio, indicates the most stable part of the vowel /u/.

be estimated easily by measures of the acoustic waveform without additional instrumental observations of vocal-fold vibration, such as laryngeal stroboscopy or electroglottography.

The N/S ratio was calculated over the spectral region between the 1st and 16th harmonics. Generally, the harmonic structure of a voice signal shows greater distortion in higher harmonics than in lower harmonics. The modulation effect of source perturbation increases in proportion to the order of harmonics. Therefore, the higher the harmonic, the greater the measured noise-to-signal ratio. However, voice signals were not preemphasized and we analyzed the vowel

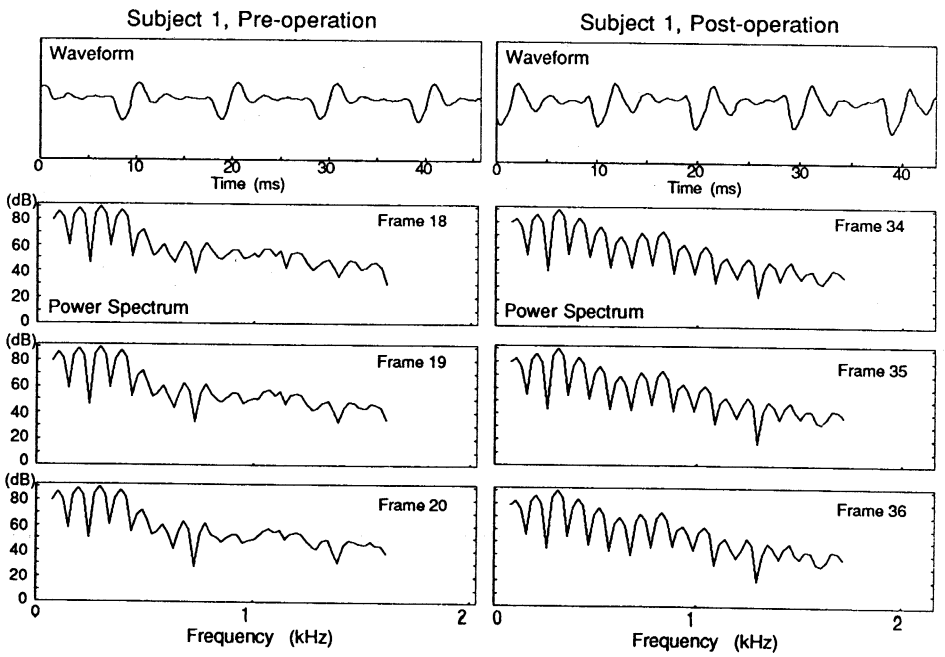

FIG. 12. Waveforms and power spectra for the selected three frames, which showed the minimum N/S ratio, from the first preoperative reading (left) and the first postoperative reading (right) by subject 1.

TABLE II. Analysis results of the N/S ratio and the fundamental frequency for six subjects before and after laryngeal surgery.

	Preoperation				Postoperation			
	Reading 1		Reading 2		Reading 1		Reading 2	
Subject	*F*0(Hz)	N/S(dB)	*F*0(Hz)	N/S(dB)	*F*0(Hz)	N/S(dB)	*F*0(Hz)	N/S(dB)
1	97.5	− 25.8	97.1	− 24.8	103.8	− 29.0	97.2	− 33.4
2	136.0	− 20.1	135.6	− 29.6	140.5	− 34.3	136.9	− 31.4
3	144.5	− 29.8	135.9	− 25.9	131.3	− 32.4	131.6	− 35.9
4	234.5	− 34.9	233.3	− 36.0	248.6	− 40.4	252.5	− 42.0
5	202.4	− 29.0	212.7	− 30.8	237.1	− 42.5	228.3	− 41.4
6	195.9	− 34.5	200.6	− 26.0	209.3	− 36.7	219.9	− 36.9

/u/, whose first and second formant frequencies are among the lowest of the five Japanese vowels. The vowel spectra were thus dominated by low frequencies, so the analysis covered most of the acoustic power of the voice. The calculated N/S ratio would, therefore, be expected to be insensitive to the particular choice of analysis parameter, such as sampling rate and number of harmonics analyzed, over a fairly wide range.

However, spectral differences between source signals, such as an increase or decrease of higher harmonics, may affect the N/S ratio. The pathological characteristics of the source spectrum, such as weakness of higher harmonics in breathy voices, may be evaluated from the present pitch-synchronous spectrum, if we can assume that the effect of the vocal tract resonance was the same for the given voice samples.

