Analysis of Autocorrelation-based Parameters in Creaky Voice

Carlos Toshinori Ishi *

JST/CREST, Human Information Science Laboratories, Department 1, ATR
2-2 Hikaridai, "Keihanna Science City", Kyoto, 619-0288 Japan

*e-mail: carlos@atr.jp

Keywords: creaky voice, phonation type, voice quality, autocorrelation.

1. Introduction

"Creaky voice" has many other terminologies, such as "creak", "vocal fry", "glottal fry", "laryngealization", "glottalization", and "pulse register phonation", used in several research areas (like Linguistics, Physiology, and Phonetics) [1,2].

Creaky voice is defined as "… a train of discrete laryngeal excitations, or "pulses", of extremely low frequency (7 to about 78 Hz), with almost complete damping of the vocal tract between excitations." (Hollien 66 cited in [1]). "The auditory effect is of a rapid series of taps, like a stick being run along a railing." (Catford 64 cited in [2]).

Creaky phonation carries many linguistic and paralinguistic information, depending on the language. For example, contrast between creaky and modal voicing among vowels and nasals is particularly common in some American Indian languages [3]. In [4,5], relationship between different phonation types and paralinguistic information like emotions and attitudes are reported for English. Strong correlations were reported between creaky voice and perception of relaxed/stressed, sad/happy, and bored/interested. In Japanese, expressive pressed voice that is frequently realized by creaky phonation also carries important paralinguistic information such as attitudes, emotional states and emphasis [6].

Further, in creaky segments, periodicity is disturbed and the pitch extraction becomes difficult, affecting the subsequent prosodic analysis, like intonation. Tendency of creaky segments for specific tone types is reported in [7] for phrase finals in Japanese.

In the JST/CREST ESP Project [8], one of the goals is an expressive speech synthesizer based on unit selection, using a large database of spontaneous speech. For this purpose, labels of voice qualities (phonation types) become as important as prosodic labels. With the goal of doing automatic labeling of voice quality on a large speech database, in the present research, we focus on the automatic detection of creaky phonation.

2. Acoustic features of creaky voice

Acoustical analysis of creaky voice includes disturbance of periodicity in the time domain (jitter and shimmer) [11], feature extraction in the power spectrum [3,10], and parameterization of the glottal excitation waveform obtained from vocal tract inverse filtering of the speech signal [4,9].

Jitter (perturbation in the fundamental frequency) and shimmer (perturbation in the amplitude) are two measures of perturbation in the periodicity in the time axis. There are many works [11] showing their correlation with perceptual roughness. However, direct correlation with creaky phonation is not reported.

[3] describes that creaky phonation has the following features compared to modal phonation: non-periodic glottal pulses, lower power, lower spectral slope, low F0. Among them, the spectral slope is reported to be the most important parameter to discriminate between different phonation types. In [3], the spectral slope is estimated based on harmonic components of the power spectrum. [10] also estimates the spectral slope based on harmonics of the spectrum, but considering the effects of the formants. However, this kind of method using harmonic components could not be appropriate for non-periodic signals.

Another approach for discriminating phonation types is based on speech production. The basic idea is removing the effects of the vocal tract resonances from the speech signal by inverse filtering techniques, to obtain the glottal excitation waveform. In the research field of speech synthesis based on speech production models, the glottal excitation waveform is parameterized according to the shape of each glottal pulse. [4,9] reports successful synthesis of different voice qualities, including creaky voice, by controling the parameters of the LF model.

However, automatic detection of creaky voice is not as widely reported. Perhaps because the glottal excitation is irregular and automatic detection becomes difficult.

3. Autocorrelation-based parameters

In the present research, in order to avoid the detection of excitation pulses in the temporal domain, we propose a parameterization of phonation type features based on the autocorrelation of the glottal excitation waveform.

3.1. Estimation of the glottal excitation waveform

First, LPC coefficients are estimated after pre-emphasis of the input signal. Then, the LPC residual signal is obtained by inverse filtering of the input signal using the estimated LPC coefficients. The residual signal is treated as the glottal excitation waveform hereinafter.

3.2. Normalized Autocorrelation Function (*NACF*)

Before estimating the autocorrelation function (*ACF*), the glottal excitation waveform is low-pass filtered at 2 kHz, in order to make the *ACF* peak detection easier.

An important point to be taken into account is the window size for *ACF* estimation. Since creaky voice usually appears in low fundamental frequencies, the window size should be long enough to cover at least two excitation pulses. On the other hand, a too long window size is not appropriate for segments with high and changing pitch. Therefore, we decided to use an analysis window with variable length.

