ANN.BULL.RILP(1991)No.25. 109-119

# Acoustical and Perceptual Characteristics of Pathological Voices: Rough, Creak, Fry and Diplophonia

Satoshi Imaizumi and Jan Gauffin[a)]

## Introduction

Although several perceptual scales[1-5] have been proposed to describe pathological voice quality, there is still a lot of debate on the number of scales needed, their physiological/acoustical/perceptual definitions, usage and reliability.

For instance, the Voice Committee for Phonatory Function Tests of the Japan Society of Logopedics and Phoniatrics has recommended use of the GRBAS scale[2] which consists of five 4-grade scales, i.e., for Hoarse (Overall Grade of Hoarseness), Rough, Breathy, Asthenic and Strained voice quality. On the other hand, Hammarberg et al.[3] have proposed 11 seven-grade scales, i.e., for Breathy, Aphonia, Hyperfunctional/Tense, Hypofunctional/Lax, Diplophonia, Voice breaks, Grating/Harsh, Rough/Coarse, Vocal fry/Creak, Register and Pitch. Eskenazi et al.[4] suggest 6 seven-grade scales, i.e., Overall severity, Hoarseness, Breathiness, Roughness, Vocal fry, and Excellence of normal voice.

The lack of clear definitions for these scales has been one of the biggest causes of discrepancy among research findings. For instance, although scales such as "Rough", "Creak", "Fry" and "Diplophonia" have been believed to correlate with pitch/amplitude perturbation in the voice waveform, no theory has been advanced to explain how and why these voice qualities differ from each other and how pitch/amplitude perturbations contribute to them.

This paper reports some results from our preliminary experiments to clarify the acoustical/perceptual characteristics of the voice qualities described by the scales "Rough", "Creak", "Fry" and "Diplophonia". Based on acoustic analyses of 102 voice samples, a synthesis model of pitch/amplitude perturbations was constructed. Through perceptual experiments on synthetic voice samples generated using the model, the following hypotheses H1-2 were tested. H1) The "Rough" quality is perceived when listeners holistically perceive the effect of perturbations as one coherent quality. H2)Other qualities are perceived when listeners analytically perceive the effect of perturbations as two or more separate sets of frequency components ("Diplophonia"), an additional sensation of repeating impulses corresponding to perturbation frequency ("Fry"), and the special case of "Fry" observed at final parts of voice ("Creak").

[a)] Department of Speech Communication and Music Acoustics
Royal Institute of Technology, Po Box 70014, S-100 44 Stockholm, Sweden

## Method

### Acoustic Analysis

Voice samples recorded from 102 patients with various laryngeal diseases were analyzed. Voice samples for /e/ were digitized through a 16-bit A/D converter at a sampling rate of 40 kHz and stored on a disk controlled by a computer. A 0.5 s segment was extracted by excluding the initial and final portions from each sample.

At first, using a peak picking method, local maximum points which could correspond to vocal excitation epochs were detected successively. Here, we write L(i) for the i-th pitch location, p(i)=L(i)-L(i-1) for the i-th pitch period, e(i) for the amplitude at L(i), for i=0, 1, .., I, where I was the total number of pitch periods extracted.

Then, cycle by cycle perturbation quotients pq(i) for p(i) and eq(i) for e(i) were calculated as

$$pq(i) = 100.0\{p(i)/p(i_M)-1.0\} \quad (\%) \quad (1)$$

$$eq(i) = 100.0\{e(i)/e(i_M)-1.0\} \quad (\%) \quad (2)$$

$$p(i_M) = \sum_{m=-M/2}^{M/2} w(m)p(i+m) \quad (3)$$

$$e(i_M) = \sum_{k=-M/2}^{M/2} w(m)e(i+m) \quad (4)$$

where w(m) was a hamming window of length M+1.

The pitch perturbation quotient, PPQ(%) and amplitude perturbation quotient APQ(%) were calculated as the following

$$PPQ = \frac{1}{I-M} \sum_{i=M/2}^{I-M/2} abs(pq(i)) \quad (5)$$

$$APQ = \frac{1}{I-M} \sum_{i=M/2}^{I-M/2} abs(eq(i)) \quad (6)$$

The coefficient of determination R between pq(i) and eq(i) was calculated as the following

$$r = CORR(pq(i), eq(i), I-1) \quad (7)$$
$$R = r*r$$

Here, CORR(X(n1), Y(n2), K) indicates the correlation coefficient between the two variables of length K, X(n1) and Y(n2).

