2—Q—9

Speech Production Process Analysis by Clustered Line-Spectrum Modeling

◎Yoshinori TAKAHASHI †, Mikio TOHYAMA † ‡, David M. Howard ‡

† Waseda University(Tokyo, Japan) ‡ Media Engineering, University of York(York,UK)

Introduction

This paper investigates the characteristics of the speech pressure waveform that correspond to the phase changes of vocal fold motion using Clustered Line-Spectrum Modeling (CLSM)[1]. The closed(open) quotient of the vocal fold motion is a key aspect of the source waveform in speech and singing voice analysis[2, 3]. This phase information can be estimated using the output(Lx) of the electrolaryngograph[4, 5]. However, in practice, it can be very inconvenient to obtain the Lx waveform in situ, particularly due to the need of an elastic neckband. CLSM can decompose speech waveforms into resonant types of responses, in order to find the timing information closely related to the phase changes estimated from Lx which is measured synchronously with the speech pressure waveform recording.

1. Speech Waveform Analysis by CLSM[1]

CLSM represents a waveform as a summation of sinusoidal components clustered around spectral peaks frame by frame. Those clustered components can be estimated by obtaining the least error solutions in the frequency domain by including the effects of the time-window function.

Figure 1 shows examples of analysis of vowels (A: [ah] in English and B: in Japanese) using CLSM. Three dominant spectral peaks in every short frame, each of which is composed of three sinusoidal components, are extracted such that the residual components might be small as shown in plots (f) and (i). Some of those spectral peaks can be characterized as formants.

2. Phase Changes of Vocal Fold Vibration

The waveform or spectral characteristics of vocal fold vibration might be closely related to voice quality[2-4, 6]. Quantifiable aspects of vocal fold vibration such as the larynx closed quotient can be estimated on a cycle-by-cycle basis from Lx[4-5]:

$$CQ \equiv (CP/T_X) \times 100 \quad (\%) \qquad (1)$$

where CP denotes the duration of the closed phase and T_X gives the period of a cycle.

This quotient is particularly important for singing voice analysis and characterizing the effect of singer's training [3]. It could be useful to estimate this quotient from the sung pressure waveform without synchronously recording Lx.

Figures 2(a) and (b) show the speech waveforms as in Fig.1 (a) and(b) along with the time-aligned simultaneously recorded Lx waveform in Fig.2(d). The solid line shows the start of the closed phase which corresponds to a positive peak in the differential (e) of the Lx waveform[4]. The solid line can be seen to correspond to the points where the 1st component extracted by CLSM in Fig.2b has a negative peak. The dashed line shows the end of the closed phase where the negative-going Lx waveform crosses a preset level where the peak-to-peak amplitude is divided into 4:3[3]. This phase change seems to be closely related to the positive pulse occurrences in Fig.2(b). That is, the closed phase might be located in the time interval between the negative and positive peaks of the 1st CLSM component shown in Fig.2b.

The occurrences of the positive peaks can be alternatively illustrated by the energy decay curve defined on a cycle-by-cycle basis by:

$$E(\tau) = \int_{\tau}^{T_X} f^2(\tau) d\tau \qquad (2)$$

where $f(t)$ represents the waveform in Fig.2b.

Several stages in the decay curve are visible. The broken line seems to mostly correspond to the initial portion of the second stage of the decay. If we estimate the closed quotient from the positions marked by open circles in Fig. 2(c), then we get the CQ estimates shown. These can be compared with CQ values calculated from Lx in Fig.2d. No pair of CQ values is more than 2.4 % adrift.

Summary

We have shown that CLSM is able to provide the timing information from speech waveforms, which identifies the phase changes of vocal fold vibration. In particular the low frequency component extracted by CLSM might be useful to estimate the larynx closed-phase quotient. The authors acknowledge Prof. Y. Yamasaki for his proposal of cooperation work and encouragement.

References

[1] M. Kazama, et al. J. Audio Eng. Soc. 51(3) pp.123-137(2003)

[2] D. H. Klatt, J. Acoust. Soc. Am. 87(2), pp.820-857(1990)

[3] D. M. Howard, J. Voice 9(2) pp. 163-172 (1995)

[4] D. M. Howard et al, J. Voice 4(1) pp. 205-212 (1990)

[5] A. K. Krishnamurthy et al. ASSP 34(4) pp. 730-743(1986)

[6] A. E. Rosenberg, J. Acoust. Soc. Am. 49 pp.583-590 (1971)