A two-step *ACF* estimation is used to adjust the window length adaptively. First, *ACF* is estimated in an 80 ms window. And then, the time lag of the maximum peak is extracted and multiplied by four, to be used as the new window size. Here, the new window size was clipped to lie in the interval between 16 ms and 80 ms.

The obtained ACF is normalized according to the following expression:

$$NAC(L) = \frac{N}{N-L}\frac{Rxx(L)}{Rxx(0)}, \qquad (1)$$

where N is the number of samples of the frame window, L is number of samples of the autocorrelation lag, and Rxx is the autocorrelation function.

Figure 1 shows examples of glottal excitation waveforms and normalized autocorrelation functions obtained using the methods described above.

For modal phonation (a), a clear periodicity can be observed; the *NACF* peaks are close to 1 value, and there are no small peaks between the time lag 0 and the first big peak. (b) and (c) show examples of creaky voice with big-small-big-small and short-long-short-long sequences (jitter/shimmer) of the glottal pulses. (b) shows a smaller peak between the time lag 0 and the maximum peak. The magnitude of this small peak becomes lower and the width of this peak becomes larger, as the jitter/shimmer becomes stronger, such that it divides in two, as shown in (c). In the modal phonation example, it can also be observed that the first two peaks (closer to time lag 0) have values close to 1. (d) shows an example of (non double-periodic) creak, where only one big *NACF* peak can be observed. However, a narrow width is observed for this peak, because of the impulse-like shape of the glottal excitation for creak phonation.

3.3. *NACF*-based parameters

Based on visual inspection of the *NACF* of the glottal excitation waveforms of modal and creaky phonations as described in the previous section, we decided to use the first two peaks (called *P1* and *P2*, from the time lag 0) in the *NACF*, to characterize different phonation types. A threshold of 0.2 is used to detect peaks in *NACF*.

The following parameters are proposed based on these two peaks (*P1*, *P2*).

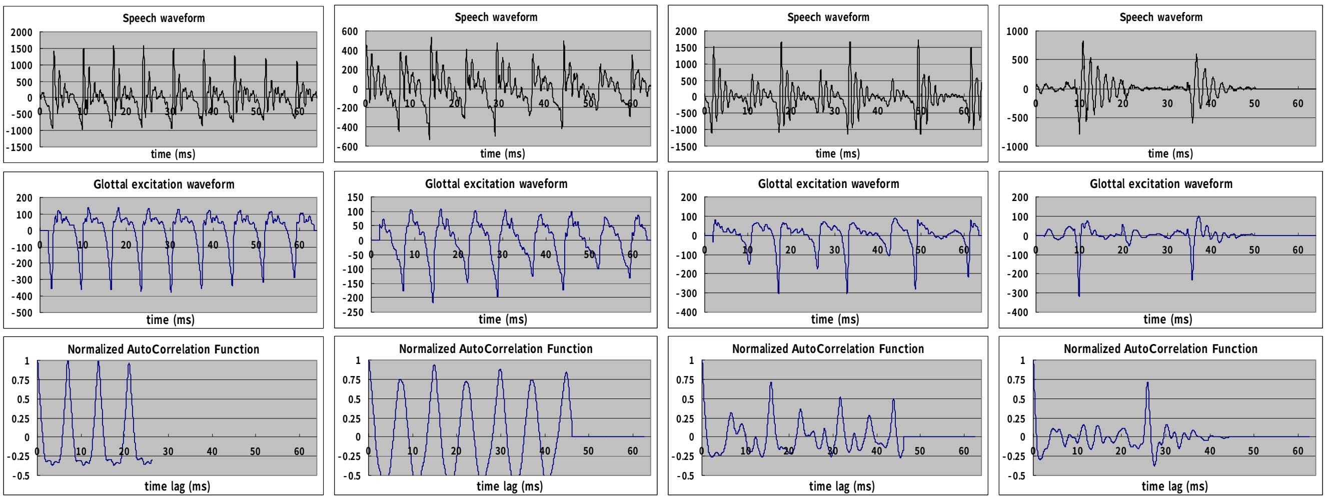
- **Peak magnitude (*NAC* value) ratio:**
  $NACR = NAC(P2) \,/\, NAC(P1) \qquad (2)$
- **Peak position (time lag) ratio:**
  $TLR = TL(P2) \,/\, (2{*}TL(P1)) \qquad (3)$
- **Peak width ratio:**
  $WR = W(P2) \,/\, W(P1) \qquad (4)$
- **Maximum peak magnitude:**
  $NACmax = NAC(Pmax) \qquad (5)$
- **Maximum peak position:**
  $TLmax = TL(Pmax) \qquad (6)$
- **Maximum peak width:**
  $Wmax = W(Pmax) \qquad (7)$

For single-periodicity, all the ratios (*NACR*, *TLR*, *WR*) and *NACmax* are expected to have values close to 1. For double-periodicity, $NACR > 1$; $NACmax < 1$; if jitter is strong, $TLR \neq 1$; and if jitter or shimmer is strong, $WR < 1$.

