SINUSOIDAL AND ENVELOPE MODULATION MODELING OF SIGNALS -A SIGNAL THEORETIC APPROACH TO ACOUSTIC EVENTS RENDERING-

Mikio Tohyama

Kogakuin University,
2665-1 Nakano-machi, Hachioji-shi, Tokyo 192-0015, Japan

ABSTRACT

This study investigates a signal theoretic approach to rendering acoustic events. Acoustic events modeling language (AEML) is an essential part of acoustic events rendering using an acoustic-data stream based on structured audio representation. This article mainly describes sinusoidal and envelope-modulation modeling for intelligible speech modification and reverberation signals rendering. The sinusoidal modeling is useful for constructing intelligible speech using only a few dominant components, and narrow-band envelopes such as 1/4-octave-band-speech envelopes are the key to representation of speech intelligibility. Envelope modulation modeling using the dominant sinusoidal carriers enables modification of the talker's pitch and speech-rate without sacrificing intelligibility. The narrow-band envelope can be estimated by clustered line-spectrum modeling (CLSM) based on the LSE-criterion in the frequency domain. Sinusoidal modeling with a decaying envelope is also a key technology for reverberation sound rendering based on the modal statistics of a reverberation field.

1. INTRODUCTION

An immersive communication system requires the rendering of acoustic events to enable effective collaboration through an interactive sound field network (ISFN)[1-3]. The acoustic events modeling language (AEML) is essential to acoustic events rendering using an acoustic data stream based on structured audio representation[4]. Figure 1 shows a tentative construction for AEML that describes a set of procedures for audio scene description and functions for modeling.

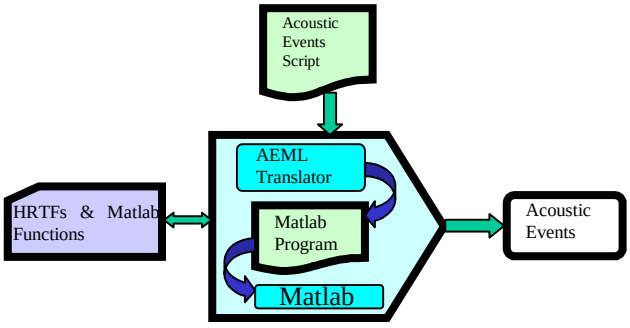


Fig. 1. The construction of AEML.

This article will take a signal-theoretic approach to rendering acoustic events and describe sinusoidal and envelope-modulation modeling for intelligible speech modification and reverberation sounds rendering. The author will demonstrate that the talker's pitch and speech-rate can be modified without sacrificing intelligibility and that a natural-sounding reverberation sound can be rendered.

Fourier analysis is a powerful tool for signal analysis from a mathematical point of view; however, it is not always suitable as a means of representing acoustic signals from the point of view of signal information analysis such as speech-intelligibility. Speech intelligibility is closely related to the envelopes of narrow-band speech signals, such as 1/4-octave-band-filtered speech-components[5-6]. Therefore signal-envelope modeling might provide especially important insights into speech intelligibility and contents analysis, and the superposition of envelope-modulated narrow-band signals can be used to generate an intelligible speech signal.

The envelope of a narrow-band signal component can be interpreted as the instantaneous magnitude of the narrow-band filtered signal[7]. Thus this instantaneous magnitude can be represented by a short-term Fourier magnitude spectrum (or by a long-term Fourier phase spectrum[8-9]). Clustered line-spectrum components whose frequencies are close to each other are the fundamental elements used to construct the envelope.

The immersive audio technology recently developed for computer network communication requires that acoustic events modeling include 3-D spatial sound effects [1-4]. Reverberation is important for 3-D effects creation as well as for sound-image localization. Artificial reverberators are quite effective and convenient for adding reverberation sounds to a direct sound [10] without known room-boundary conditions. However, the sound-field parameters, such as the modal and reverberation densities, cannot be easily adjusted, so conventional reverberators are not flexible enough to be used as 3-D reverberation-rendering tools.

This article will investigate intelligible speech modification and reverberation sounds rendering based on sinusoidal and envelope-modulation modeling. First, it examines how well narrow-band envelopes(or short-term magnitude spectrum) represent an intelligible speech signal. Second, it shows how the superposition of envelope-modulated narrow-band signals whose carriers

are replaced by the dominant sinusoidal signals can be used to modify the talker's pitch and speech-rate without sacrificing intelligibility [11][7-8]. Third, it explains the formulation of the clustered line-spectrum modeling (CLSM) based on the least-squares-error (LSE) criterion in the frequency plane, instead of in the time domain where conventional sinusoidal modeling is applied[12], to represent the narrow-band signal envelope. After that, it demonstrates that natural-sounding reverberation sound can be rendered based on the superposition of the decaying envelope-modulation modeling of sinusoidal carriers whose frequencies are determined based on the reverberation sound field statistics.

