Ann.Bull.RILP(1988)No.22, 85-96

A PITCH-SYNCHRONOUS ANALYSIS OF HOARSENESS IN RUNNING SPEECH

Hiroshi Muta*, Thomas Baer*, Hiroyuki Fukuda**,
and Shigeji Saito**

Introduction

A degraded change in voice quality, generally called hoarseness, is one of the major symptoms of benign laryngeal diseases such as vocal cord polyps. Quantitative studies of the acoustic characteristics associated with laryngeal pathology have focused on two different kinds of measures[1]; the periodicity of the glottal source and the signal-to-noise ratio of the source signal.

Description of the glottal source periodicity in a sustained vowel, such as measures of cycle-to-cycle perturbation of pitch period[2] and amplitude[3], has objectively indicated the degree of hoarseness either directly from the audio signal or from the glottal source signal calculated by inverse filtering[4]. However, these measures do not always show significant glottal source perturbation in a hoarse voice associated with a benign disease or an early cancer[5].

Sound spectrographic analysis of sustained vowels shows less conspicuous harmonic structure in hoarse voices than in normal voices[6]. This phenomenon, low intensity of the harmonic component relative to the background, has been explained both as a decrease of higher harmonics in the source spectrum[1], and as an increase of additive noise in the source signal[7]. The modulation effect of cycle-to-cycle perturbation of the glottal source may 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[7,8,9,10] or in the time domain[11]. 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. Reliability of these methods thus depends on the subjects' ability to produce a long sustained vowel at constant pitch and intensity.

We have developed a method of pitch-synchronous analysis, which 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. This method calculates a noise-to-signal (N/S) ratio from the power spectrum, which indicates the depth of valleys

* Haskins Laboratories
** Keio University

between harmonic peaks. A precise pitch-synchronous spectrum is calculated from a discrete 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.

Analysis process

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. The size for the four fundamental periods, M, is temporarily set according to the preliminarily estimated fundamental period, $K_0\Delta t$:

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

The Hanning window function for this analysis frame is defined as

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

where $T = M\ \Delta t$. The windowed speech signal, $s_W(t)$, is defined by

$$s_W(t) = w(t)\ s(t), \qquad (\ 0 \le t \le T\ ). \tag{4}$$

The auto-correlation 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 peak, $R(K)$. If K is not equal to K_0, then set $K_0 = K$, and repeat (2) to (5) until the frame size, M, consists of four fundamental periods. The fundamental frequency, f_0, is given by

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

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, $F_h(f,x)$, which describes the spectrum spread of the hth harmonic, as a function of the error in fundamental frequency, x, given the measured amplitude of the hth harmonic, $|X(4h)|$.

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

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

$$W(f) = \int_0^T w(t)\ e^{-j2\pi ft}\, dt$$

$$= 0.5T \left[\frac{\sin\pi fT}{\pi fT} + 0.5\left\{ \frac{\sin\pi(fT-1)}{\pi(fT-1)} + \frac{\sin\pi(fT+1)}{\pi(fT+1)} \right\} \right] e^{-j\pi fT}. \tag{9}$$

A better estimate of the fundamental frequency is obtained by searching for the value of x for which the difference between $|F_h(f)|^2$ 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 hth harmonic, $E_{Lh}(x)$ and $E_{Hh}(x)$, are defined as

$$E_{Lh}(x) = |X(4h-1)|^2 - |F_h(hf_0-1/T,x)|^2 \tag{10}$$

$$E_{Hh}(x) = |X(4h+1)|^2 - |F_h(hf_0+1/T,x)|^2 \tag{11}$$

The total square error, $G(x)$, from the first to the Lth 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 errors are calculated up to the 16th harmonic, 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 $f_R = f_0 + x$.

