Acoustic Sensing With Artificial Intelligence

Bowon Lee

Department of Electronic Engineering
Inha University
Incheon, South Korea
bowon.lee@inha.ac.kr
bowon.lee@ieee.org

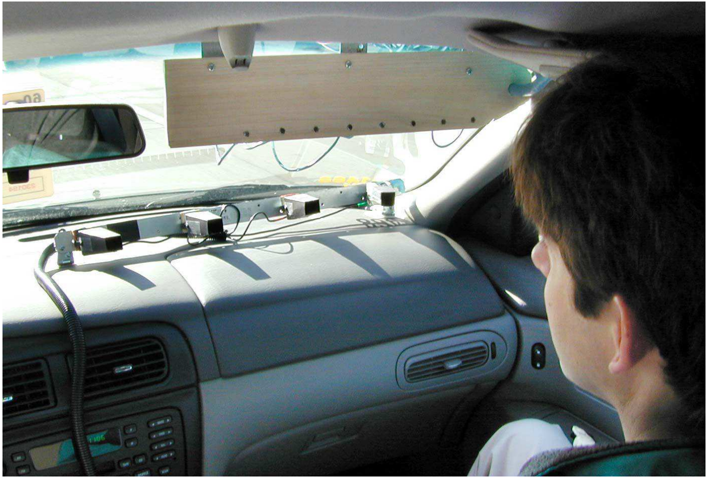
NVIDIA Deep Learning Day
Seoul, South Korea
October 31, 2017

Outline

Telephone (1876)

Home Assistants (Current)

AVICAR Project at UIUC (2002 – 2006)



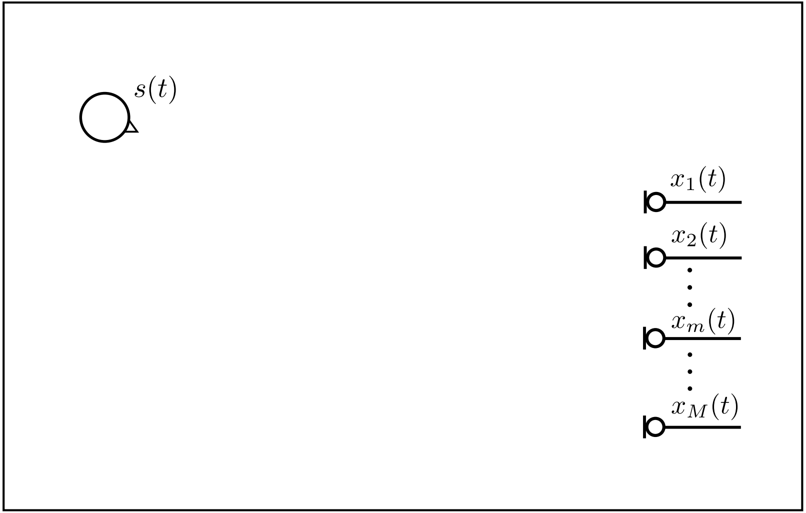
AVICAR Video

ConnectUs Project at HP Labs (2008 – 2011)