2. MODIFICATION OF SPEECH SIGNAL USING ENVELOPE MODULATION MODELING

2.1. Envelopes for Intelligible Speech

Speech intelligibility is closely related to narrow-band speech signal envelopes- not to the entire signal envelope. Figure 2a shows a speech signal waveform. Figure 2b shows a waveform obtained by modulating a wide-band binary noise by the entire signal envelope, while Fig.2c is the waveform after superposition of 1/4-octave-band signals modulated by the corresponding 1/4-octave-band envelope. We can understand the speech sentence by listening to the Fig.2c waveform even though no verbal information is obtained from the Fig.2b signal.

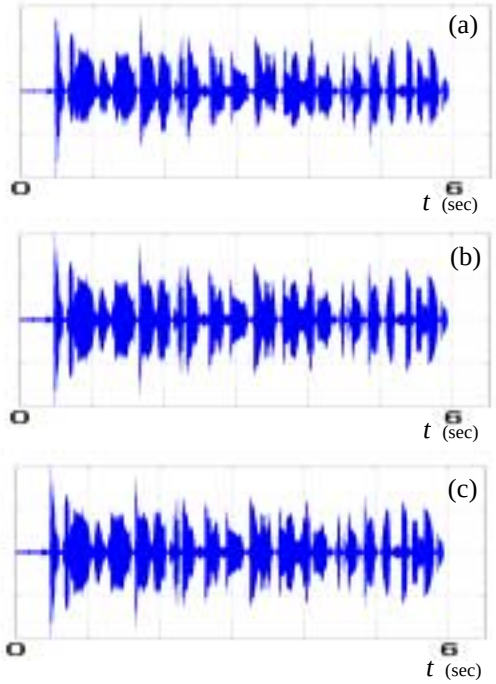


Fig. 2. Speech signal representation using binary noise ((a) original, (b) synthesized by a wide band binary noise modulated by the entire signal envelope, (c) synthesized by superposition of 1/4 oct. binary noise modulated by the corresponding 1/4 oct. band envelope

2.2.Envelopes and Short-Term Power Spectrum

The envelope of a narrow-band component can be interpreted as the temporal change of the magnitude spectrum for the narrow-band components. The significance of the magnitude(or the phase) spectrum for speech intelligibility depends on the frame length used for the spectrum analysis [9][13-14]. Figure 3 shows a block diagram of a method for deriving two types of hybrid signal (types A and B) by a cross-wise combination of the magnitude and phase spectra. Primary signal 1 is from male speech and signal 2 is from female speech.