The Hanning window is re-defined 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

$$T_R = 4 / f_R. \tag{14}$$

The Hanning window function is defined as

$$w_R(t) = \begin{cases} 0.5(\ 1 - \cos 2\pi t/T_R\ ), & (\ 0 \le t \le T_R\ ). \\ 0, & (\ \text{otherwise}\ ). \end{cases} \tag{15}$$

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

$$s_R(t) = w_R(t)\ s(t), \qquad (-\infty \le t \le \infty), \tag{16}$$

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

$$s_R^*(n) = \begin{cases} w_R(n\Delta t)\ s^*(n), & (n = 0,1,2,\ldots, M_R), \\ 0, & (\text{all other } n), \end{cases} \tag{17}$$

where M_R is the largest integer which 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

$$X_R(f) = \sum_{n=-\infty}^{\infty} s_R^*(n)\ e^{-j2\pi fn\Delta t}. \tag{18}$$

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

$$P(k) = |X_R(k/T_R)|^2 = \left| \sum_{n=0}^{M_R} s_R^*(n)\ e^{-j2\pi kn\Delta t/T_R} \right|^2. \tag{19}$$

Because the Hanning window now covers exactly four fundamental periods, harmonic peaks and valleys appear in every four steps of k. If the signal consists of pure harmonics, the hth main lobe consists of $P(4h-1)$, $P(4h)$ and $P(4h+1)$, and no side lobes appear in the valley, $P(4h+2)$. The shallower the valley, the larger the amounts of non-harmonic components. The smallest value of the signal power, $P(k)$, over the hth harmonic peak and valley, $4h-1 \le k \le 4h+2$, is taken as the power of the noise component for the hth harmonic peak, P_{Nh}. Therefore, the estimated power spectrum of the noise component, $P_N(k)$, is defined as

$$P_N(k) = P_{Nh} = \min_{i=-1,0,1,2} P(4h+i), \qquad (4h-1 \le k \le 4h+2), \tag{20}$$

where $h = 1,2,3,\ldots, L$. The noise-to-signal ratio, R_{NS}, is defined as

$$R_{NS} = 10\log\left(\sum_{k=3}^{4L+2} P_N(k) \Big/ \sum_{k=3}^{4L+2} P(k) \right). \tag{21}$$

Results

1. Analysis of synthesized voices

In order to study the sensitivity of the *N/S* ratio, voices synthesized by the SPEAK program[12] were analysed by the present method. A parameterized model of the glottal flow waveform was used, and this source model was noninteractive with the vocal tract. Voice samples were created with varying amounts of jitter, shimmer, additive noise, amplitude modulation, and frequency modulation. Voice samples were synthesized with 20,000 waveform samples per second.

Figure 1 shows the waveform and the power spectrum for an analysis frame of the synthesized voices, with 1%, 4%, and 16% additive noise in the glottal source. As expected, the valleys in the power spectrum become shallower as the additive noise increases. Figure 2 shows the *N/S* ratio for synthesized voices with varying amounts of additive noise. Each point is the mean result from 25 frames, shifted 6.4 ms each. Standard deviations are indicated by error bars. The figure shows that the *N/S* ratio varies with the amount of additive noise in the glottal source signal.

Figure 3 shows the averaged power spectrum of 25 frames with 1%, 4%, and 16% amplitude perturbation and 1/4%, 1%, and 4% pitch perturbation of the glottal source. The valleys in the power spectrum become shallower as the perturbation increases. Figure 4 shows the *N/S* ratio for synthesized voices with varying amounts of amplitude perturbation and pitch perturbation. The *N/S* ratio varies with the amount of the amplitude or pitch perturbation of the glottal source. 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.

2. Analysis of pre- and post-operative voices

Six subjects with benign laryngeal disease, who had mild or moderate hoarseness, were selected for study. All subjects underwent microscopic laryngeal surgery with sufficient perceptual voice quality after surgery to completely satisfy both surgeons and patients. The subjects were requested to read the Japanese sentence, /aoi uono eo kaita/, "I drew a picture of a blue fish". The sentence was read twice in both pre- and post-operative sessions. The recorded voice was digitized with 12-bit precision at a sampling rate of 10,000 samples per second. Figure 5 shows the waveform for the pre- and post-operative utterances of Subject 1. The sentence was read rather slowly and distinctly.

