Acoustic Events Modeling Language (for Immersive Communication Network)

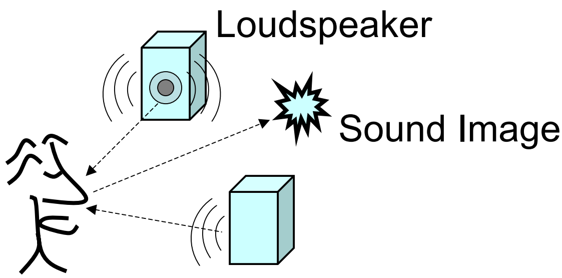
M. Tohyama

Dec. 4, 5. 2003

Acoustic Events Modeling: Wave Theoretic Approach

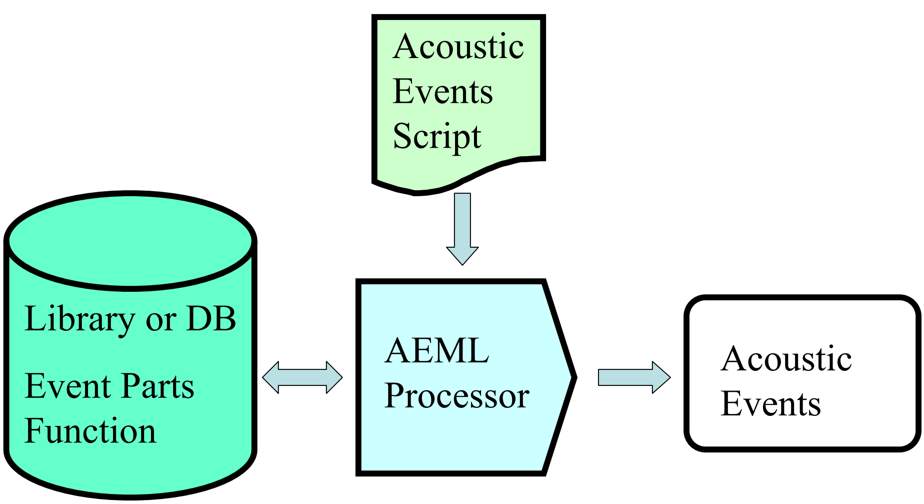
- Source Waveform Signature Analysis and Modeling
  - Speech Waveform Analysis according to the Envelope
  - Talker Information Extraction
  - Signature Analysis for Vocal Fold Motion
- Transmission Path Information Rendering
  - Reverberation Rendering based on the Transfer Function Statistics
  - Representation for HRTF for Moving Source Rendering
- Signal Perception Modeling
  - Precedence Effects and Signal Envelopes

What is an Acoustic Event?



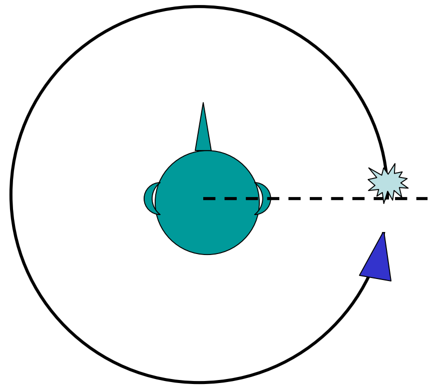
Acoustical phenomena
we can listen to and perceive
as a distinct sensation.

Acoustic Events Modeling Language

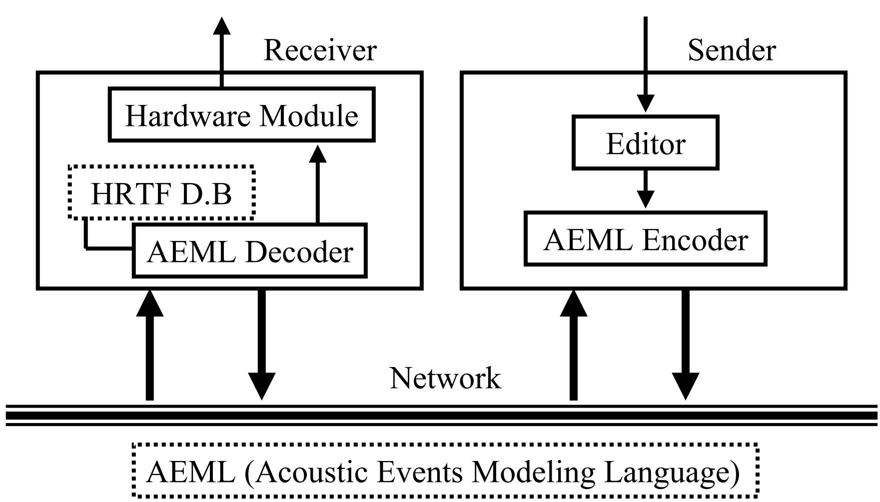


Example of Acoustic Events Script

```
s1=sound(test.wav)
ld=listener(dummy01)
sp=point(-1,0,0)
db=point(0,0,1)
lp=point(0,0,0)
ld:locate(lp)
s1:goround(0,12,0.5,lp,1,db,sp,0.5,off)
play(ld)
```