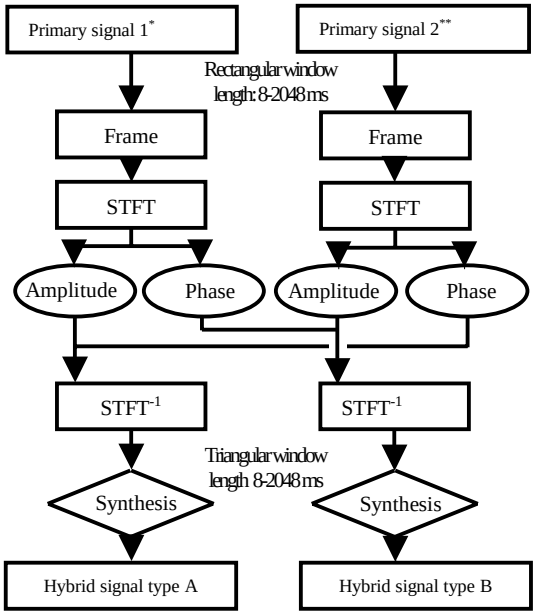


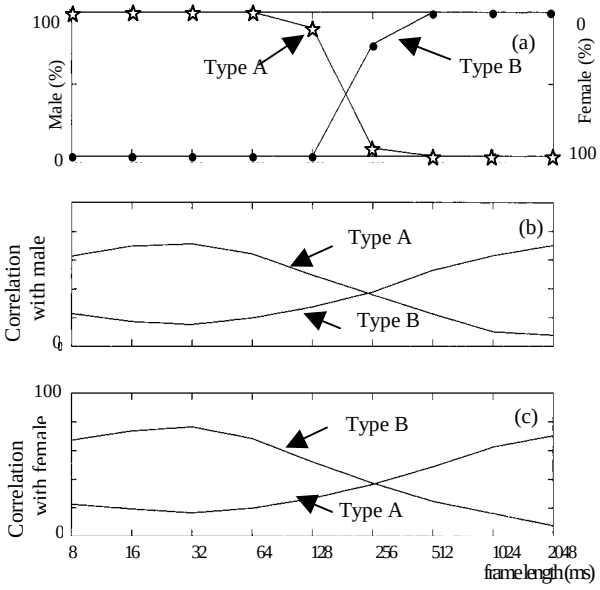
Fig. 3. Block diagram illustrating the method for deriving two types of hybrid signals from two primary signals, by a cross-wise combination of the amplitude and phase spectra in the STFT overlap-add procedure
*: Original female speech for experiment 2
**: Original noise for experiment 2

Figure 4(a) shows the hearing results from six subjects for six sentences spoken by three male and three female speakers. "Male 100%" ("Female 100%") on the left (right) vertical axis indicates that the subjects judged all the sentences to be spoken by a male (female). For 8- to 128-ms-long frames, the type-A signals preserved the male speech information better than the female. The opposite was true for the type-B signals. For 256- to 2048-ms-long frames the relationships were reversed – type-B signals better preserved male speech information and type-A signals better preserved the female. Thus, for the 8- to 128-ms-long frames, the magnitude spectrum was the most significant, whereas for the 256- to 2048-ms-long frames the phase spectrum was the most significant.

Figures 4(b) and (c) show the correlation between the narrow-band envelopes of the type-A or -B hybrid signals and the primary signals for the male and female speech, respectively. For short frame lengths, the synthesized

hybrid signal was highly correlated with the original speech whose magnitude spectrum was used in the reconstruction. In contrast, for long frame lengths, the correlation was highest between the hybrid signal and the original speech whose phase spectrum was preserved in the reconstruction. The two correlation curves cross at a frame length of approximately 256 ms.

The hearing tests (Fig. 4a) and the correlation analysis of the narrow-band envelopes (Figs. 4b and c) show the same trend with respect to the dominance of either the male or the female speech in the hybrid signals, and a similar frame-length effect. This suggests that the speech-like character is closely related to the degree of preservation of the original (narrow-band) envelopes. The critical frame length for determining the significance of either the magnitude or the phase spectrum is about 256 ms.



Type A: synthesized by using male-speech amplitude and female-speech phase
Type B: synthesized by using male-speech phase and female-speech amplitude

Fig.4. . Listening test and envelope correlation analysis for synthesized signal from male and female speech
((a) percentage of decisions stating male or female dominance, (b) averaged narrow-band envelope correlation between synthesized and original male speech, (c) averaged narrow-band envelope correlation between synthesized and original female speech)

2.3. Speech Representation by Envelope Modulation

As we explained in the previous section, an intelligible speech signal can be synthesized from short-term amplitude spectrum records, and the results of the envelope-correlation analysis indicate that the narrow-band temporal envelopes convey significant information that can be used to reproduce intelligible speech. 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 the talker's pitch and speech-rate without losing intelligibility or sacrificing voice quality.

Figure 5 shows a basic example of envelope-modulation modeling of speech signals. A speech signal can be synthesized as

$$s(n) = \sum_k e_k(n)\cos\phi_k(n), \qquad (1)$$

where $e_k(n)$ and $\phi_k(n)$ denote, respectively, the envelope and the instantaneous phase in the *k*-th frequency band. If we can substitute simple carriers, such as sinusoidal signals, for the narrow-band carriers, we can easily modify a speech signal.

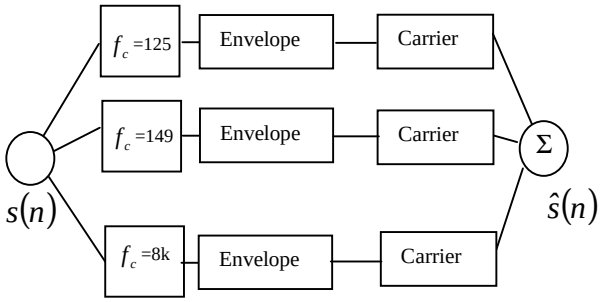


Fig. 5. The block diagram of envelope modulation modeling (f_c : center frequency of each band)

Figure 6 shows samples of instantaneous phase differences between synthesized and original speech after subtracting central-frequency phase-components, where the carriers were replaced frame-by-frame by sinusoidal signals estimated from the greatest magnitude spectrum components in each frequency band. The instantaneous phase can be approximated fairly well and by listening we can confirm that almost perfectly intelligible speech can be reconstructed by listening.