Figure 6 shows the time domain results for the pre- and post-operative voice samples of Subject 1. In order to extract

the vowel /u/ from the voice samples, three successive frames, whose averaged *N/S* ratio showed the minimum value, were taken as the representatives for each sample. Figure 7 shows the waveforms and power spectra for the selected three frames from the pre- and post-operative samples of Subject 1. The post- operative spectrum shows better harmonic structure than the pre-operative spectrum.

Table 1 shows the analysis results of the *N/S* ratio and the fundamental frequency for 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 8 shows the averaged *N/S* ratio of each pair (first and second readings) of pre- and post-operative 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.

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 hoarse voice in relation to both the perceptual impression of a listener and the pathological state of a larynx.

Classifications of listeners' impressions in multiple dimensions, such as rough, breathy, asthenic, and strained, have been proposed[13], and acoustic parameters associated with different kinds of voice quality have been studied[14,15]. 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. Relative strength of harmonics also contributes to the contrast between 'asthenic' and 'strained' voices.

The above kinds of acoustic parameters do not bear a simple relationship with pathological modes of vocal-fold vibration, 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 would be necessary to accurately extract individual glottal cycles from the acoustic signal and separate the glottal excitation signal from the non-specific spectral noise in each.

Inverse filtering has been proposed as a method for extracting source characteristics from the acoustic signal[4].

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 variations in formant patterns appeared to be caused by actual variations in source characteristics[16].

If we are to fully understand the acoustic characteristics of hoarse voice, we will have to learn much more about the relationship between 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 findings[6]: for whatever reason, a hoarse voice has a greater non-harmonic component and a less pure harmonic component than a normal voice.

The *N/S* ratio was calculated over the spectral region between the 1st and 16th harmonics. However, the voice signals were non-preemphasized and we analyzed the vowel /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 conditions, such as the sampling rate and the number of harmonics chosen, were wide enough to cover the most of the acoustic power of the voice. Improvement of these conditions did not affect the *N/S* measures for actual voice samples.

Acknowledgment

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

Table 1. The analysis result of the N/S ratio and the fundamental frequency.

Subject	Name	Age	Sex	Diagnosis	Pre-operation F0 (Hz)	Pre-operation N/S (dB)	Post-operation F0 (Hz)	Post-operation N/S (dB)
1	N.O.	39	M	Polyp	97.5 97.1	-25.8 -24.8	103.8 97.2	-29.0 -33.4
2	K.I.	46	M	Polyp	136.0 135.6	-20.1 -29.6	140.5 136.9	-34.3 -31.4
3	F.I.	29	M	Polyp	144.5 135.9	-29.8 -25.9	131.3 131.6	-32.4 -35.9
4	K.I.	35	F	Cyst	234.5 233.3	-34.9 -36.0	248.6 252.5	-40.4 -42.0
5	N.K.	30	F	Nodules	202.4 212.7	-29.3 -30.8	237.1 228.3	-42.5 -41.4
6	M.U.	46	F	Polyp	195.9 200.6	-34.5 -30.0	209.3 219.9	-36.7 -36.9

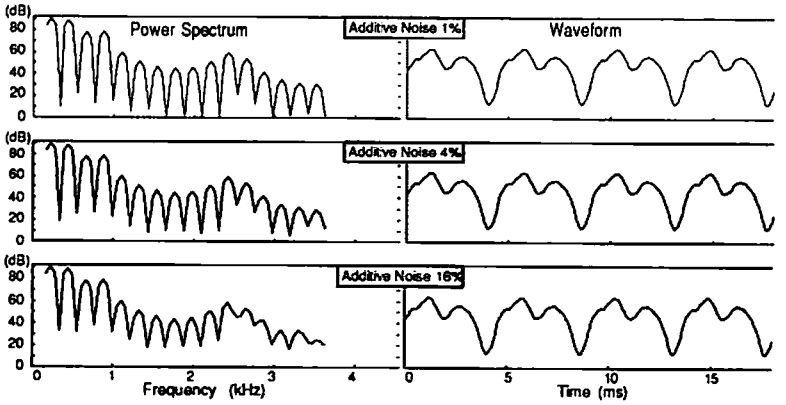

Fig.1. Waveform and power spectra for an analysis frame of the synthesized voices, vowel /u/, F0=220 Hz, with 1%, 4%, and 16% additive noise in the glottal source. The valleys in the pwer spectrum becomes shallower as the additive noise increases.