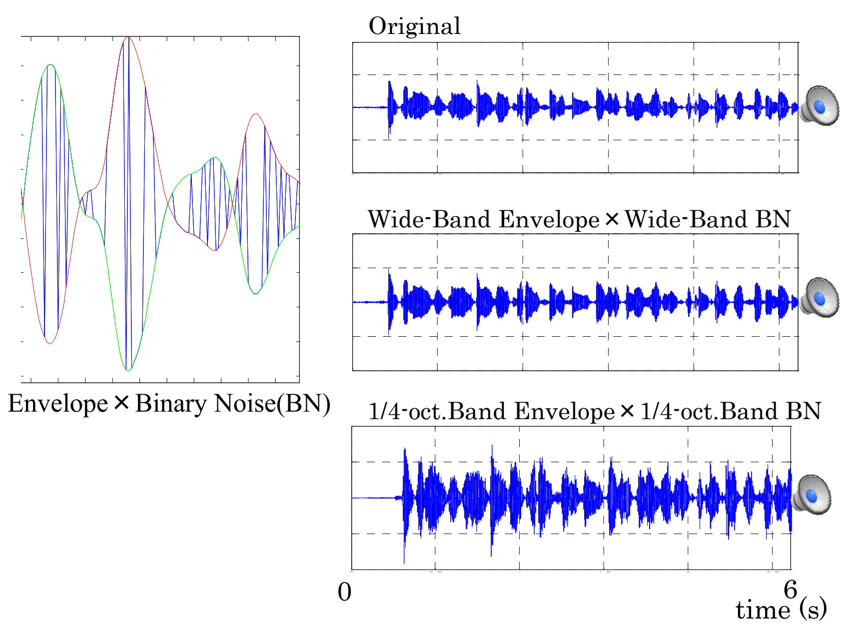
Interactive Sound Field Network



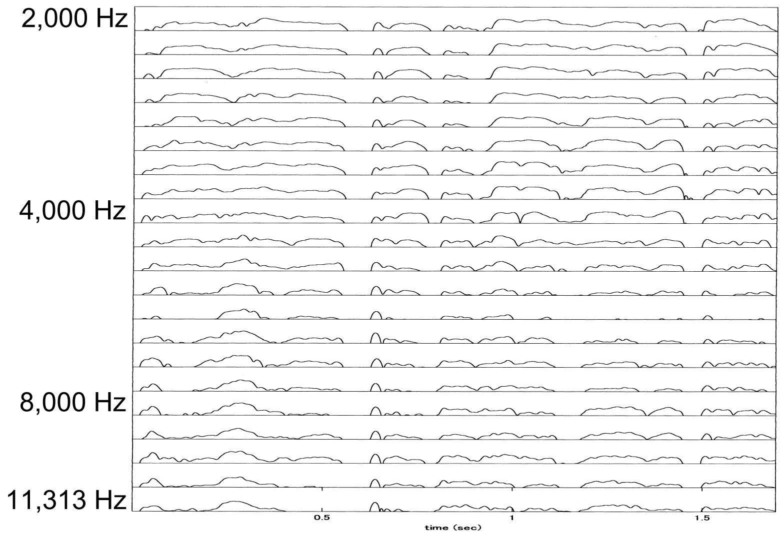
Source Waveform Signature Analysis using Signal Envelopes

- Talker Information Extraction using Signal Envelopes
  - Envelope Correlation Matrices for Higher Frequency Bands over 2kHz
- CLSM: Clustered Line-Spectrum Modeling for Envelope Analysis
  - Peak-picking and Least Square Solution in the Frequency Domain
  - Magnitude and Phase Information for Envelope Reconstruction
- Speech Source Waveform Analysis
  - CQ: Closed Quotient Analysis for Vocal Fold Motion
  - Higher Formant Analysis for Closed and Open Phase