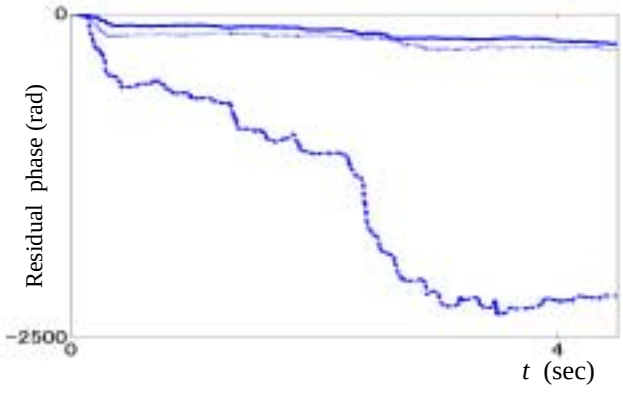


Fig. 6. Instantaneous phase difference between synthesized and original speech signal
(center frequency is ——250 Hz, - - - - 500 Hz, —·—· 4000 Hz)

2.4. Pitch and Speech-Rate Modification

The combination of envelopes and carriers in narrow bands is important for speech-signal control. The author expected the carrier to affect the talker's voice pitch as heard by a listener and tested whether this was so by multiplying the carrier frequencies by a constant factor. Figure 7 illustrates a tentative scheme for this process.

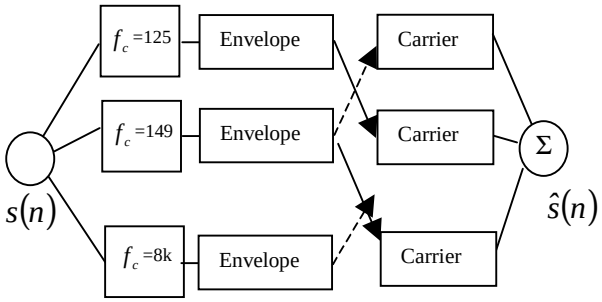
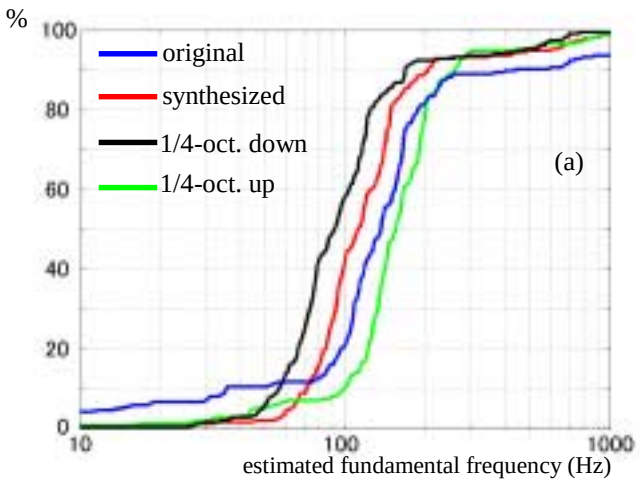


Fig. 7. Schematic representation of carrier frequency conversion

Figure 8 shows the accumulated distributions of fundamental frequencies estimated frame-by-frame for original speech, speech synthesized through envelope-modulation modeling without carrier modification, and speech modified by carrier-frequency manipulations for a 1/4-octave shift higher or a lower, and similarly for a 1/2 octave shift higher or lower. The fluctuation of short term fundamental frequencies might be a crucial aspect of reproducing a natural-sounding speech sentence. Similar distributions were observed around the expected fundamental frequencies as was perfectly intelligible modified speech with pitch conversion. Lowering the pitch might be an effective means of assisting listeners with severe hearing loss [8]; however, the talker's voice quality in the lowered-pitch sample changed from the original female voice to a more masculine voice.



