2006 IEEE International
Symposium on Signal Processing
and Information Technology

RESONANT BANDWIDTH ESTIMATION OF VOWELS USING CLUSTERED-LINE SPECTRUM MODELING FOR PRESSURE SPEECH WAVEFORMS

O. Yasojima [1] *, Y. Takahashi* [2] *, and M. Tohyama*[1,2]

Waseda University
Global Information and Telecommunication Studies
Shinjuku Tokyo, Japan

ABSTRACT

The estimation of resonant frequency bandwidths is a fundamental issue related to the quality of spoken vowels and vocal-tract acoustics. In this article, we discuss our analysis of bandwidths using clustered line-spectrum modeling (CLSM) of the pressure waveforms of vowels on a cycle-by-cycle basis with reference to Lx waveforms from an electrolaryngograph recorded at the same time as the speech signal. We used CLSM to decompose the waveforms into three dominant resonant (modal) oscillations with almost exponentially decaying envelopes. The modal (so-called formant) frequencies were observed in a wide frequency range from 100 (Hz) to over 4 (kHz). The modal bandwidths were estimated from the decaying constants of the modal oscillations and were wider than those reported in the literature under the closed glottis condition. The bandwidths increased for both male and female speakers as the formant frequencies became higher. The bandwidths for females, however, were wider with greater variances than those for males. We could effectively represent a cycle of a vowel record shorter than 10 (ms) by CLSM based on the least squares error criterion in the frequency domain. We thus confirmed that cycle-by-cycle analysis using CLSM is a practical approach to characterizing vowel sounds in terms of dominant frequencies using their modal bandwidths.

1. INTRODUCTION

Formants and their frequency bandwidths are fundamental properties for characterizing a vowel produced through the vocal tract [1]. The bandwidths have been conventionally estimated by the sweep tone method [2][3] under the closed glottis condition. However, we are interested in characterizing naturally pronounced vowels from sound pressure records in a normal conversational situation from the point of view of speech quality representation. In this paper, we describe a way to estimate the frequency bandwidths of vowels from their pressure waveforms through clustered line-spectrum modeling (CLSM) [4] on a cycle-by-cycle basis.

If we assume that a vowel record in a single period is a transient response of the vocal tract, we can interpret formant frequencies and their bandwidths as resonant frequencies and modal bandwidths in terms of modal vibration analysis. CLSM is a method for representing a short interval of a signal, including the envelope, using the dominant sinusoidal components on a least squares error (LSE) basis in the frequency domain.

In this article, we confirm that CLSM decomposes a vowel signal record into its dominant frequency components, which are distributed in a wide frequency range from 100 Hz to 4 kHz, even though a single period of a vowel is less than 10 ms. These dominant components of vowels, which correspond to modal resonant responses of the vocal tract, allow us to estimate decay constants by fitting with exponentially decaying functions. Modal frequency bandwidths can be derived from the decaying constants according to the modal vibration theory [3].

2. CLSM FOR SHORT SIGNAL INTERVALS

Spectral peak picking is an effective way to represent a signal observed in a short interval without artifacts due to a signal-windowing function [4]. If several sinusoidal components are located closely together in the frequency domain, however, it is difficult to represent the signal along with its envelope in a short observation interval.

CLSM was developed by Kazama, Yoshida, and Tohyama [4] to estimate the true sinusoidal components from a clustered spectral record that cannot be estimated through the peak-picking process. If a target signal is composed of a finite number of clustered sinusoidal components, the components can be estimated by obtaining an LSE-based solution using the over-determined simultaneous equations in the frequency domain [5][6] instead of the time region where conventional sinusoidal modeling is applied.