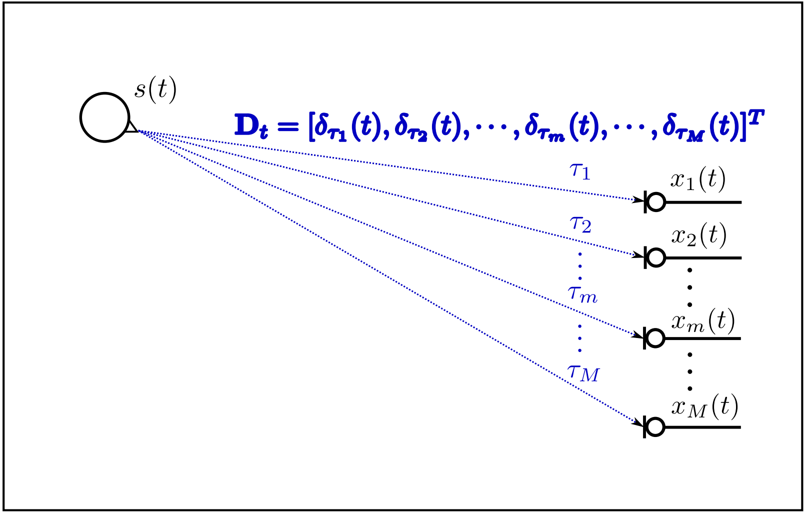
ConnectUs Video

Outline

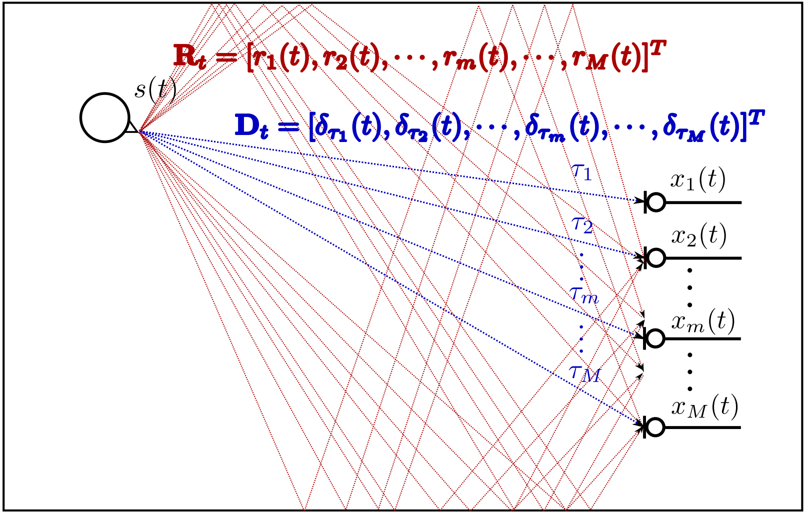
Acoustic Sensing With a Microphone Array



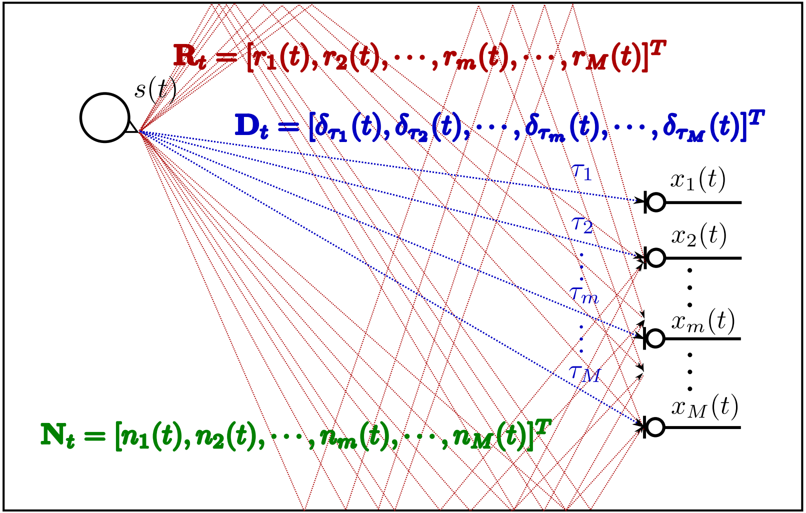
Acoustic Sensing With a Microphone Array



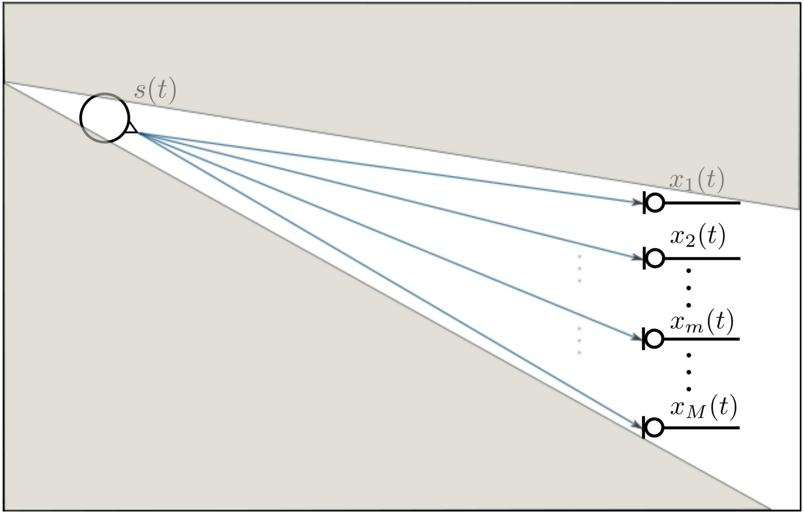
Acoustic Sensing With a Microphone Array