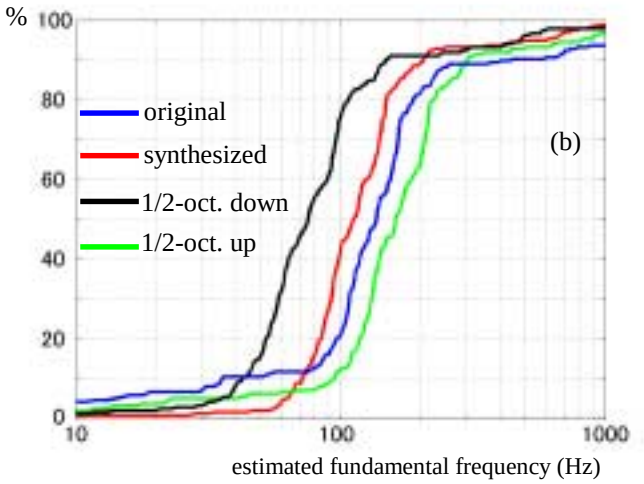


Fig. 8. Accumulated distributions of estimated fundamental frequencies for pitch-modified speech signals
((a) 1/4-oct. up or down, (b) 1/2-oct. up or down)

Using sinusoidal carriers to replace the narrow band carriers is a convenient means of pitch modification and can be used in envelope modification; e.g., in stretching and shrinking for speech rate control. Figure 9 shows examples of the fundamental-frequency distribution when slowing down to half the original speed and speeding up to twice the original speed. These results confirmed that intelligible speech with a modified speech rate could be synthesized while preserving the talker's voice characteristics and the pitch sensation.

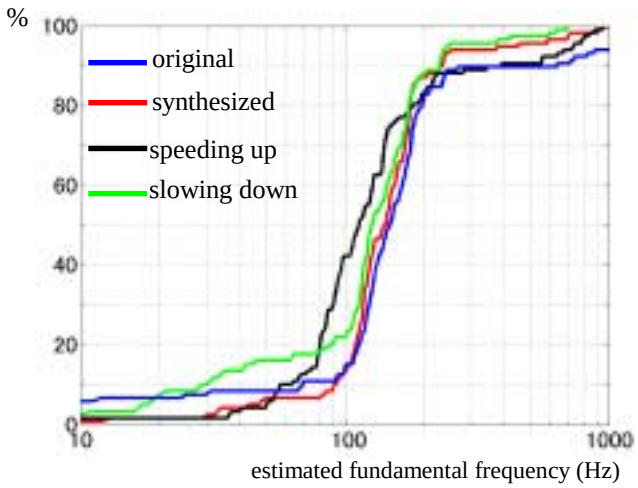


Fig. 9. Distributions for the estimated fundamental frequencies for speech-rate modified signals

Envelope modulation modeling is thus an effective tool for intelligible speech signal representation. It was possible to modify the talker's pitch with no need to change envelopes, and to convert the speech-rate by stretching or shrinking the narrow-band envelopes without having to modify the carriers.

3. SIGNAL REPRESENTATION BY CLUSTERED LINE-SPECTRUM MODELING

The envelope for a narrow band speech component is significant in intelligible speech representation. Therefore, the author has developed a technique, -clustered line-spectrum modeling (CLSM)- that can be used, to estimate the true signal components from clustered spectrum records. CLSM makes it possible to describe signals that include 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 in the time region where conventional sinusoidal modeling is performed[12].

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 also that the target signal is expressed in an analytic form as

$$x_a(n) \equiv \sum_{k=1}^{K} A_k e^{j2\pi f_k n} + \varepsilon_k(n),$$

in the narrow frequency-band 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, K is the number of dominant sinusoidal components, and $\varepsilon_K(n)$ denotes the residual component including the modeling error and external noise. 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 $k = k_{p-l+L-1}$ $k = k_{p-l+L-1}$ as

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

where

$$\begin{pmatrix} X(k_{p-l}) \\ \vdots \\ X(k_{p-l+L-1}) \end{pmatrix} \equiv \mathbf{x}_{observe}$$

and

$$\begin{pmatrix} X_S(k_{p-m}) \\ \vdots \\ X_S(k_{p-m+P-1}) \end{pmatrix} \equiv \mathbf{x}_{signal}$$

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

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

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

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

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

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

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. 10. Figure 10a 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. 10b. Figure 10c shows the synthesized waveform obtained by applying CLSM where L = 7 observations for P = 5 clustered signal components centered at the dominant peak. Figure 10d shows the residual signal and Fig. 10e shows the residual power spectrum. Thus, CLSM is clearly an effective means of signal representation that includes a smooth envelope.