Envelope Requires Frequency-Band Information

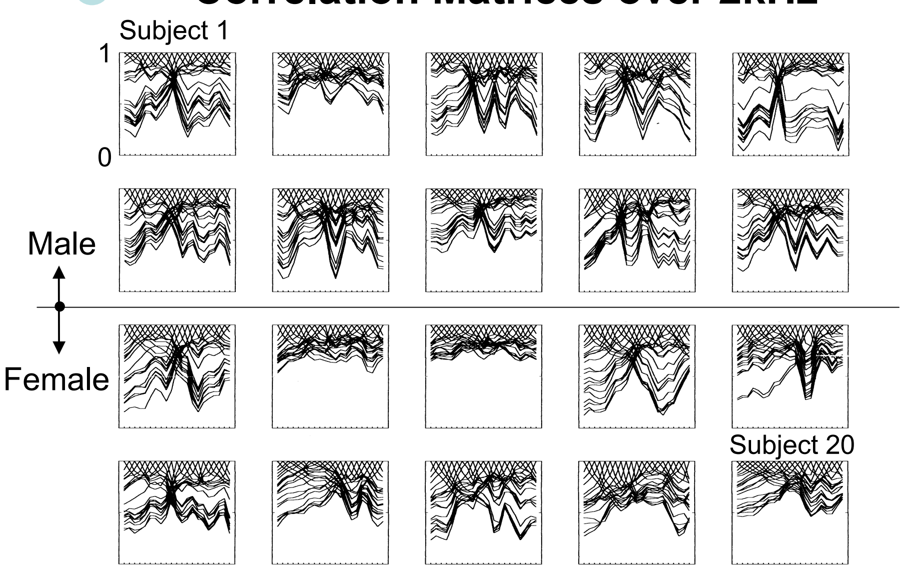


Example of narrow-band envelope

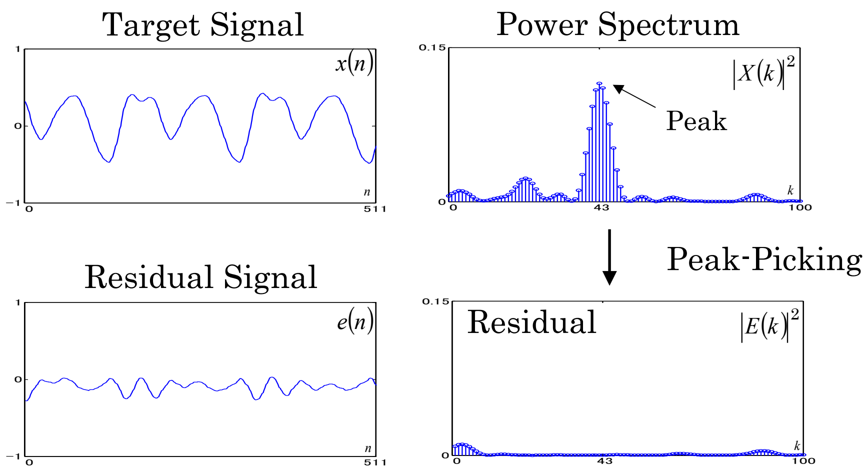


“Konojiwa tohkukara mienikui”

Correlation Matrices over 2kHz

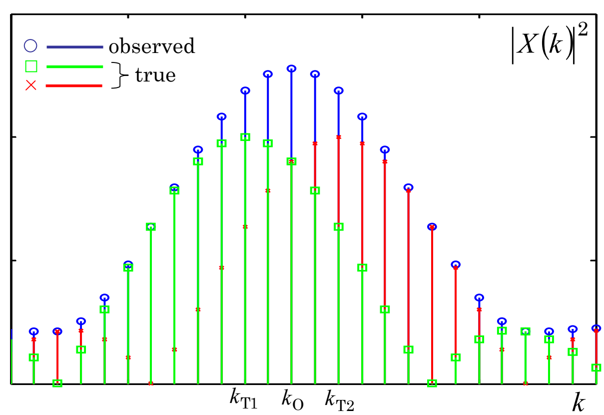


Signal Representation by Spectrum Peak-picking



Clustered Line Spectrum Modeling (CLSM)