Next, pq(i) and eq(i) were modeled by a AR process as follows.

$$pq(i) = bp*up(i)-\sum_{j=1}^{J} ap(j)*pq(i-j) \quad (8)$$

$$eq(i) = be*ue(i)-\sum_{j=1}^{J} ae(j)*eq(i-j) \quad (9)$$

where up(i) and ue(i) were Gaussian noise time series with a zero mean and unit variance, i.e., N(0, 1). ap(j) and ae(j) were AR coefficients estimated using the Burg method. The perturbation magnitudes bp and be were estimated as the root mean square values of the residue signals.

Setting z=exp(-2paifT0), where T0 is the average of p(i) for i=1,2,.., I, the spectrum of pq(i) and eq(i) were estimated from
the following

$$PQ(z) = bp/\{1+\sum_{j=1}^{J} ap(j)*z^{-j}\} \quad (10)$$

$$EQ(z) = be/\{1+\sum_{j=1}^{J} ae(j)*z^{-j}\} \quad (11)$$

As the parameters representing pq(i) and eq(i), pole frequencies and their bandwidths of PQ(z) and EQ(z) were calculated by solving the following

$$1+\sum_{j=1}^{J} ap(j)*z^{-j} = 0 \quad (12)$$

$$1+\sum_{j=1}^{J} ae(j)*z^{-j} = 0 \quad (13)$$

## Perturbation Synthesis

To synthesize the pitch period time series p'(i) and the amplitude time series e'(i) with perturbation, pq'(i) and eq'(i) were generated as

$$pq'(i) = bp*up(i)-\sum_{j=1}^{J} ap(j)*pq'(i-j) \quad (14)$$

$$eq'(i) = be*ue(i)-\sum_{j=1}^{J} ae(j)*eq'(i-j) \quad (15)$$

where bp and be were specified to control the perturbation magnitude, and ap(j) and ae(j) were specified as pole frequencies and their bandwidths to control the perturbation spectra.

Then, the time series of the pitch periods without perturbation pt(i) were generated using the F0 generation model known as the Fujisaki model[6]. The pitch period time series p'(i) was the following

$$p'(i) = pt(i)*\{1+0.01*pq'(i)\} \quad (16)$$

The amplitude time series e'(i) was generated as follows

$$e'(i) = et(i)*\{1+0.01*[sgn(r)*R*w*pq'(i)+(1-R)*eq'(i)]\} \quad (17)$$

where et(i) was the gross trend in the amplitude without perturbation and was set as 0.9 in this experiment. The coefficient of determination R was introduced in Eq. 17 to represent the correlation between pq(i) and eq(i).

In Eq. 17, however, R represents the correlation between the pitch period and the amplitude of voice source, whereas the correlation between pq(i) and eq(i) in Eq. 7 represents that of speech signal. The actual correlation between the pitch period and the amplitude of synthetic speech may be changed via resonance-harmonics interaction as pointed out by Horii(1989)[7]. This issue should be investigated in the future.

## Perceptual Experiment

As a preliminary perceptual experiment, six normally hearing subjects rated voice quality of the 102 voice samples using the GRBAS scale in the same way as reproted previously[8,9]. Based on the rating results, a few voice samples were selected and modelled for the following perceutual experiment using synthetic voice samples.

Two normally hearing subjects served as the listeners for the perceutual experiment using synthetic voice samples. They were medical doctors familiar with the terms "Rough", "Creak", "Fry" and "Diplophonia".

Eight synthetic voice samples were prepared to simulate various types of pitch perturbation. Table I shows the parameters varied in this experiment.

Voice samples 1 and 2 were synthesized so as to have multiple peaks in the perturbation spectra of pq'(i) and eq'(i), whose bandwidths were large. Voice samples 3 and 4 were synthesized to have a sharp spectrum peak located at F0/3 or F0/4. Voice samples 5 and 6 had a very low F0, 80/110Hz and one sharp spectrum peak at F0/2. Voice samples 7 and 8 were set to have a falling pitch contour and to have a large perturbation only at the final portion of the voice.