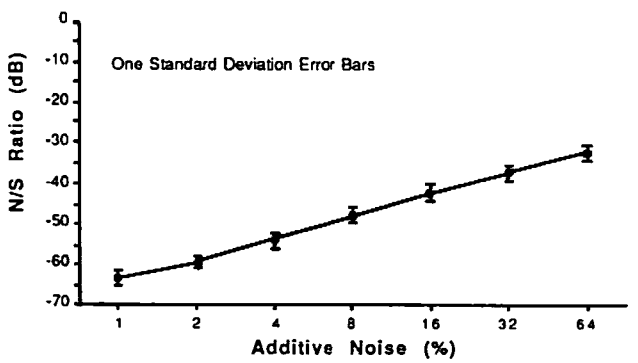

Fig.2. The N/S ratio for synthesized voices with varying amounts of additive noise in the glottal source. The N/S ratio varies with the amount of additive noise in the glottal source signal.

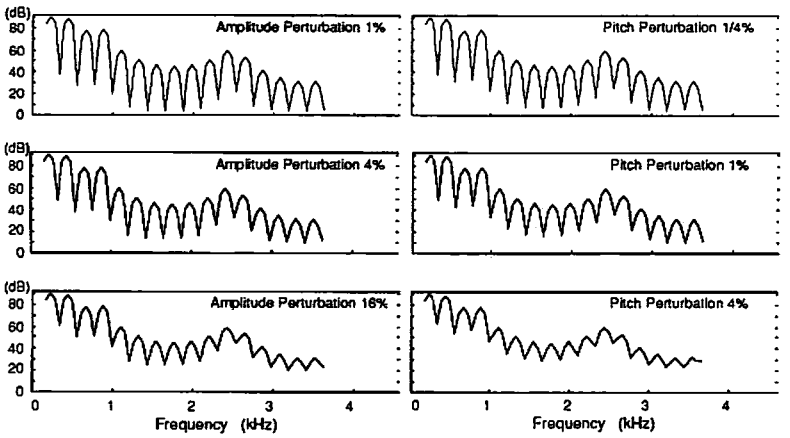

Fig.3. Averaged power spectra of 25 frames for the synthesized voices, vowel /u/, F0=220 Hz, with 1%, 4%, and 16% amplitude perturbation and 1/4%, 1%, and 4% pitch perturbation of the glottal source. The valleys in the power spectrum become shallower as the perturbation increses.

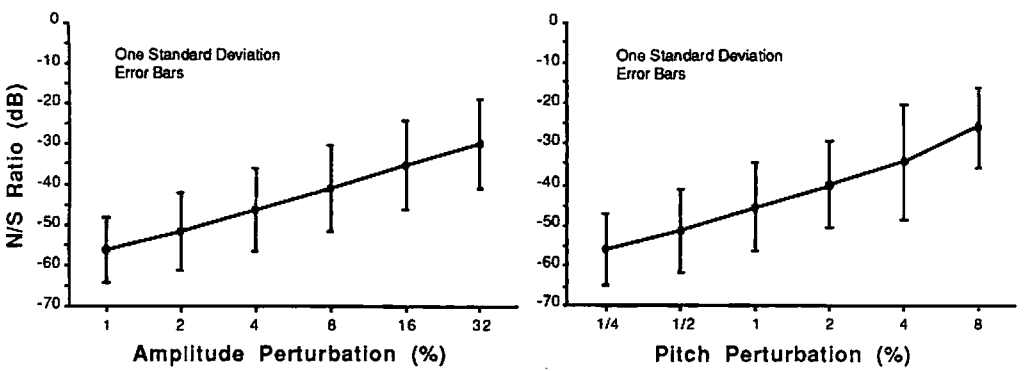

Fig.4. The N/S ratio for synthesized voices, vowel /u/, F0=220 Hz, with varying amounts of amplitude perturbation (left) and pitch perturbation (right) of the glottal source. The N/S ratio varies with the amount of amplitude or pitch perturbation.