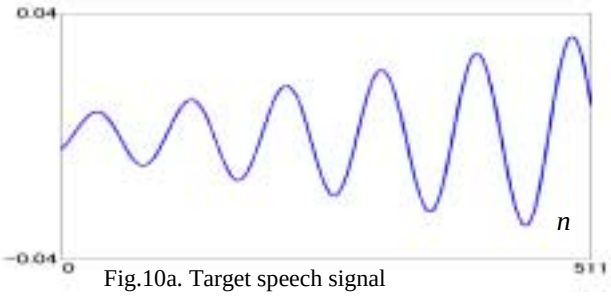


Fig.10a. Target speech signal
(1/4-oct. band, center frequency is 500 Hz.)

(dB)

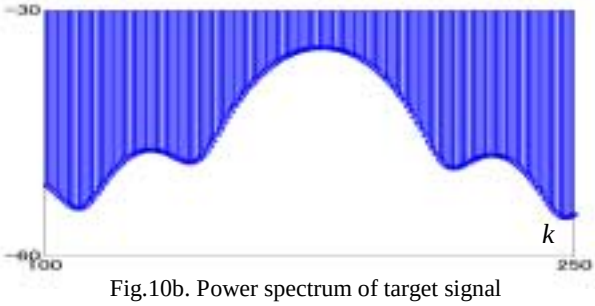


Fig.10b. Power spectrum of target signal
(DFT length is 16384)

Fig.10c. Synthesized signal by applying CLSM where L =7 observations for P =5

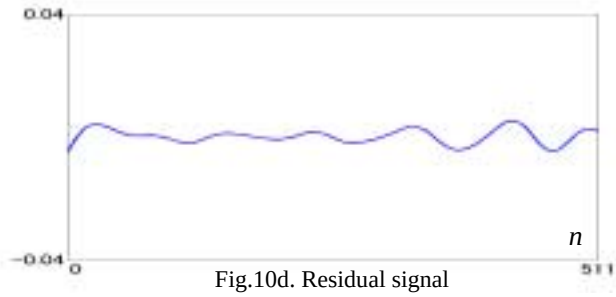


Fig.10d. Residual signal

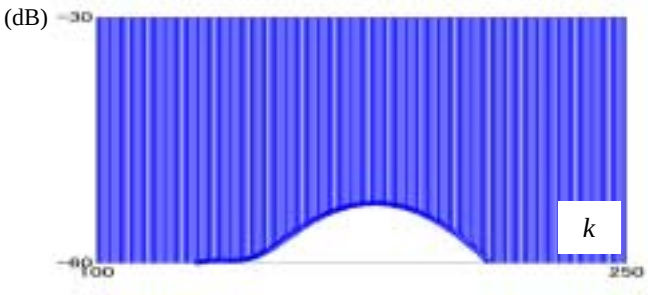


Fig.10e. Power spectrum of residual signal

Fig. 10. Signal reconstruction with a smooth envelope using CLSM

4. 3D-REVERBERATION RENDERING BASED ON RANDOM SOUND FIELD STATISTICS

Artificial reverberators are used in the immersive and structured audio technology applied for computer network communication that requires the modeling of acoustic events that include 3-D spatial sound effects[1-4]. Conventional reverberators, however, are not easily adapted to a room's acoustic parameters such as the reverberation-time characteristics, room volume, and modal density. Therefore, the author has been developing an algorithm for rendering binaural room impulse responses based on sinusoidal modeling with exponentially decaying envelopes [15].

The sinusoidal components needed to render the impulse response that represents the reverberation of a sound field can be determined from the transfer function statistics, including the modal-spacing distribution and angular distribution statistics for binaural reverberation sound. A collection of eigen frequencies, which are the frequencies of free vibrations, can be modeled as an occurrence of random events on a frequency scale[16-17].

The spacing statistics of the eigen frequencies (i.e., the modal spacing statistics) follow a Wigner distribution (Fig. 11) for an irregularly shaped room[16-18]. A Wigner distribution is written as

$$p(x) = 4xe^{-2x}\, x \geq 0\,,$$

where x denotes a normalized modal-spacing variable whose average is unity.