An example of clustered two-components



Power spectrum

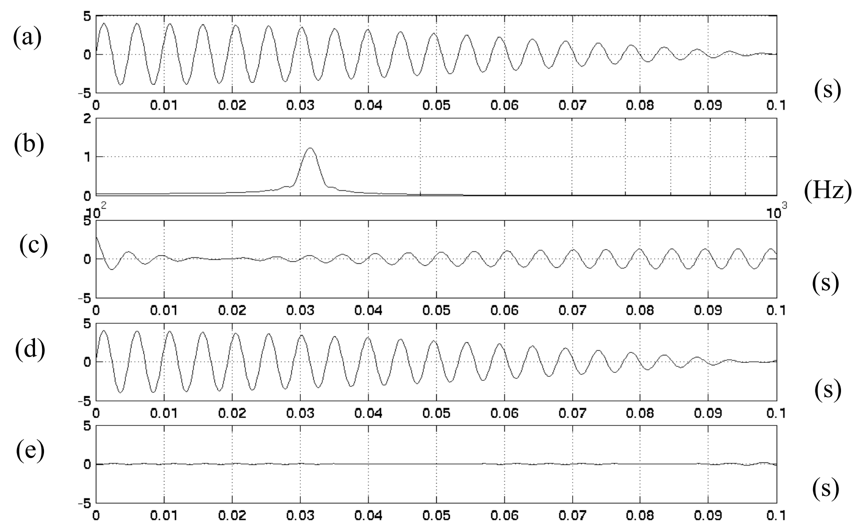
Quatieri (1990), Maher (1990)

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

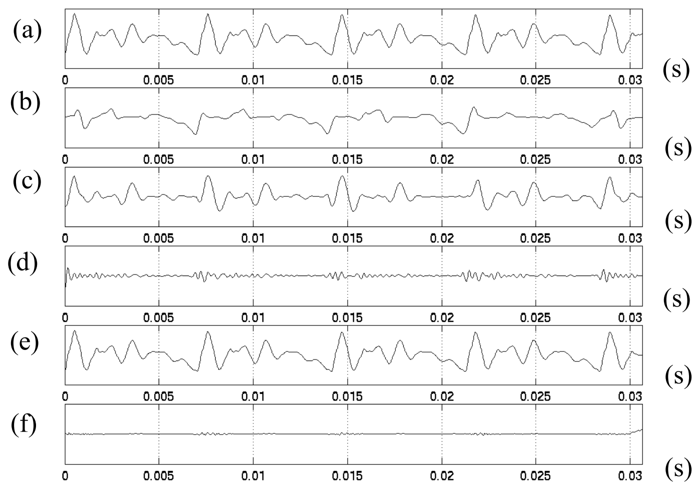
$$\underbrace{\begin{bmatrix} X_O(k_0) \\ X_O(k_1) \\ \vdots \\ X_O(k_{L-1}) \end{bmatrix}}_{L \text{ Observations}} = \underbrace{\begin{bmatrix} W(k_0-k_i) & W(k_0-k_{i+1}) & \cdots & W(k_0-k_{i+K-1}) \\ W(k_1-k_i) & \ddots & & \vdots \\ \vdots & & & \\ W(k_{L-1}-k_i) & \cdots & & W(k_{L-1}-k_{i+K-1}) \end{bmatrix}}_{\text{Window function}} \cdot \underbrace{\begin{bmatrix} \hat{X}_T(k_i) \\ \hat{X}_T(k_{i+1}) \\ \vdots \\ \hat{X}_T(k_{i+K-1}) \end{bmatrix}}_{K \text{ Assumed Sinusoidal components}}$$

$L > K$

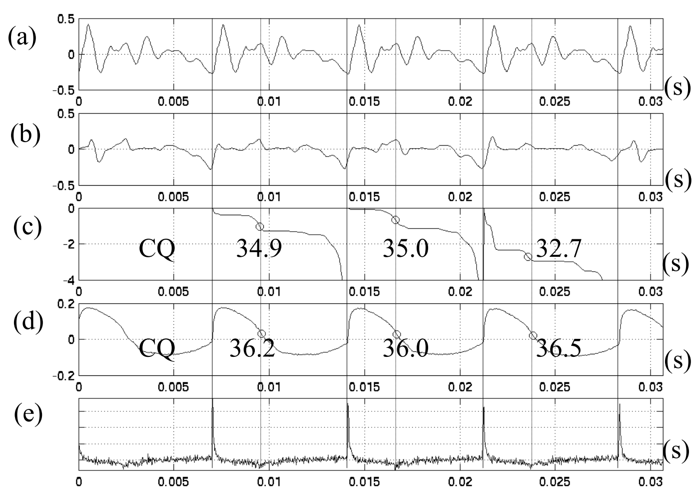
Linear equations in the frequency domain