[1] Graduate School of Global Information and Telecommunication Studies, Waseda University

[2] Kogakuin University

[1] Global Information and Telecommunication Institute, Waseda University

[2] University of York, UK

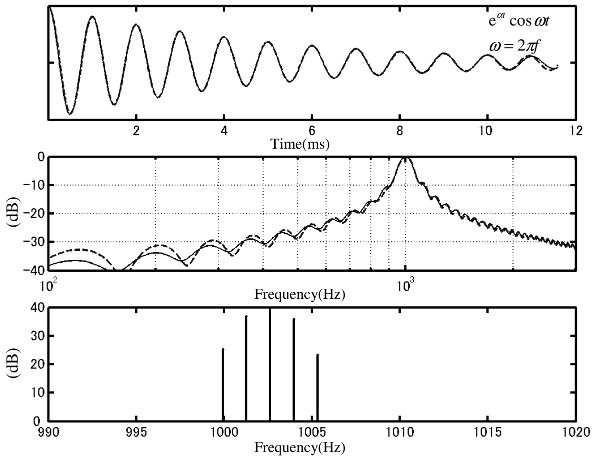

Fig. 1. An example of CLSM for a decaying signal: (a) time-waveform, $f = 1(\mathrm{kHz}), \alpha = 200$, (b) FFT power spectrum, (c) CLSM record for $P = 5$ and $L = 7$. Solid line: original, broken line: reconstructed

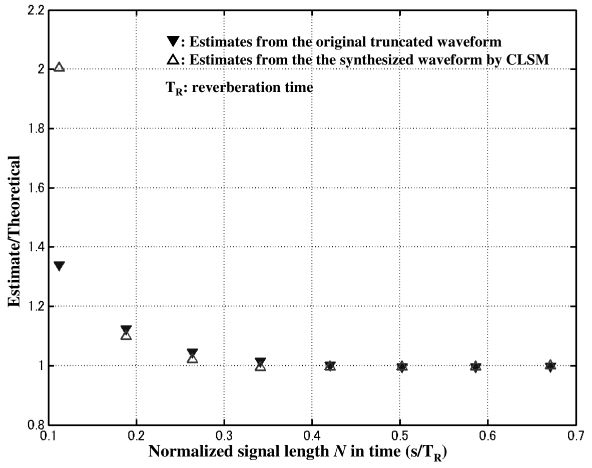

Fig. 2. An example of estimating the decay constant for the waveform shown in Fig. 1 truncated at length N. △: estimates from truncated original waveforms. ▼: from the CLSM synthesized waveforms after truncation.

Suppose we have a signal with a record length of N and the interpolated spectrum is analyzed by taking the M-point FFT after zero-adding. We assume that the target signal can be expressed in an analytic form for the narrow frequency band as

$$x_a(n) \equiv \sum_{k=1}^{K} A(k)e^{j2\pi f(k)n} + \epsilon_K(n), \qquad (1)$$

where the K components are clustered around the peak at $k = k_p$, A_k and f_k respectively denote the k-th sinusoidal component's complex magnitude and frequency, and $\epsilon_K(n)$ denotes the residual component including the modeling error and external noise. If we attempt to represent the signal by clustered $P(\leq K)$ sinusoidal components between $k = k_p - m$ and $k = k_p - m + P - 1$, the P parameter sets can be estimated based on the LSE criterion by using a set of linear equations for L observation frequency points between $k = k_p - l$ and $k = k_p - l + L - 1$:

$$X_o = WX_s. \qquad (2)$$

Here, X_o denotes the observed spectrum at L frequency points, X_S is the P spectrum components (magnitude and phase) to be estimated for the signal, and W gives the matrix representing spurious spectra due to the window function $w(n)$ with a length of N. W^T denotes the transpose of W, and here we set

$$W_{NM}(q) \equiv \frac{1}{N}\sum_{n=0}^{N-1} w(n)e^{-j\frac{2pikn}{M}}\bigg|_{k=q} \qquad (3)$$

and $L > P$, $l > m$

$$m \equiv \begin{cases} \frac{P-1}{2} & P : odd \qquad l \equiv \frac{L-1}{2} \qquad L : odd \\ \frac{P}{2} & P : even \qquad l \equiv \frac{L}{2} \qquad L : even \end{cases} \qquad (4)$$

.

An example of how CLSM makes it possible to suitably describe a waveform including the envelope is shown in Fig. 1. The solid line in Fig.1a shows a decaying sinusoidal waveform whose frequency is 1 kHz, and the broken line is the waveform synthesized through CLSM by setting $P = 5$ and $L = 7$ around the center of the FFT spectral peak of the solid line in Fig. 1b. The broken line in Fig.1b shows the FFT spectrum for the CLSM synthesized signal. Figure 1c shows the line spectral components extracted by CLSM. CLSM is clearly an effective method for signal representation that includes a smooth envelope.