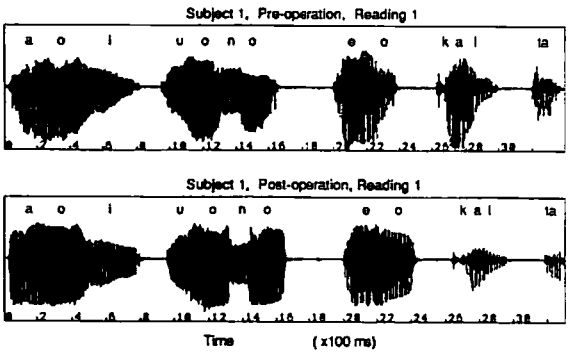

Fig.5. Waveforms of the sentence, /aoi uono eo kaita/, for the first pre-operative reading (top), and the first post-operative reading (bottom) by Subject 1.

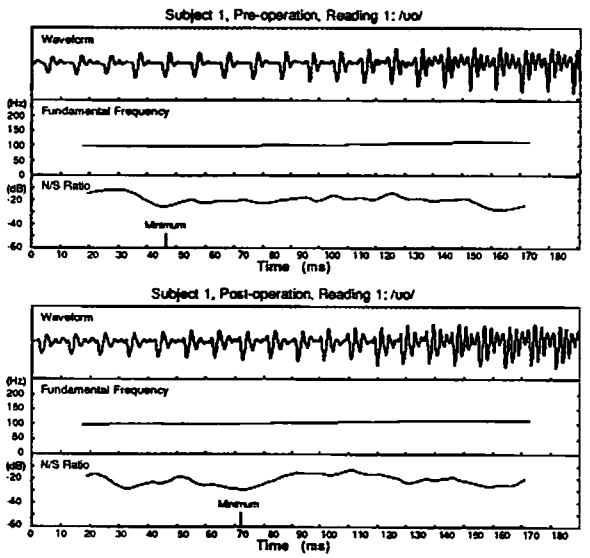

Fig.6. Time domain results for the first pre-operative reading (top) and the first post-operative reading (bottom) by Subject 1. The vertical bar in each bottom panel, which shows the minimum of the smoothed N/S ratio, indicates the most stable part of the vowel /u/.

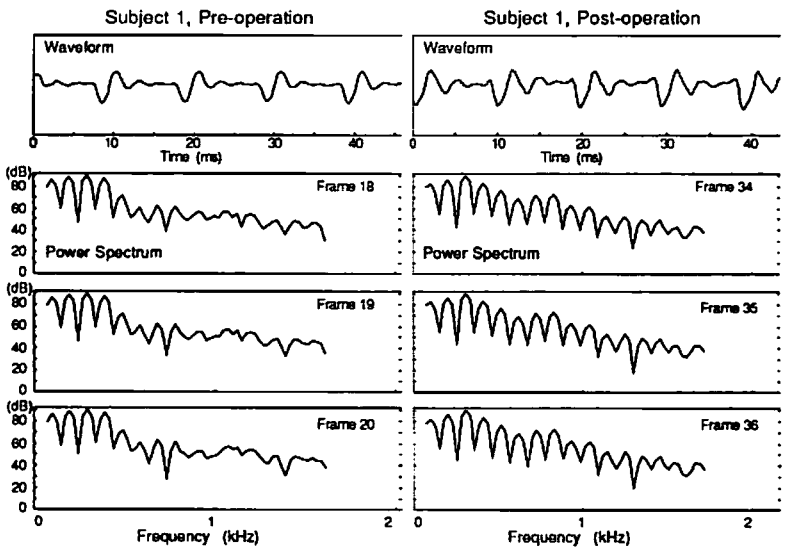

Fig.7. Waveforms and power spectra for the selected three frames, which showed the minimum N/S ratio, from the first pre-operative reading (left) and the first post-operative reading (right) by Subject 1.