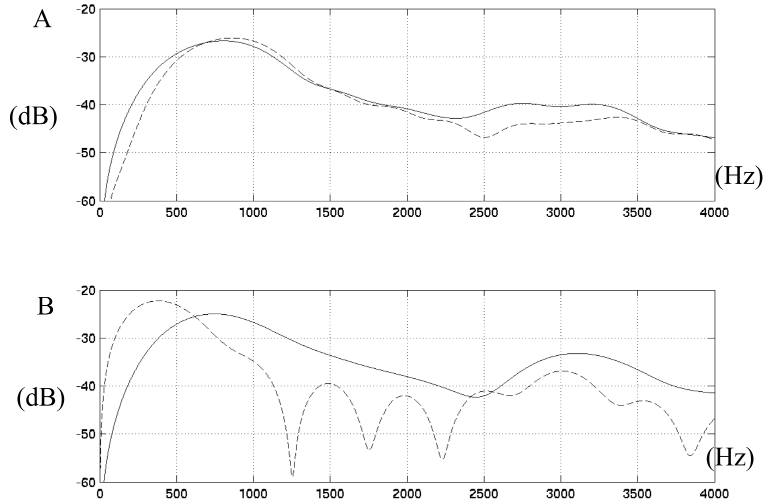
An example of CLSM, (a) original waveform by 5 sinusoidal components, (b) power spectrum, (c) residual after subtracting the peak spectrum component, (d) reconstructed waveform by 3 components based on CLSM (e) residual errors by (a)-(d)



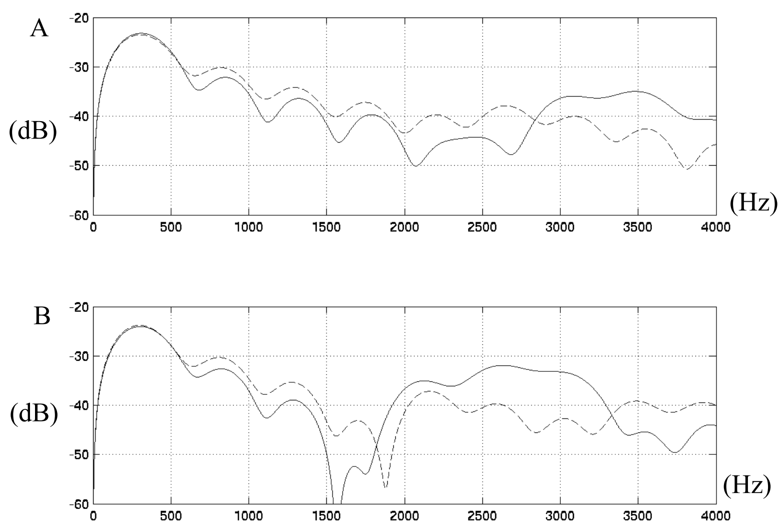
CLSM analysis using 3 sinusoidal components every 0.2 ms of frames, Speech [ah], (a) original waveform, (b)(c) (d) lower, middle, and higher components, (e) reconstructed waveform by CLSM, (f) residual components (a)-(e)



CQ estimation for the Speech [ah], (a) original waveform, (b) lower component corresponding, (c) energy decay curve of plot (b), (d) time-aligned simultaneously recorded Lx waveform, (e) differential of the Lx



Power spectrum with an exponential window for closed and open phase, solid line: closed-phase spectrum, broken line: open-phase spectrum, A: speech [ah], B: sung [ah]

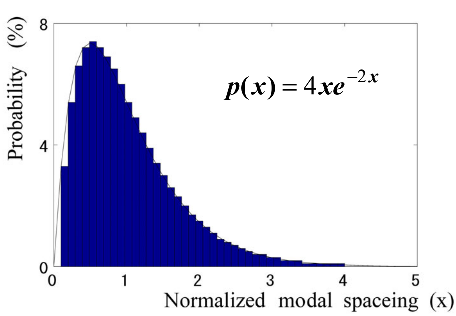


Power spectrum for A: speech [ee], B: [ee]