Perceptual evaluation was carried out by a two alternative forced choice procedure. For instance, to evaluate "Rough", subjects were asked to select one of a pair of voice samples which was deemed "Rougher" than the other. The selection percentage was calculated as the "Rough" score for each sample and by which each sample was ranked.

## Results and Discussion

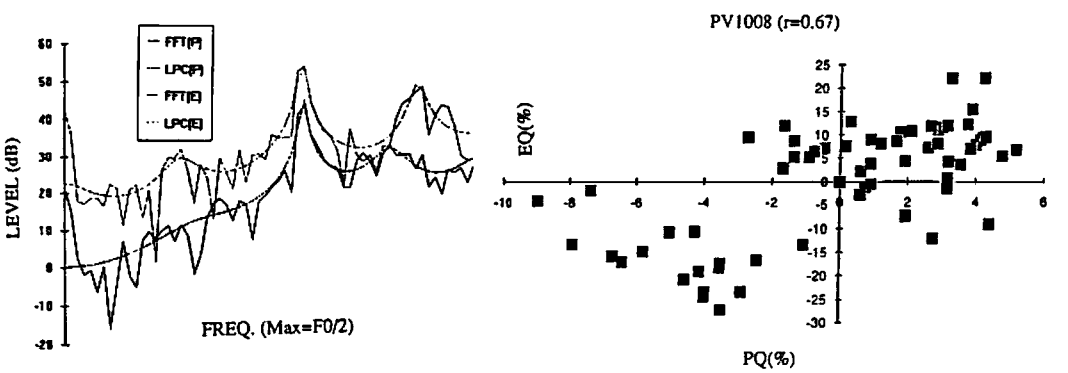
The acoustic results for the three pathological voice samples are shown in Figures 1-3. These voice samples were evaluated as strongly "Rough" in the preliminary perceptual experiment, and some of the listening subjects reported these voices as examples of "Diplophonia". Each figure shows the voice waveform, the FFT and LPC power spectra of pq(i) and eq(i), and the scatter diagram between pq(i) and eq(i) together with the value of r. Here, as noted above, r is the correlation coefficient between pq(i) and eq(i), pq(i) is the time series of the pitch perturbation quotient, and eq(i) is that of the amplitude perturbation quotient.

As shown in these figures, these voice samples have one or two high peaks in their FFT and LPC power spectra for both pq(i) and eq(i). The peak locations for pq(i) and eq(i) are very close to each other. The frequency of the highest peak varies between the voice samples. In Fig. 3, the highest peak is located at the maximum frequency, i.e., F0/2, where F0 is the average fundamental frequency and is the sampling frequency for pq(i) and eq(i) in this study.

The correlation between pq(i) and eq(i) also varies between the voice samples. The voice sample shown in Fig. 1 shows a positive correlation, r=0.67, whereas that of Fig. 2 shows a negative correlation r=-0.54. The voice sample shown in Fig. 3 shows a large positive correlation, r=0.94.

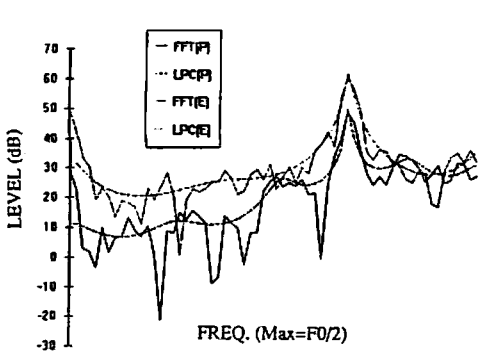
Table I. Synthesis parameters for 8 voice samples for the perceptual experimant. pt(F0) indicates the average fundamental frequency, and c indicates that F0 is constant whereas v indicates that F0 is variable. "m" in PQ indicates the multiple spectrum peaks with large bandwidths, whereas "1/n" indicates one sharp peak at a frequency of F0/n. "bp" is the magnitude of the pitch perturbation. The average amplitude was set as constant, and the amplitude perturbation was generated to correlate with the pitch perturbation by setting R=0.9. The formant frequencies and bandwidths were typical values for Japanese /e/.

