SINUSOIDAL AND ENVELOPE-MODULATION MODELING OF ACOUSTIC SIGNALS

PACS REFERENCE: 43.71.An, 43.72.Ar, 43.72.Ja

Mikio TOHYAMA
Kogakuin University Bldg. #5 Room 601
2665-1 Nakano-machi, Hachioji-shi, Tokyo, 192-0015 Japan
Tel: +81-3-3342-1211 (Ext. 3422)
Fax: +81-3-3348-3486
E-mail: ctdeneuv@sin.cc.kogakuin.ac.jp

ABSTRACT

An acoustic signal representation based on sinusoidal and envelope-modulation modeling is described. Such studies are crucial to acoustic events modeling for structured audio-signal representation and rendering through interactive networks. This paper confirmed that; (1) sinusoidal modeling is useful for constructing an intelligible speech using only a few dominant components, (2) envelope modulation modeling enables modification of a talker's pitch and speech rate without sacrificing intelligibility through use of the simple carriers, and (3) the narrow-band envelope could be also estimated by clustered line-spectrum modeling (CLSM) based on the least square error criterion in the frequency domain.

INTRODUCTION

An immersive communication system requires rendering of acoustic events to enable effective collaboration through an interactive network – i.e., an interactive sound field network (ISFN) [1-3]. An acoustic events modeling language (AEML) is essential to acoustic events rendering based on structured audio representation [4]. This article describes a signal theoretical approach to rendering acoustic events. Signal theory itself can be an effective tool for acoustic engineering, since the human ear might be sensitive only to a resultant signal waveform. However, a signal theory for acoustic events must be deeply and effectively related to the source signal content or information that ensures a listener understands acoustic events through the signal sensation. The author will focus on speech intelligibility as the fundamental signal information that underlies acoustic events. This is because speech communication through acoustic events is a critical function, although the source signal information of acoustic events varies. This article, which mainly focuses on the representation of intelligible speech from the viewpoint of envelope-modulation modeling, should contribute significantly to AEML development.

SINUSOIDAL ANALYSIS BY ITERATIVE SPECTRUM-PEAK PICKING BY FFT

Signal representation, rendering, or morphing requires signal modeling that uses the fewest possible parameters, and spectrum-peak picking is a possible means of such signal representation [5]. Let us assume a target signal can be expressed in an analytic form as

$$x_a(n) \equiv \sum_{k=1}^{K} A_k e^{j2pf_k n} + e_k(n), \qquad (1)$$

where A_k and f_k, respectively, denote the k-th sinusoidal component's complex magnitude and frequency, K is the number of dominant sinusoidal components, and $e_K(n)$ denotes the residual component including the modeling error and external noise.

The dominant components can be chosen by the following procedure.

Step 1: Take the M-point FFT of the signal in the analytic (complex) form after zero-adding.

Step 2: Find the component $X(k_p)$ that has the maximum in the power spectrum record obtained in Step 1.

Step 3: Subtract the maximum component from the original signal

$$e_a(n) \equiv x_a(n) - X(k_p)e^{j\frac{2pk_p n}{M}} \quad \text{for } n = 0,1,\cdots N-1,$$

and newly set $x_a(n) \leftarrow e_a(n) \quad \text{for } n = 0,1,\cdots N-1.$

Step 4: Repeat steps 1 to 3 until $\sum_{n=0}^{N-1} |e_a(n)|^2 < E,$ where E denotes the allowable error. Figure 1 shows an example of short-frame speech analysis. This demonstrates that peak-picking is a good way to estimate the short-time spectrum.

SPECTRUM ESTIMATION BY CLUSTERED LINE-SPECTRUM MODELING

The principle, that is called clustered line-spectrum modeling (CLSM), was developed to estimate the true signal components from the clustered spectrum records. This CLSM makes it possible to describe signals including the envelope. If a target signal is composed of a finite number of sinusoidal components, the signal components can be estimated by obtaining the LSE-based solution using the over-determined simultaneous equations in the frequency domain instead of the time region where conventional sinusoidal modeling is performed [6].