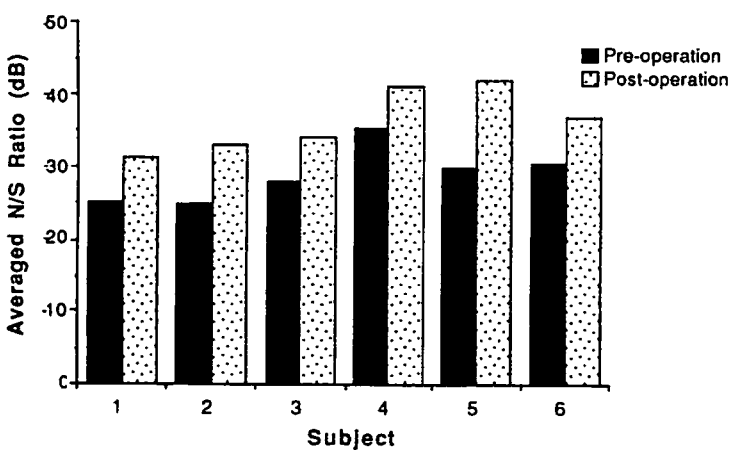

Fig.8. Averaged N/S ratio of each pair (first and second readings) of pre- and post-operative voice samples.

References

1) Isshiki, N., Yanagihara, N., and Morimoto, M.: Approach to the objective diagnosis of hoarseness. Folia Phoniat., 18, 393-400, 1966.
2) Lieberman, P.: Perturbations in vocal pitch. J. Acoust. Soc. Am., 33, 597-603, 1961.
3) Koike, Y.: Vowel amplitude modulations in patients with laryngeal diseases. J. Acoust. Soc. Am., 45, 839-844, 1969.
4) Davis, S. B.: Computer evaluation of laryngeal pathology based on inverse filtering of speech. SCRL Monograph 13., 1976.
5) Ludlow, C. L., Bassich, C. J., Connor, N. P., Coulter, D. C., and Lee, Y. J.: The validity of using phonatory jitter and shimmer to detect laryngeal pathology. In T. Baer, C. Sasaki and K. Harris (Eds.), Laryngeal function in phonation and respiration: Boston, Little, Brown and Company, pp.492-508, 1987.
6) Yanagihara, N.: Significance of harmonic changes and noise components in hoarseness. J. Speech Hear. Res., 10, 531-541, 1967.
7) Kasuya, H., Ogawa, S., Mashima, K., and Ebihara, S.: Normalized noise energy as an acoustic measure to evaluate pathologic voice. J. Acoust. Soc. Am., 80, 1329-1334, 1986.
8) Kojima, H., Gould, W. J., Lambiase, A., and Isshiki, N.: Computer analysis of hoarseness. Acta Otolaryngol., 89, 547-554, 1980.
9) Kitajima, K.: Quantitative evaluation of the noise level in the pathologic voice. Folia Phoniat., 33, 115-124, 1981.
10) Hiraoka, N., Kitazoe, Y., Ueta, H., Tanaka, S., and Tanabe, M.: Harmonic-intensity analysis of normal and hoarse voices. J. Acoust. Soc. Am., 76, 1648-1651, 1984.
11) Yumoto, E., Gould, W. J., and Baer, T., Harmonics-to-noise ratio as an index of the degree of hoarseness. J. Acoust. Soc. Am., 71, 1544-1550, 1982.
12) Titze, I. R.: Three models of phonation. J. Acoust. Soc. Am., Suppl. 1, 79, S81, 1986.
13) Hirano, M.: Clinical examination of voice, Wien, Springer-Verlag, pp.81-84, 1981.
14) Imaizumi, S.: Acoustic measure of roughness in pathological voice. J. Phon., 14, 457-462, 1986.
15) Imaizumi, S.: Clinical application of the acoustic measurement of pathological voice qualities. Ann. Bull. RILP, 20, 211-216, 1986.
16) Muta, H., Muraoka, T., Wagatsuma, K., Horiuchi, M., Fukuda, F., Takayama, E., Fujioka, T., and Kanou, S.: Analysis of Hoarse Voices Using the LPC Method. In T. Baer, C. Sasaki and K. Harris (Eds.), Laryngeal function in phonation and respiration: Boston, Little, Brown and Company, pp.463-474, 1987.