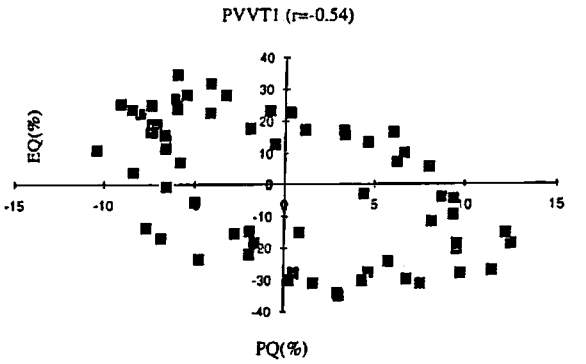
| Voice Samples | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| pt(F0) | 120 | 120 | 120 | 120 | 80 | 110 | 110 | 180 |
| | c | c | c | c | v | v | v | v |
| PQ | m | m | 1/3 | 1/4 | 1/2 | 1/2 | 1/2 | 1/2 |
| bp | 1.0 | 0.5 | 0.8 | 0.8 | 1.6 | 1.6 | 1.6 | 1.6 |
| | | | | | | | v | v |




(a) (b)

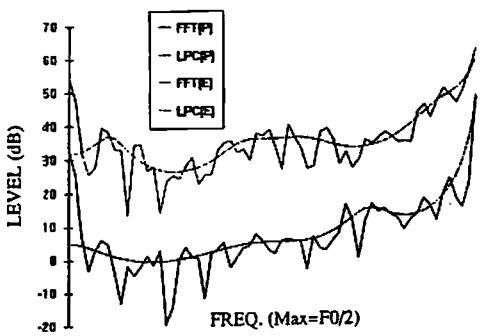
Fig. 1. Acoustic analysis results for pathological voice P1. (a) FFT and LPC power spectra of pq(i) and eq(i), and (b) scatter diagram between pq(i) and eq(i) with the value of r.

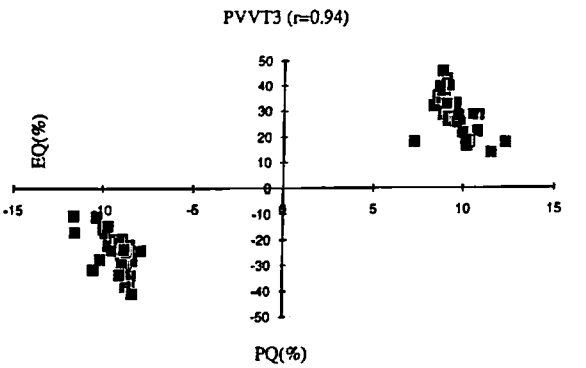



(a) (b)

Fig. 2. Same as Fig.1, but for pathological voice P2.




(a) (b)

Fig. 3. Same as Fig.1, but for pathological voice P3.

Fig. 4 shows a voice sample which was evaluated as "normal" in the perceptual study. The peaks in the FFT and LPC spectra of pq(i) and eq(i) are lower in dB than the voice samples shown in Fig. 1-3, and the bandwidths of the peaks are also larger than the voice samples shown in Figs. 1-3. The correlation coefficient is very low, r=0.01, for this voice sample.

These results indicate that pathological voice samples perceived as strongly "Rough" or "Diplophonia" tend to have large perturbations both in pitch period and in amplitude. The FFT and LPC spectra of pq(i) and eq(i) have poles whose bandwidths are smaller than those of "normal" voice samples. The pole frequencies vary among voice samples. These findings agree with the findings of the correlation analyses reported by Imaizumi(1985, 1986)[8,9] and those of the spectrum analyses reported by Hirama et al(1989)[10].

Fig. 5 shows the results of the perceptual experiment. The "Rough" score is relatively large for voice samples 1, 2 and 8, but does differ largely among the voice samples. The "Diplophonia" score is large for voice samples 2 and 3, and is not so large for the others. The "Fry" score is large for 5 and 4, and small for the others. The "Creak" score is large for voice samples 8 and 7, and small for the other samples.