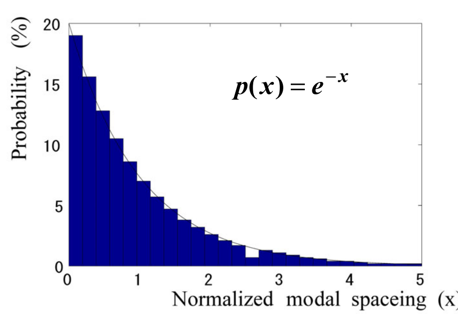
Transmission Path Information Rendering

- Reverberation Rendering based on the Transfer Function Statistics
- Distribution of Poles and Zeros and Transfer Function Statistics
- Room Acoustics Chaos and Sound Ray Propagation
- DFT Equation for Extrapolation of Reverberation Responses
- Representation of HRTF for Moving Source Rendering
- Convolution for Time Invariant Systems
- Interpolation of HRTF for Moving Source rendering

Distribution of Poles Modal Spacing Statistics



(a) Irregular room



(b) Rectangular room

Rendering Process

Wigner Distribution — Modal Frequency Generation

↓

Random Variables — Superposition of Modal Oscillation

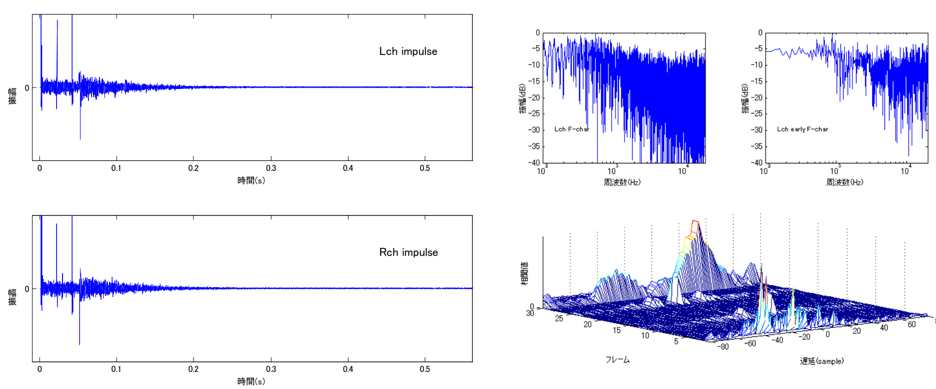
↓

Freq. characteristics — Direct Sound and Early Reflections

↓

Angular Distribution — Binaural Reverberation Rendering

Reverberation Example



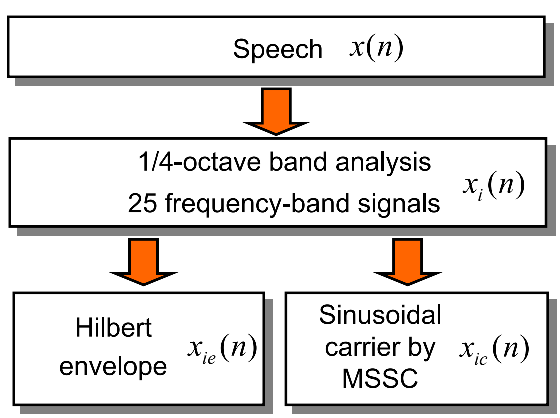
Dog: →

Signal Perception Modeling

- Precedence Effects and Signal Envelopes
  - Precedence Effect on Speech Signal
  - Speech Signal Representation using Envelopes and Carriers
  - Envelope vs Carrier Delays
- Selective Listening
  - Dereverberation and Phase Equalizatio
  - Blind Source Separation

Procedure

- The precedence effect speech signal
  - Characteristics of speech signal → Envelope and carrier



Amplitude modulation $\hat{x}_i = x_{ie}(n) \times x_{ic}(n)$

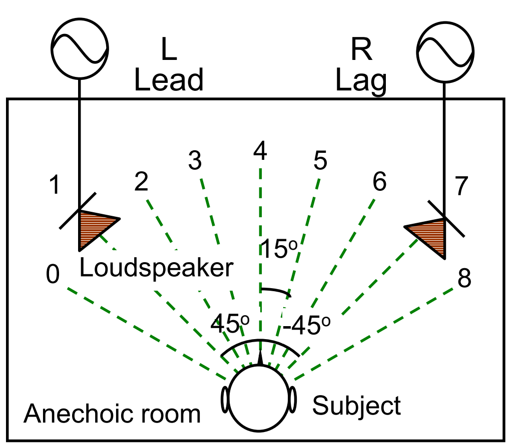
Full-band signal $\hat{x}(n)$

- Signal carrier

  MSSC: most significant sinusoidal carrier $x_{ic}(n)$

  - MSSC is a signal with the greatest magnitude in the frequency band
  - The spectrum was estimated frame by frame