Suppose that a signal with a record length of N and the interpolated spectrum is analyzed by taking the M-point FFT after zero-adding. Assume again that the target signal is as expressed by Eq. (1) in the narrow frequency-band where the K components are clustered around the peak at $k = k_p$. If we attempt to represent the signal by clustered P (<=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 $k = k_{p-l+L-1}$ as

$$\mathbf{x}_{observe} = W\mathbf{x}_{signal}$$

, where

$$\begin{pmatrix} X(k_{p-l}) \\ \vdots \\ X(k_{p-l+L-1}) \end{pmatrix} \equiv \mathbf{x}_{observe} \quad \text{and} \quad \begin{pmatrix} X_S(k_{p-m}) \\ \vdots \\ X_S(k_{p-m+P-1}) \end{pmatrix} \equiv \mathbf{x}_{signal}$$

denotes the observed spectrum at L frequency points and the P spectrum components for the signal, respectively,

$$W_{NM}(q) \equiv \frac{1}{N}\sum_{n=0}^{N-1} w(n) e^{-j\frac{2pkn}{M}} \Big|_{k=q},$$

$$W \equiv \begin{pmatrix} W_{NM}(k_{p-l} - k_{p-m}) & \cdots & W_{NM}(k_{p-l} - k_{p-m+P-1}) \\ \vdots & \ddots & \vdots \\ W_{NM}(k_{p-l+L-1} - k_{p-m}) & \cdots & W_{NM}(k_{p-l+L-1} - k_{p-m+P-1}) \end{pmatrix}$$

for $L>P$, $l>m$,

$$m \equiv \frac{P-1}{2}, \quad P:\text{odd}, \quad \equiv \frac{P}{2}, \quad P:\text{even} \quad l \equiv \frac{L-1}{2}, \quad L:\text{odd}, \quad \equiv \frac{L}{2}, \quad L:\text{even}.$$

The spectrum estimates can be obtained by solving the LSE solutions as

$$\hat{\mathbf{x}}_{signal} = (W^{\mathrm{T}}W)^{-1}W^{\mathrm{T}}\mathbf{x}_{observe}.$$

An example of CLSM making it possible to suitably describe a waveform including the envelope is shown in Fig. 2. Figure 2a shows a ¼-octave-band speech sample with a smooth envelope whose center frequency is 500 Hz. The dominant peak in the power spectrum is shown in Fig. 2b. Figure 2c shows the synthesized waveform obtained by applying the CLSM where $L = 7$ observations for $P = 5$ clustered signal components centered at the dominant peak. CLSM is clearly an effective means of signal representation that includes a smooth envelope.

ENVELOPE MODULATION MODELING

The temporal envelope contains, however, a signal signature directly related to speech intelligibility rather than a sinusoidal spectrum component [7]. This section describes a form of speech-signal representation that uses narrow-band temporal envelopes and carriers. Envelope modulation modeling makes it possible to modify a speaker's pitch and speech-rate without losing intelligibility or sacrificing voice quality. Figure 3 shows a basic example of envelope-modulation modeling of speech signals. A speech signal can be synthesized as

$$s(n) = \sum_k e_k(n)\cos \boldsymbol{f}_k(n)$$

where $e_k(n)$ and $\boldsymbol{f}_k(n)$ denote, respectively, the envelope and instantaneous phase in the k-th frequency band. The narrow-band carriers can be replaced by sinusoidal carriers estimated from the greatest magnitude spectrum components by peak-picking frame-by-frame in each frequency band with almost perfect intelligibility and without loss of a speaker's voice tonality.

In the envelope-modulation method of speech modeling, the speech-rate and pitch of a speaker can be modified. Figure 4 shows examples of distributions for the fundamental frequencies estimated by the auto-correlation analysis frame-by-frame for the original (a) and the synthesized (b). Similarly (c) and (d) illustrate the distribution examples for the modified signals that were obtained by lowering (c) (or raising (d)) the

frequencies of every peak-picked sinusoidal carrier frame by a constant ratio. We confirmed the pitch-conversion by listening, although a speaker's individuality or tonality was not completely maintained. Figure 5 shows again distribution examples similar to Fig.4 for the modified speech signals that were obtained by 2-times-stretching (a) (or ½-shrinking (b)) the narrow-band envelopes. Intelligible speech with a modified speech rate can be synthesized while preserving a speaker's pitch characteristics.