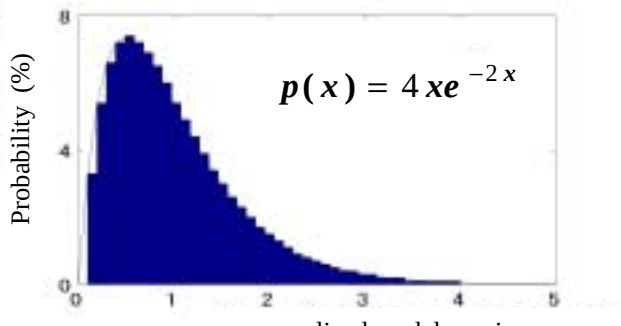


Fig. 11. Modal-frequency-spacing histogram

The average number of eigen frequencies (i.e., the modal density) is given by

$$n(f) \cong \frac{4\pi f^2}{c^3} V \equiv \frac{1}{d(f)}\,,$$

where V is room volume (m^3), f is frequency (Hz), c is sound speed (m/s), and $d(f)$ is average modal spacing (Hz). The normalized random variable x can be written using the modal spacing δf (Hz) as

$$x \equiv \delta f \,/\, d(f) \cong \frac{4\pi f^2}{c^3} V \delta f\,.$$

The modal density increases in proportion to the square of the frequency, as the frequency of interest is increased. This increase in modal density makes individual wave analysis difficult. The modal overlap is defined as

$$M(f) \equiv n(f) B(f)$$
$$\cong n(f)\frac{6.9}{2T_R} \cong 13.8\pi \frac{Vf^2}{c^3 T_R}$$

which shows the average number of modes simultaneously excited within the modal bandwidth, $B(f)$, of an individual modal response function. Here, T_R denotes the reverberation time in seconds, and the modal bandwidth is given by

$$B(f) \equiv \frac{\int |H(f)|^2\, df}{|H(f)|^2_{\max}} \cong \frac{6.9}{2T_R}(\text{Hz})$$

where $H(f)$ denotes the modal response function, which can be interpreted as a resonance response function for the modal frequency of interest. When the modal overlap is high, an individual modal pattern is no longer observed in a space because many modes with the same frequency are simultaneously excited. As a result, individual modal analysis is not effective when there is high modal overlap. Therefore, the statistical approach is attractive and a reasonable choice for sound-field analysis.

The rendering process is outlined in Fig. 12.

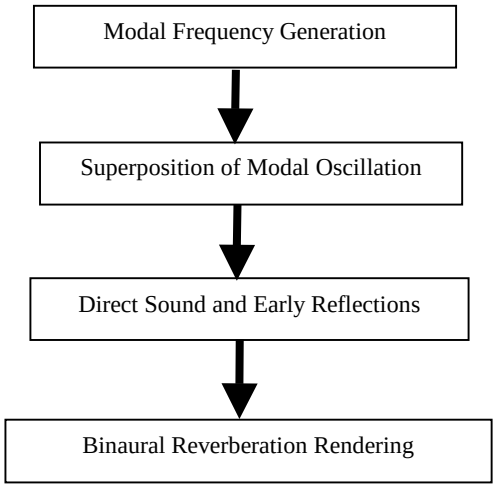


Fig. 12. Rendering process

First, we generate a series of modal frequencies following the modal spacing histogram shown in Fig. 11. Here, we have to take only the non-overlapping modes under high-modal-density conditions. Figure 13 illustrates a generated sequence of modal frequencies.

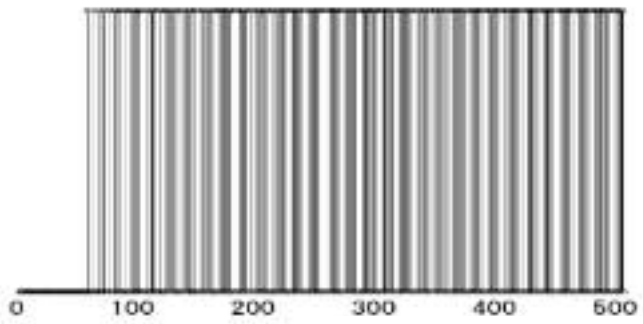


Fig.13. Generated modal-frequency-sequence

The impulse response is obtained through a summation of all modal responses expressed by decaying sinusoidal components with random magnitude and phase. Both direct sound and early reflections can be added to the generated impulse response. The levels, time gaps, and frequency characteristics of the early reflections are appropriately controlled. The impulse response including the early reflections is shown in Fig. 14a, and its magnitude frequency characteristics are shown in Fig. 14b.