3. ENERGY DECAY CONSTANTS AND FREQUENCY BANDWIDTHS

If we have a resonant decaying waveform of $y(n)$ for length N, the energy decaying curve in terms of room reverberation acoustics can be written as[8]

$$E(n) = \sum_{m=n}^{N-1} y^2(m). \qquad (5)$$

We would normally expect this type of curve to follow an exponential decay trend. The decay constant α in such a case can sometimes be estimated by fitting an exponential decay curve on a least squares error basis, but the length of the observation signal N is not always sufficient for estimating the decay constant.

Figure 2 shows an example of estimating the decay constant for the waveform shown in Fig. 1 truncated at length N. The estimates can be compared to the theoretical one for the original waveform. We also plotted the estimates from the CLSM synthesized waveforms after the truncation. The signal length N was normalized by the reverberation time, which represented the 60-dB decay time given by

$$T_R \cong 6.9/\alpha. \qquad (6)$$

The results indicate that we can get good estimates for both the original and CLSM waveforms provided that length N is longer than $T_R/3$. The modal bandwidth B(Hz) can be described in hertz as

$$B = \alpha/\pi \qquad (7)$$

following the standard definition [3].

4. ESTIMATION OF FORMANT FREQUENCY BANDWIDTHS FOR VOWELS FROM PRESSURE WAVEFORMS

The formant frequency bandwidth [1][2][3], which can be interpreted as the modal bandwidth for the resonant frequency of the transfer function through the vocal tract, is a fundamental property that can be used to characterize the quality of a vowel. If we want to estimate bandwidths from pressure waveforms obtained for normally pronounced vowels, we have to analyze the modal resonances from the records of utterances on a cycle-by-cycle basis. The range of modal frequencies, however, extends from below 100 Hz to 2 or 3 kHz. For the low frequency components, the period of a vowel, about 10 ms, is not long enough for the conventional frequency analysis. This means CLSM might be a good approach to get estimates based on the inherent nature of the modal vibration of the vocal tract.

If a cycle of a vowel waveform represents the transient response of the vocal tract, then a decaying waveform reconstructed from CLSM around a dominant spectral peak might give us an estimate of the resonant decay constant and the resonant bandwidth. Figure 3 illustrates an example of CLSM analysis for a single vowel cycle. The sample was recorded in a sound booth along with Lx waveforms from an electrolaryngograph, so that we could easily carry out a cycle-by-cycle analysis of vowels. The plots in the left column are time waveforms, where panel (a) is the original waveform, (c) shows the first dominant component extracted by CLSM, and (e) displays the residual component after subtraction of the dominant component in panel (c) from the original waveform. The curves in the right column similarly display the respective FFT power spectral records, and panel (g) shows the magnitude of the clustered line spectra extracted by CLSM. By repeating the CLSM analysis process three times, we obtained three dominant components.

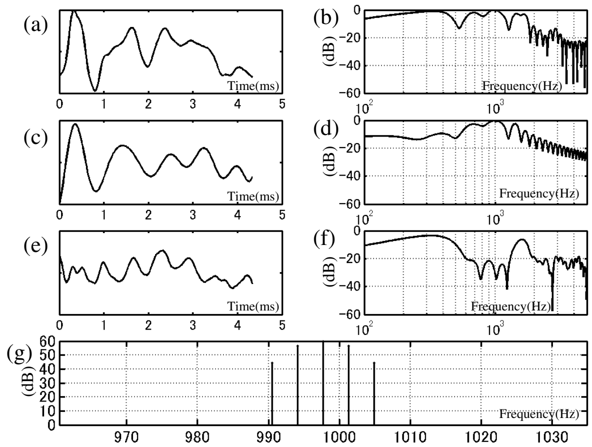

Fig. 3. Examples of CLSM analysis for female-spoken vowel ”a”: (a) time-waveform, (c)the first-dominant component extracted by CLSM, (e) the residual waveform after (c) is subtracted from(a). In addition, (b), (d), and (e) are FFT power spectra of (a), (c), and (e), respectively, and (g) shows the clustered line spectra extracted for the first dominant component by CLSM

Figure 4 shows samples of the energy decay curves calculated following Eq. (5)for the dominant resonant response extracted by CLSM for a male (Fig. 4a) and a female (Fig. 4b) speaker. The slope of the broken line indicates the least squares error estimate of the decay constant. The slope for the female speaker was steeper than that for the male, indicating that the formant bandwidth was wider for the female speaker. The reverberation time for the decaying curve for the male speaker was slightly longer than 10 ms, so the ratio of the signal length and the reverberation time exceeded 1/3.