CONCLUSION

In this article the author has described a signal theoretical approach to acoustic-signal modeling of acoustic events that renders particular speech-signal modeling. The envelope is a key consideration in the understanding of signal content through signal-waveform analysis. Sinusoidal and envelope-modulation modeling of acoustic signals have been demonstrated. Consequently, clustered line-spectrum modeling (CLSM) including an iterative process and a method of intelligible speech representation that uses narrow-band envelopes and their sinusoidal carriers were presented. This envelope-modulation method enables modification of the speaker's voice pitch and speech rate without sacrificing intelligibility.

ACKNOWLEDGEMENTS

The author is extremely grateful to Prof. Tammo Houtgast of Amsterdam Free University, The Netherlands. The author thanks him for his valuable discussions, suggestions, and able guidance. This research was partly supported by Telecommunication Advanced Research Organization, Japan, and the International Communications Foundation, Japan.

REFERENCES

[1] L. Sacioja et al., Creating Interactive Virtual Acoustic Environments, J. Audio Eng. Soc., vol. 47 pp. 675-705 (1999)

[2] C. Kyriakakis et al., Surrounded by Sound, IEEE Signal Processing Magazine, vol. 1, pp. 55-66 (1999)

[3] C. Kyriakakis, Fundamental and Technological Limitations of Immersive Audio Systems, Proc. IEEE 86 pp. 941-951 (1998)

[4] B.L. Vercoe, W.G. Gardner, and E.D. Scheirer, Structured Audio: Creation, Transmission, and Rendering of Parametric Sound Representations, Proc. IEEE 86 pp. 922-940 (1998)

[5] M. Kazama and M. Tohyama, Estimation of Speech Components by ACF Analysis in a Noisy Environment, J. Sound & Vib. 240 pp. 41-52 (2001)

[6] E. B. George, and M. J. T. Smith, Analysis-by-Synthesis/Overlap-Add Sinusoidal Modeling Applied to the Analysis and Synthesis of Musical Tones, J. Audio Eng. Soc., 40 pp. 497-516 (1992)

[7] R. Drullman, Temporal Envelope and Fine Structure Cues for Speech Intelligibility, J. Acoust. Soc. Am, 97(1) pp.585-592

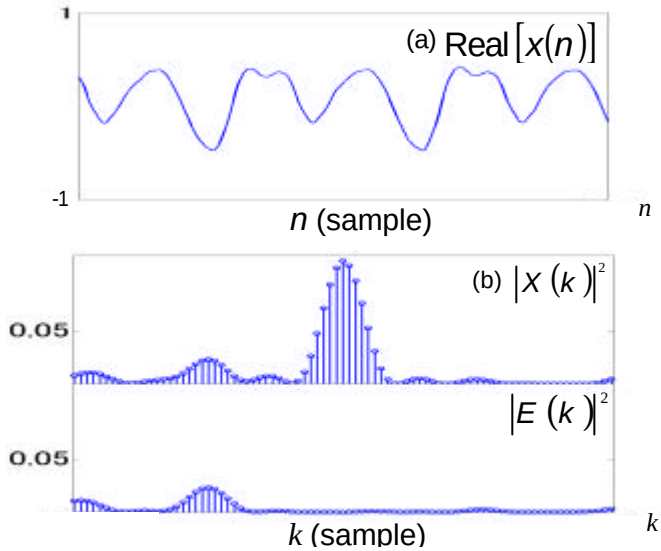

Fig. 1. Speech representation by interpolated spectrum peaks
(a) a short time frame of speech
(b) observed (upper)and subtracted (lower) spectrum records by peak-picking sampling frequency is 48kHz