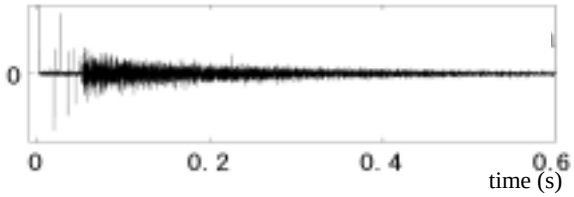


Fig. 14a. Impulse response.

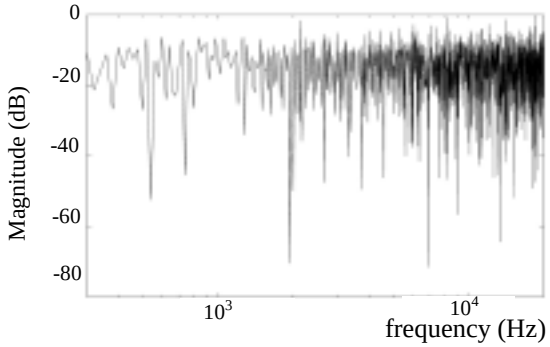


Fig. 14b. Frequency response of the impulse response

Fig. 14. Frequency response and the generated impulse response

A binaural response can be derived by introducing an angular distribution into the reverberation field statistics [19].

5. CONCLUSION

The author has described a signal theoretic approach to acoustic-signal modeling of acoustic events that renders a particular speech-signal and reverberation sound 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 that include reverberation sound have been demonstrated. In addition, reverberation sound rendering based on the random sound field statistics, clustered line-spectrum modeling (CLSM) for envelope representation, and a method of intelligible speech representation that uses narrow-band envelopes and their sinusoidal carriers were presented. The envelope-modulation method enables modification of the speaker's voice pitch and speech rate without sacrificing intelligibility. This article, which has mainly focused on the representation of signals from the viewpoint of envelope-modulation modeling, should contribute significantly to AEML development. The author is

extremely grateful to Prof. Tammo Houtgast of Amsterdam Free University, the Netherlands, for his valuable discussions, suggestions, and able guidance. This research has been partly supported by the Telecommunications Advancement Organization of Japan and the International Communications Foundation.

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

[6]R.V. Shannon, F.G. Zeng, and J. Wygonski, "Speech Recognition with Altered Spectral Distribution of Envelope Cues," *J. Acoust. Soc. Am.* 104 pp. 2467-2476 (1998)

[7]P.M. Clarkson and S.F. Bahgat, "Envelope Expansion Methods for Speech Enhancement," *J. Acoust. Soc. Am.* 89 (3) pp. 1378-1382 (1991)

[8]D.A. Vickers et al, *J. Acoust. Soc. Am.* 110, pp.1164-1175 (2001)

[9]M. R. Schroeder and H.W. Strube, "Flat-Spectrum Speech," *J. Acoust. Soc. Am.* 79 pp.1580-1583 (5) (1986)

[10]D. Rocchesso, "Maximally Diffusive Yet Efficient Feedback Delay Networks for Artificial Reverberation," *IEEE SP Letters* 4 pp. 252-255 (1997

[11]J. Laroche and M. Dolson, "New Phase-Vocorder Techniques for Real Time Pitch Shifting, Chorusing, Harmonizing, and Other Exotic Audio Modifications," *J. Audio. Eng. Soc.* 47, pp.928-936 (1999)

[12]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

[13]A.V. Oppenheim and J.S. Lim, "*The Importance of Phase in Signals,"* Proc. IEEE, 69 pp.529-541 (1981)

[14]M. Kazama, M. Toyama, and T. Houtgast, "Speech Reconstruction by Using Only Its Magnitude Spectrum Or Only Its Phase," 17th *Int. Congress on Acoustics* 7p. 51 (2001)

[15]M. Toyama, M. Kazama, and Y. Kamiya, "3-D Reverberation Sound Rendering Based on Distribution Statistics of Poles and Residues in Transfer Functions," 17th *Int. Congress on Acoustics* 4B.11.04

[16] R. H. Lyon, Statistical Analysis of Power Injection and Response in Structures and Rooms, J. Acoust Soc. Am, 45 pp. 545-565 (1969)

[17]M. R. Schroeder, J. Audio Eng. Soc. vol. 37 pp.795-808 (1989)

[18]Y.Fujisaka, Y. Takahashi, and M. Tohyama, Eigenmode Analysis and Degree of Freedom in Chaotic Semi-Stadium Sound Fields, Presented at IEEE ICASSP 2001 Audio-P2.9

[19] Y. Ando, "Concert Hall Acoustics," Springer Verlag, (1985)