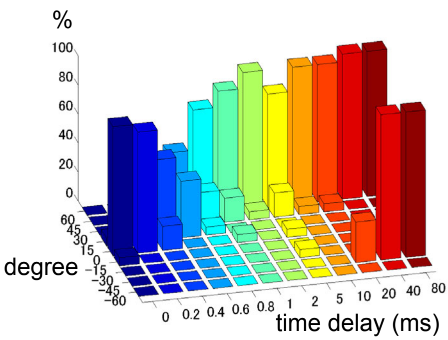
Experimental setup



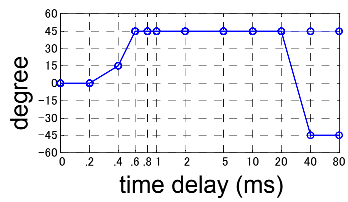
- Subjects indicate the direction of sound image by choosing corresponding numbers from 0 to 8

Results

Original, E+C

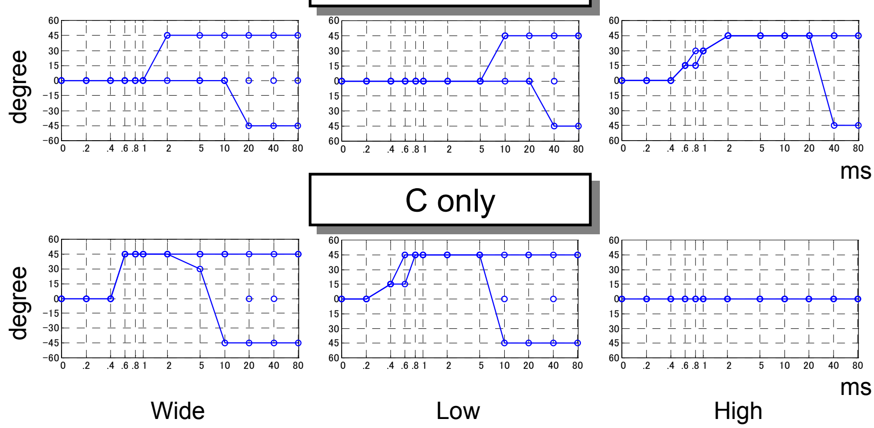


- The results for “original” and “E+C” are almost same.



E only

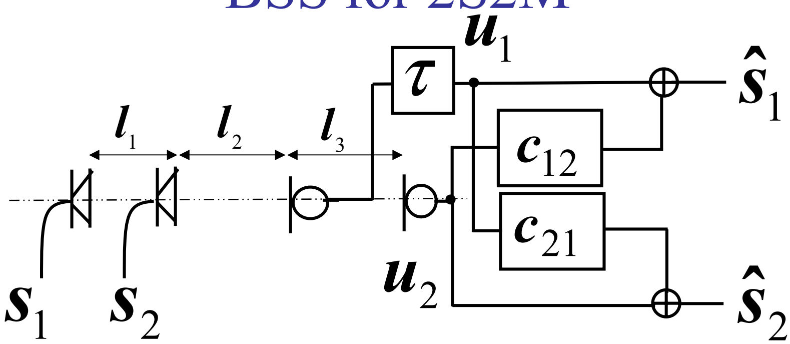
C only



Envelope vs Carrier

- Envelope delay → Sound-image separation
  Carrier delay → Sound-image localization and separation
- Low-frequency range → Carrier is a significant cue
  High-frequency range → Envelope is a significant cue
- Sound localization is dominated by low-frequency components of the signals

BSS for 2S2M



$$\begin{pmatrix} \hat{\boldsymbol{s}}_1 \\ \hat{\boldsymbol{s}}_2 \end{pmatrix} = \begin{pmatrix} 1 & c_{12} \\ c_{21} & 1 \end{pmatrix} \begin{pmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{u}_1 + c_{12}\boldsymbol{u}_2 \\ \boldsymbol{u}_2 + \boldsymbol{c}_{21}\boldsymbol{u}_1 \end{pmatrix}$$

unknown: $\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{l}_1, \boldsymbol{l}_2, \boldsymbol{l}_3$ $\begin{pmatrix} 1 & c_{12} \\ c_{21} & 1 \end{pmatrix}$: separation matrix