The results shown in Fig. 5 indicate the following. 1)The synthesis model is capable of generating the differences between the four voice qualities denoted "Rough", "Diplophonia", "Fry" and "Creak." 2) The "Rough" score was relatively large for the voice samples in which the multiple spectrum peaks were set as dull, i.e., the bandwidths of the poles were large. However, since the "Rough" score was not so small even for the other voice samples, the "Rough" quality may have been perceived more or less for all the voice samples. 3) The "Diplophonia" score was relatively large for the voice samples, in which the peak location was set as F0/3 or F0/4, that is as lower than F0/2. 4) The "Fry" score was large for voice sample 5 which had a very low F0, 80Hz, and one sharp spectrum peak at F0/2. 5)The "Creak" score was large for voice samples 7 and 8, for which pitch was set with a falling contour and a large perturbation only at the final portion of the voice.

These tendencies suggest that our hypotheses H1-2 are valid. H1) The "Rough" quality is perceived when listeners holistically perceive the effect of perturbations as one coherent quality. H2) Other qualities are perceived when listeners analytically perceive the effect of perturbations as two or more separate frequency components ("Diplophonia"), an additional sensation of repeating impulses corresponding to the perturbation frequency ("Fry"), and the special case of "Fry" observed at the final parts of voice ("Creak").

The "Rough" score was relatively large for voice samples 1 and 2, for which the

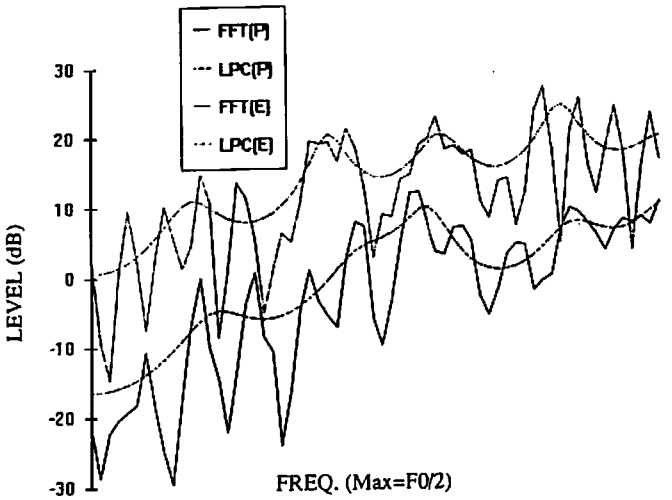



Fig. 4. FFT and LPC power spectra of pq(i) and eq(i) for normal voice N1.