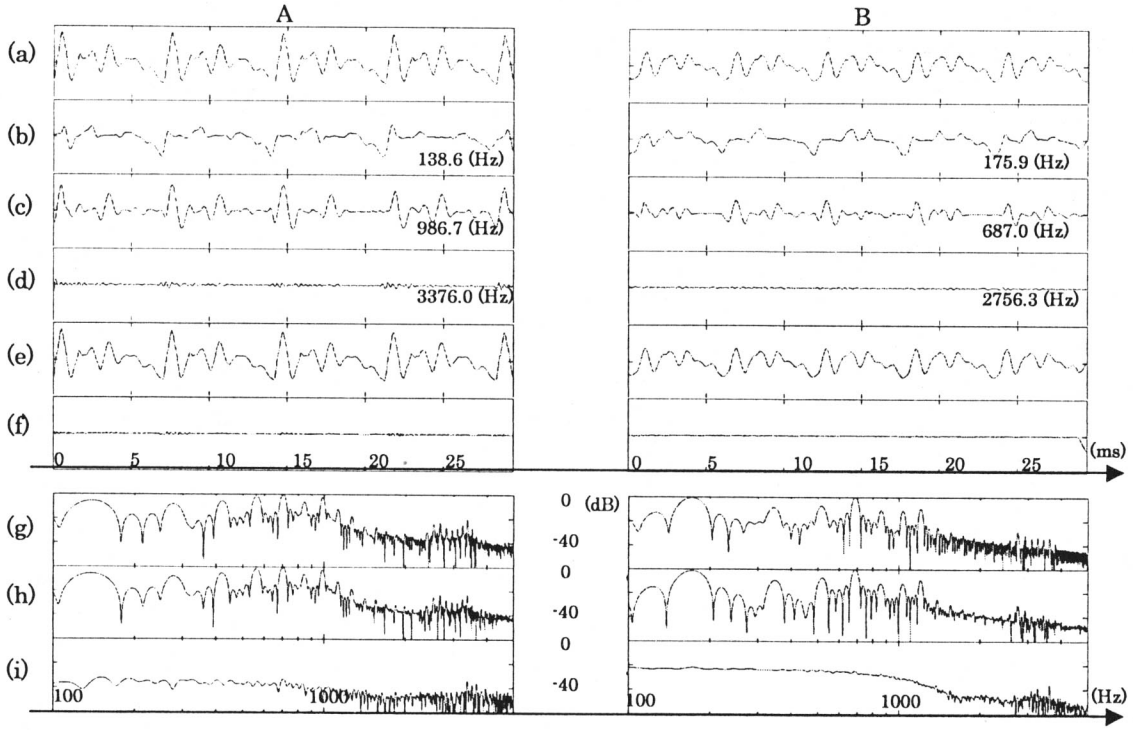

Fig.1 Samples of speech waveform analysis and synthesis by CLSM
Fig.1A: 「ah」 in English, Fig.1B: 「ah」 in Japanese
(a) Original speech waveform, (b) 1st component (lowest frequency), (c) 2nd component, (d) 3rd component (highest frequency), (e) Synthesized by (b)+(c)+(d), (f) Residual by (a)-(e), (g) Power spectrum of (a), (h) Power spectrum of (e), (i) Power spectrum of (f)

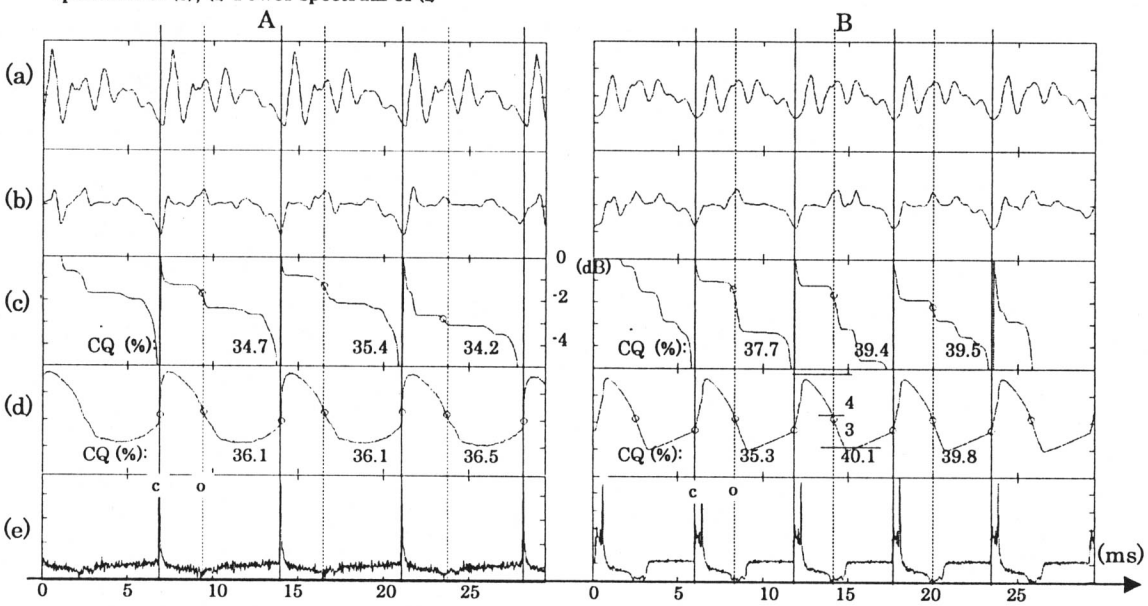

Fig.2 Comparison of the estimates of glottal closed phase from Lx and from CLSM analysis
Fig.2A: 「ah」 in English, Fig.2B: 「ah」 in Japanese
(a) Original speech waveform, (b) 1st component, (c) Energy decay curve of (b), (d) Time-aligned Lx waveform, (e) Differential of (d). The markings "c" and "o" indicate the start of the closed and open phase, respectively. On the Lx waveform (d), these are calculated automatically, where "c" corresponds to a positive peak in the differential (e) of the Lx waveform (d) and "o" to the point where the negative-going Lx waveform crosses a preset level which divides its peak-to-peak amplitude by 4:3.