For low F0 creaky phonation with non-double periodicity, i.e., large interval between excitation pulses (big *TLmax*), there are cases where only one peak can be detected. Therefore, the ratio-based parameters cannot be used to represent these signals. However, a small value of *Wmax* is expected in these cases, since creaky phonation has narrow (impulse-like) excitation pulses.

3.4. Preliminary evaluation of the proposed parameters

As speech data for evaluation, we used the same dataset



(a) modal phonation *(b) low jitter creaky phonation* *(c) high jitter creaky phonation* *(d) low F0 creaky phonation*

Figure 1: *Speech waveform, glottal excitation waveform and NACF for Modal and Creaky phonation*

analyzed in [7], containing 404 phrase final syllables segmented from natural spontaneous speech of a female adult speaker. Each syllable was labeled in terms of Creaky (C), Aspirated (A) or Modal (M), looking at the waveform and hearing the segments. The parameters proposed in Section 3 were estimated in all frames (5619) of the annotated speech intervals.

As a preliminary evaluation, a decision tree was constructed for each of the categories {C,A,M}, using the *R* Package. The tree resulted in 91.5% of correct identification. Specifically for Creaky category, deletion error was 13.7%, while substitution error was 7.9%. However, only the parameter set {*NACmax*, *NACR*, *TLR*} was used in the constructed decision tree. A more detailed analysis is necessary to verify the contribution of each parameter for the phonation type discrimination.

4. Conclusion

Parameters based on the normalized autocorrelation function of glottal excitation waveform were investigated with the aim of automatically detecting creaky voice segments. Preliminary evaluation of the proposed parameters showed good performance in the automatic detection of creaky voice. Detailed analysis is now been conducted to verify the behavior of each parameter in the phonation type discrimination. Other topics to be investigated are to verify if these segments are really perceived as rough, and how pitch is perceived in the creaky intervals.

5. Acknowledgements

I would like to thank Nick Campbell and the whole JST/CREST group for supporting the present research. Special thanks for Parham Mokhtari (ATR) and Ken-ichi Sakakibara (NTT) for advice and motivating discussions.

6. References

[1] Gerratt, B. R., Kreiman, J., 2001. Toward a taxonomy of nonmodal phonation. *J. of Phonetics* 29, 365-381.

[2] Laver, J., 1980. Phonatory settings. In *The phonetic description of voice quality*. Cambridge University Press, Ch. 3, 93-135.

[3] Gordon, M., Ladefoged, P., 2001. Phonation types: a cross-linguistic overview. *J. of Phonetics* 29, 383-406.

[4] Gobl, C., Ní Chasaide, A., 2003. The role of voice quality in communicating emotion, mood and attitude. *Speech Communication* 40, 189-212.

[5] Klasmeyer, G., Sendlmeier, W. F., 2000. Voice and Emotional States. In *Voice Quality Measurement*, Singular Thomson Learning. Ch. 15, 339-358.

[6] Sadanobu, T. 2003. Expressive Speech and Grammar: with special reference to pressed voice in Japanese. *JST/ CREST Int. Workshop on Expressive Sp. Proc.*, 55-60.

[7] Ishi, C.T., Mokhtari, P., Campbell, N. 2003. Perceptually-related Acoustic-Prosodic Features of Phrase Finals in Spontaneous Speech. *Eurospeech 2003*, 405-408.

[8] The JST/CREST Expressive Speech Processing project, introductory web pages at: www.isd.atr.co.jp/esp

[9] Childers, D.G. 1995. Modeling the glottal volume-velocity waveform for three voice types. *J. Acoust. Soc. Am.* 97 (1), 505-519.

[10] Hanson, H. M., Stevens, K., Kuo, H. J., Chen, M., Slifka, J., 2001. Towards models of phonation. *J. of Phonetics* 29, 451-480.

[11] Buder, E.H. 2000. Acoustic Analysis of Voice Quality: A Tabulation of Algorithms 1902-1990. In *Voice Quality Measurement*, Sing. Thomson Learning, Ch. 9, 119-244.