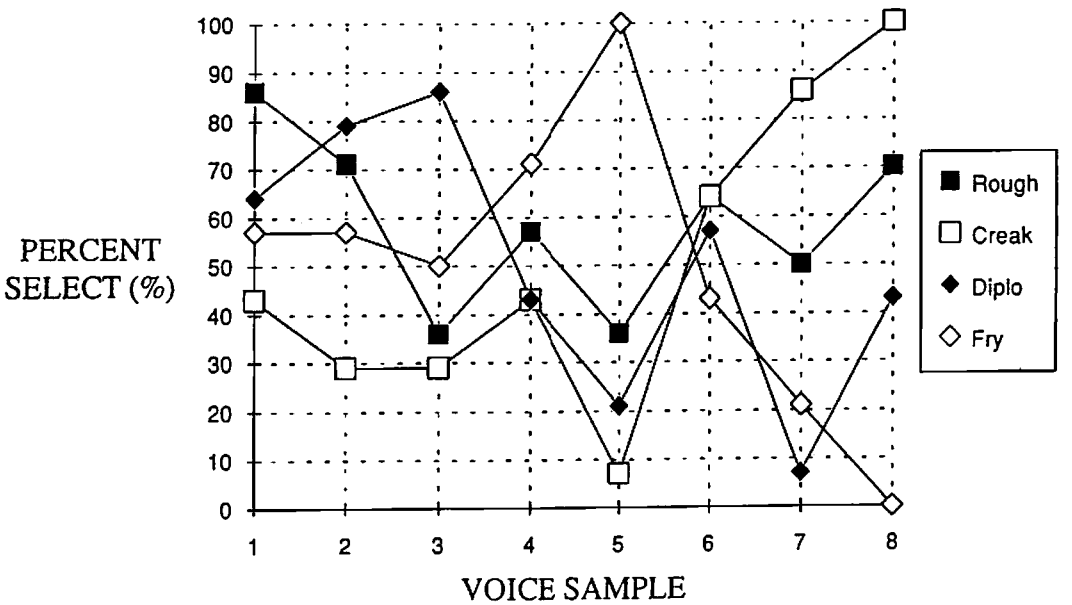



Fig. 5. Percent of selections for 8 synthetic voice samples in a paired comparison with reference to "Rough", "Fry", "Creak" and "Diplophonia."

perceptual effect of the multiple spectrum peaks may not have been clear enough to be perceived analytically. The "Diplophonia" score was relatively large for the voice samples, in which the peak location was set F0/3 or F0/4, that is as lower than F0/2. This slower perturbation may have been easier to perceive as a separate additional pitch. The "Fry" score was large for voice sample 5 which had a very low F0, 80Hz, and one sharp spectrum peak at F0/2. For this condition, the subharmonics were generated every 40Hz on the frequency axis by the perturbation. These subharmonics may have been perceived as repeating impulses at a rate of 40Hz, if the excitation was somewhat strong even in high frequency region. The "Creak" score was large for voice samples 7 and 8, for which pitch was set to have a falling contour and to have a large perturbation only at the final portion of the voice. For this condition, the same conditions as in voice sample 5 occurred only at the final portion.

## Conclusion

The acoustic correlates of "Rough", "Creak", "Fry" and "Diplophonia" in pathological voices were investigated based on an acoustic analysis of human voices and a perceptual evaluation of synthetic voices. The following results were obtained. 1) The synthesis model constructed based on the acoustic analysis was found capable of generating the differences between the four voice qualities denoted "Rough", "Diplophonia", "Fry" and "Creak." 2) The "Rough" quality seemed to be perceived when listeners holistically perceived the effect of perturbations as one coherent quality. 3)Other qualities seemed to be perceived when listeners analytically perceived the effect of perturbations as two or more separate sets of frequency components ("Diplophonia"), an additional sensation of repeating impulses corresponding to the perturbation frequency ("Fry"), and the special case of "Fry" observed at final parts of the voice ("Creak").

To generalize these results, however, more listeners must be tested since perceptual judgments might differ among listeners.

## References


1) N. Isshiki, H. Okamura, M. Tanabe and M. Morimoto:"Differential diagnosis of hoarseness.", Folia Phonaitrica, 21, 9-19(1969).
2) M. Hirano:"Clinical examination of voice.", Springer-Verlag, Wien, 81-84(1981).
3) B. Hammarberg: Clinical Routines for the perceptual-acoustical assessment of dysphonia.", Phoniatric and Logopedic Progress Report, Huddinge University Hospital, Karolinska Institute, 4, 14-29(1985).
4) L. Eskenazi, D.G.Childers and D.M. Hicks: Acoustic correlates of vocal quality.", J. Speech and Hearing Research, 33, 298-306(1990).
5) J. Laver: The phonetic description of voice quality.", Cambridge University Press, London(1980).
6) H. Fujisaki, K. Hirose and M. Sugito:"Comparison of acoustic features of word accent

in English and Japanese.", J. Acoust. Soc. Jpn. (E), 7(1), 57-63(1986).
7) Y. Horii:"Acoustic analysis of vocal vibrato: A theoretical interpretation of data.", J. Voice, 3(1), 36-43(1989).
8) S. Imaizumi:"Acoustic measures of pathological voice qualities.", Ann. Bull. RILP, 19, 179-190(1985).
9) S. Imaizumi: "Acoustic measures of roughness in pathological voice.", J. Phonetics, 14, 457-462(1986).
10) J. Hirama and Y. Kakita:"Characteristics of a pathological "Rough" voice based on the power spectrum of fluctuations." Japanese J Logopedics and Phoniatrics, 30, 225-230(1989).