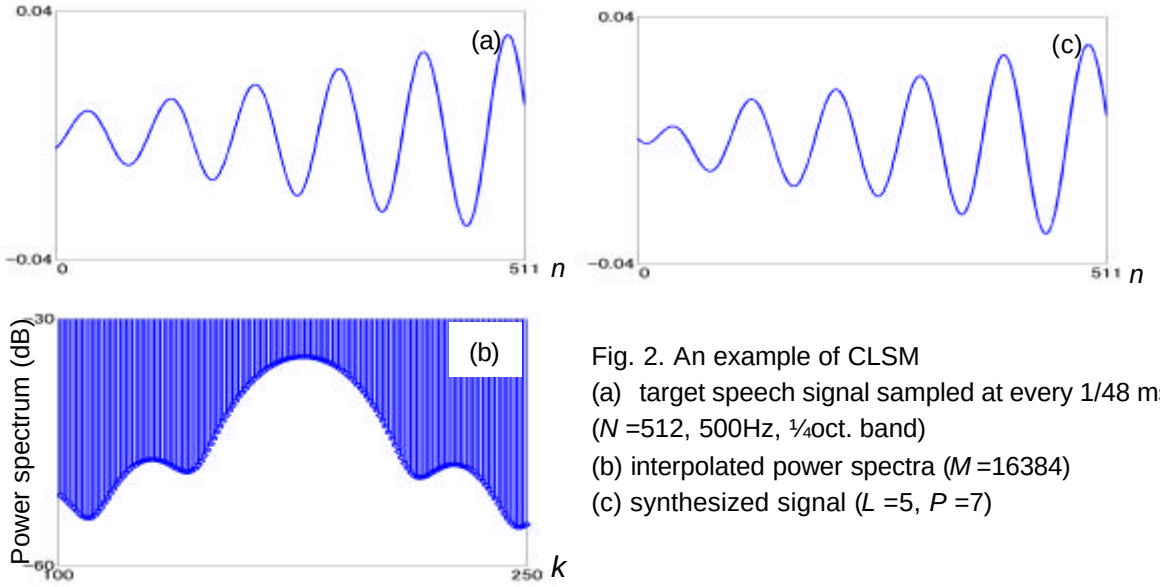

Fig. 2. An example of CLSM
(a) target speech signal sampled at every 1/48 ms (*N* =512, 500Hz, ¼oct. band)
(b) interpolated power spectra (*M* =16384)
(c) synthesized signal (*L* =5, *P* =7)