Figure 5 shows distributions of the frequency bandwidths for the first, second, and third formant in panels (a), (b), and (c), respectively; these were estimated from the slopes of the energy decay curves shown in Fig. 4. For the bandwidth estimation, we used 300 records (obtained from three males and three females for five vowels over ten 10 cycles). The formant frequencies ranged from 100 (Hz) to over 4 (kHz). The formant bandwidths estimated here for the naturally spoken vowels were greater than 100 Hz and wider than those under the closed glottis condition reported in the literature [2][3]. The bandwidths increased for both male and female speakers as the formant frequencies rose. The formant frequencies for females, however, had wider bandwidths with greater variances than those for males. This corresponds to the conventional data obtained for the closed glottis condition. A cycle of a vowel record shorter than 10 (ms) can be effectively represented by CLSM based on the least squares error criterion in the frequency domain. We thus confirmed that cycle-by-cycle analysis using CLSM is a practical approach to characterizing vowel sounds in terms of dominant frequencies with their modal bandwidths.

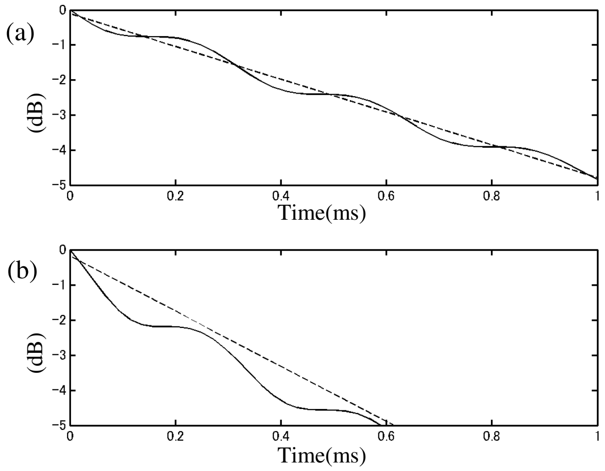

Fig. 4. Examples of energy decay curves for the modal responses of vocal tracts for the speakers: (a) male speaker and (b) female speaker. The broken lines were estimated based on LSE.

5. SUMMARY

We have shown that the formant bandwidths of vowels from naturally pronounced utterances can be estimated using clustered line-spectrum modeling (CLSM) on a cycle-by-cycle basis. If we assume that the pressure waveform of a vowel in a single period is a transient response of the vocal tract to pulse-like excitation, we can use CLSM to decompose the response into about three dominant resonant oscillations with almost exponentially decaying envelopes, and then estimate the formant frequency bandwidths characterizing vowel waveforms using the decay constants from the dominant modal responses. We estimated the formant frequencies within a wide range from 100 (Hz) to over 4 (kHz); however, the bandwidths were wider than those reported in the literature because of the glottis condition for the naturally spoken vowels. The bandwidths increased for both male and female speakers as the formant frequencies rose. The bandwidths for females, however, were wider with greater variances than those for males.

A cycle of a vowel record shorter than 10 (ms) could be effectively represented by CLSM based on the least squares er-