estimate: $\boldsymbol{c}_{12}, \boldsymbol{c}_{21}$

Separation Matrix

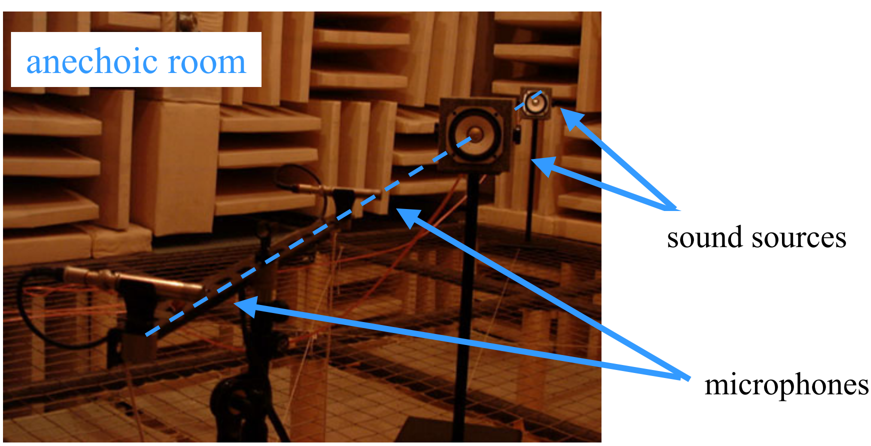
$$\hat{s}_1 \cdot \hat{s}_2 = (u_1 + c_{12}u_2) \cdot (u_2 + c_{21}u_1) = 0$$

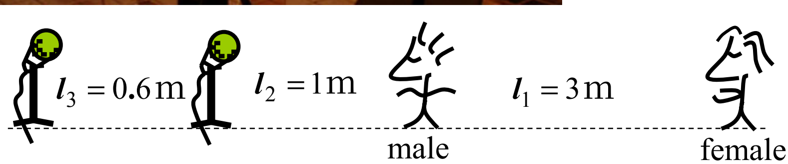
$$c_{12} \,|\, u_2 \,|^2 + c_{21} \,|\, u_1 \,|^2 + (1 + c_{12}c_{21})\, u_1 \cdot u_2 = 0$$

$$\begin{cases} \boldsymbol{u}_1^{\mathrm{n}} \cdot \boldsymbol{u}_2^{\mathrm{n}} (c_{12}c_{21} + 1) + |\, \boldsymbol{u}_1^{\mathrm{n}} \,|^2 \, c_{21} + |\, \boldsymbol{u}_2^{\mathrm{n}} \,|^2 \, c_{12} = 0 \\ \boldsymbol{u}_1^{\mathrm{n\text{-}1}} \cdot \boldsymbol{u}_2^{\mathrm{n\text{-}1}} (c_{12}c_{21} + 1) + |\, \boldsymbol{u}_1^{\mathrm{n\text{-}1}} \,|^2 \, c_{21} + |\, \boldsymbol{u}_2^{\mathrm{n\text{-}1}} \,|^2 \, c_{12} = 0 \end{cases}$$

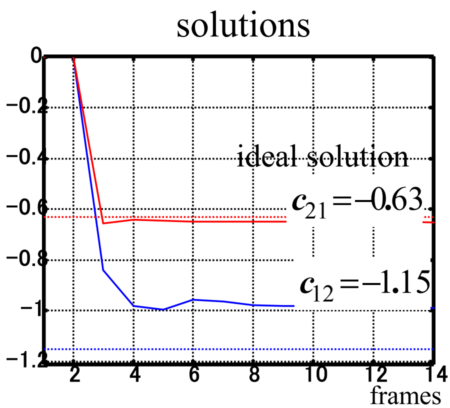
Two unknown parameters: c_{12} and c_{21}

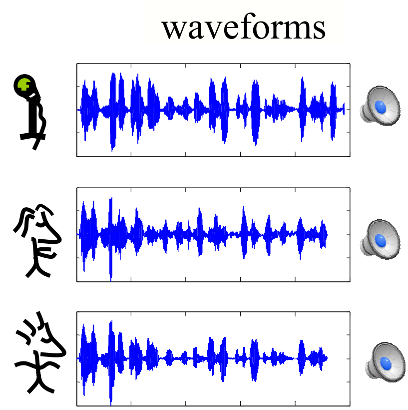
Anechoic-room Experiment





Examples





Summary
Acoustic Events Modeling

- Signal Theoretic Approach
  - —Signal Representation
  - —Linear Equation
- Wave Theoretic Approach
  - —Statistical TF Approach
- Perceptual Approach
  - —Signal Signature Analysis
  - —Binaural Modeling

Next Step: Authentic Approach

● COE Wave Communication Group

—Speech Production Process Modeling

—Spatial Sound Field Synthesizer

—Acoustic Systems and Devices