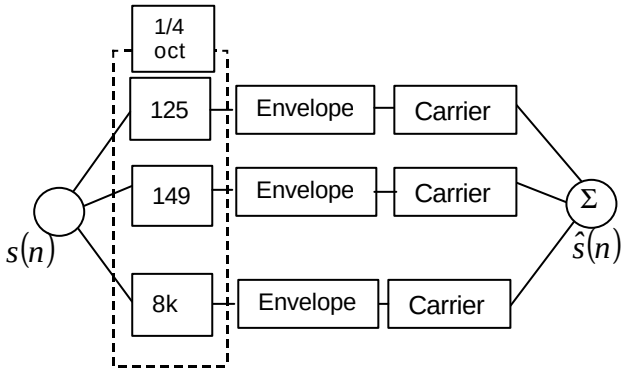

Fig. 3. A block diagram of envelope modulation modeling

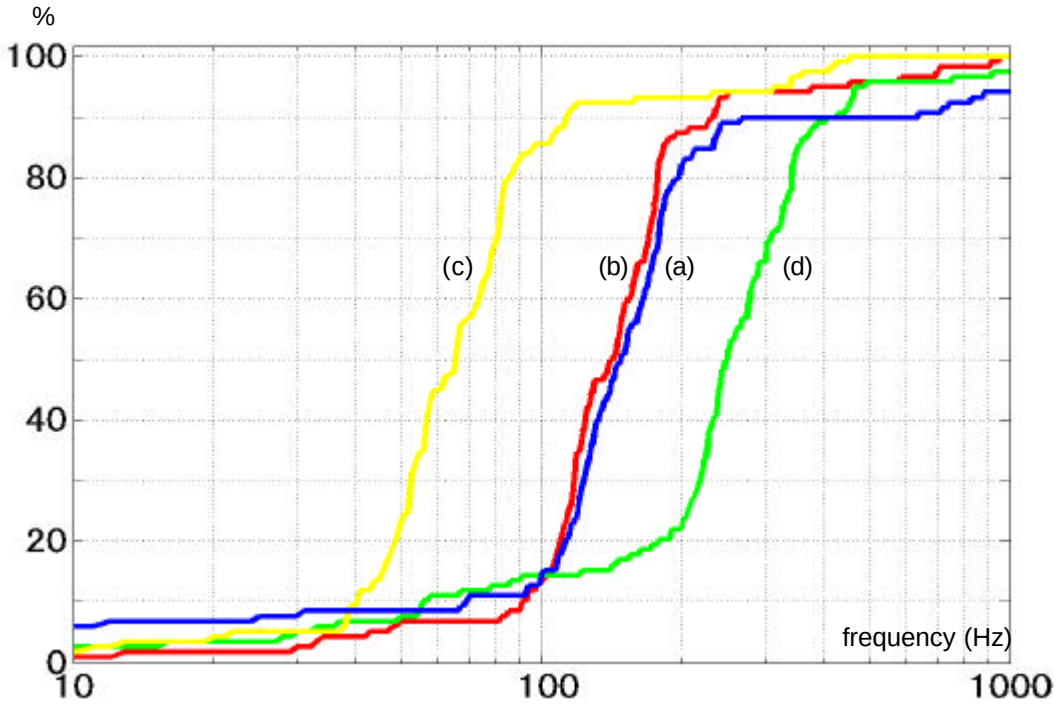

Fig. 4. Accumulated distributions of fundamental-frequency (FF) estimates for the original and synthesized speech signals.
(a) original, (b) synthesized, (c) 1/1-oct-lower shift, (d) 1/1-oct-higher shift

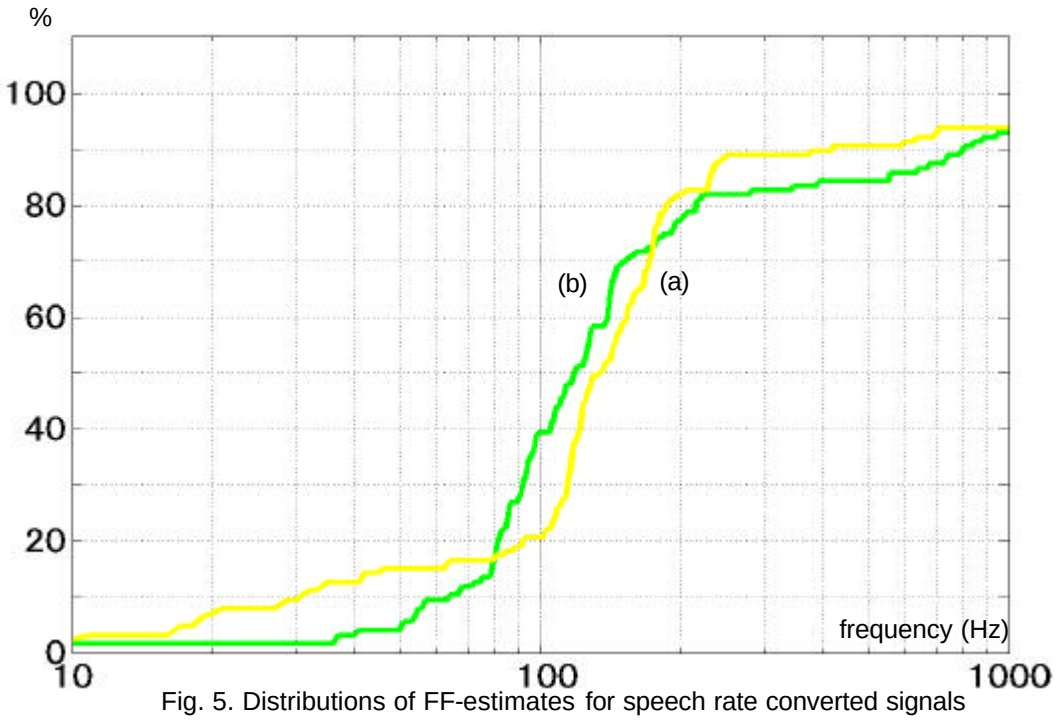

Fig. 5. Distributions of FF-estimates for speech rate converted signals
(a) 2-times-stretching, (b) ½-shrinking