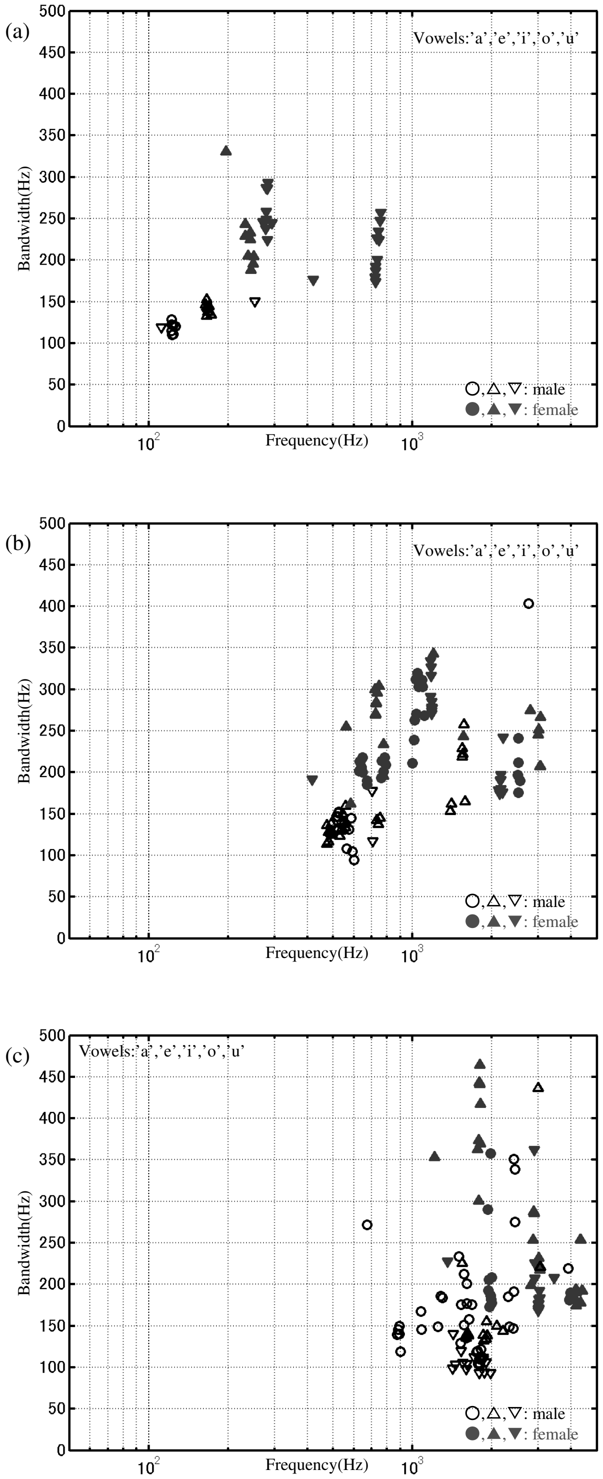

Fig. 5. Distribution of the frequency bandwidths estimated from the energy decay curves shown in Fig. 4, where ○ and ● respectively indicate male and female records: (a) the first formant; (b) and (c) bandwidth of the second and third formants, respectively.

ror criterion in the frequency domain, since the signal lengths were longer than 1/3 of the reverberation time estimated from the signal records. We confirmed that cycle-by-cycle analysis using CLSM is a possible approach to characterizing vowel sounds in terms of dominant frequencies with their modal bandwidths; however, the effect of the previous response cycle on the current analysis cycle is a problem that will have to be dealt with in the future. We thank Prof. D. Howard (University of York, UK) for his cooperation in recording the voice samples with Lx waveforms.

6. REFERENCES

[1] M.R. Schroeder, Computer Speech, Springer, 1999"

[2] O. Fujimura and J. Lindqvist, Sweep Tone Measurements of Vocal-Tract Characteristics, J. Acoust. Soc. Am. 49(2) pp. 541-558 (1971)

[3] K. N. Stevens, Acoustic Phonetics, The MIT Press (2000)

[4] M. Kazama, K. Yoshida, and M. Tohyama, Signal Representation Including Waveform Envelope by Clustered Line-Spectrum Modeling, J. Audio Eng. Soc. 51(3) pp.123-137 (2003)

[5] T. Quatieri and R. Danisewicz, An Approach to Co-Channel Talker Interference Suppression Using a Sinusoidal Model for Speech, IEEE Trans ASSP 38 pp. 56-69 (1990)

[6] R.C. Maher, Evaluation of a Method for Separating Digitized Duet Signals, J. Audio Eng. Soc. 38 pp. 956-979 (1990)

[7] Y. Hirata, A method for Monitoring Invisible Changes in a Structure Using Its Non-Stationary Vibration, J. Sound and Vib. 270 pp.1041-1044 (2004)

[8] M.R. Schroeder, Integrated-Impulse Method Measuring Sound Decay without Using Impulse, J. Acoust. Soc. Am. 66 pp.497-500 (1979)