In summary, we have developed a pitch-synchronous method for analyzing pathological voice quality. The method is sensitive to additive noise, jitter, and shimmer, and is insensitive to slower modulations in amplitude and fundamental frequency. The results of analyzing pre- and postoperative running speech, which indicate successful therapy of six patients with laryngeal disease, indicate the clinical usefulness of this method.

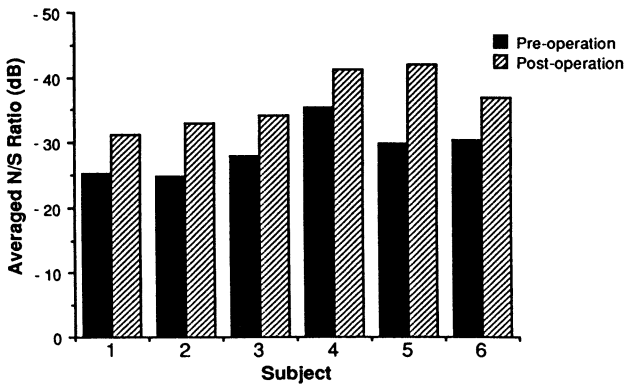

FIG. 13. Averaged N/S ratio of each pair (first and second readings) of pre- and postoperative voice samples.

ACKNOWLEDGMENT

This work was supported by NINCDS Grant 13870 to Haskins Laboratories.

Davis, S. B. (**1976**). "Computer evaluation of laryngeal pathology based on inverse filtering of speech," SCRL Monograph 13.

Hirano, M. (**1981**). *Clinical Examination of Voice* (Springer, Wien), pp. 81–84.

Hiraoka, N., Kitazoe, Y., Ueta, H., Tanaka, S., and Tanabe, M. (**1984**). "Harmonic-intensity analysis of normal and hoarse voices," J. Acoust. Soc. Am. **76**, 1648–1651.

Imaizumi, S. (**1986a**). "Acoustic measure of roughness in pathological voice," J. Phon. **14**, 457–462.

Imaizumi, S. (**1986b**). "Clinical application of the acoustic measurement of pathological voice qualities," Ann. Bull. Res. Inst. Logoped. Phoniatr. Univ. Tokyo **20**, 211–216.

Isshiki, N., Yanagihara, N., and Morimoto, M. (**1966**). "Approach to the objective diagnosis of hoarseness," Folia Phoniatr. **18**, 393–400.

Kasuya, H., Ogawa, S., Mashima, K., and Ebihara, S. (**1986**). "Normalized noise energy as an acoustic measure to evaluate pathologic voice," J. Acoust. Soc. Am. **80**, 1329–1334.

Kitajima, K. (**1981**). "Quantitative evaluation of the noise level in the pathologic voice," Folia Phoniatr. **33**, 115–124.

Koike, Y. (**1969**). "Vowel amplitude modulations in patients with laryngeal diseases," J. Acoust. Soc. Am. **45**, 839–844.

Kojima, H., Gould, W. J., Lambiase, A., and Isshiki, N. (**1980**). "Computer analysis of hoarseness," Acta Otolaryngol. **89**, 547–554.

Lieberman, P. (**1961**). "Perturbations in vocal pitch," J. Acoust. Soc. Am. **33**, 597–603.

Ludlow, C. L., Bassich, C. J., Connor, N. P., Coulter, D. C., and Lee, Y. J. (**1987**). "The validity of using phonatory jitter and shimmer to detect laryngeal pathology," in *Laryngeal Function in Phonation and Respiration*, edited by T. Baer, C. Sasaki, and K. Harris (Little, Brown, Boston), pp. 492–508.

Muta, H., Muraoka, T., Wagatsuma, K., Horiuchi, M., Fukuda, F., Takayama, E., Fujioka, T., and Kanou, S. (**1987**). "Analysis of hoarse voices using the LPC method," in *Laryngeal Function in Phonation and Respiration*, edited by T. Baer, C. Sasaki, and K. Harris (Little, Brown, Boston), pp. 463–474.

Titze, I. R. (**1986**). "Three models of phonation," J. Acoust. Soc. Am. Suppl. 1 **79**, S81.

Yanagihara, N. (**1967**). "Significance of harmonic changes 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.