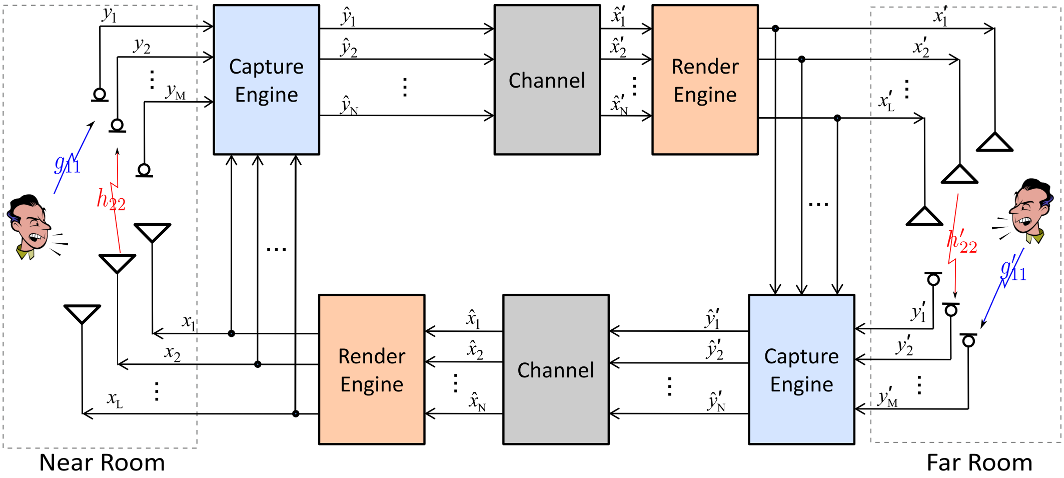
Acoustic Sensing With a Microphone Array



Acoustic Sensing With a Microphone Array



Audio Communication



Signal Processing for Audio Communication

Output signal of the capture engine

$$\hat{y}_n(t) = \underbrace{\sum_{m=1}^{M} \{f_{mn} * y_m\}(t)}_{\text{microphone array processing}} + \underbrace{\sum_{l=1}^{L} \{q_{ln} * x_l\}(t)}_{\text{echo cancellation}}$$

Topics of Acoustic Signal Processing

- Beamforming
- Blind Source Separation
- Sound Source Localization
- Multichannel Echo Cancellation
- Dereverberation, Noise Suppression, Speaker Diarization

Signal Processing for Audio Communication

Output signal of the capture engine

$$\hat{y}_n(t) = \underbrace{\sum_{m=1}^{M} \{f_{mn} * y_m\}(t)}_{\text{microphone array processing}} + \underbrace{\sum_{l=1}^{L} \{q_{ln} * x_l\}(t)}_{\text{echo cancellation}}$$

Topics of Acoustic Signal Processing

- Beamforming
- **Blind Source Separation**
- Sound Source Localization
- Multichannel Echo Cancellation
- Dereverberation, Noise Suppression, Speaker Diarization

Signal Processing for Audio Communication

Output signal of the capture engine

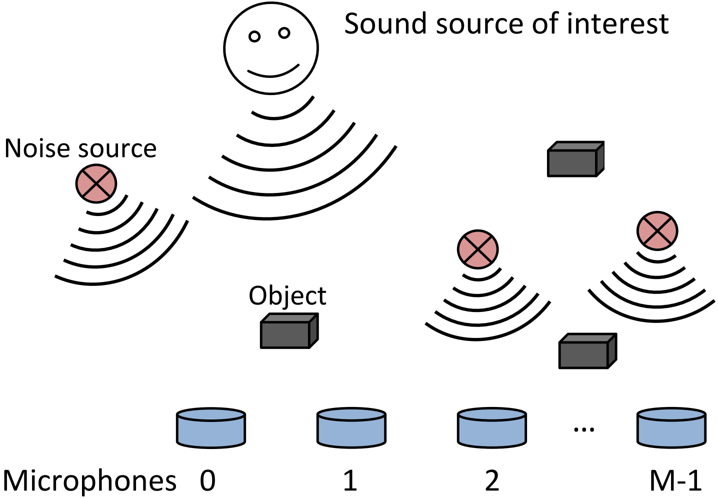
$$\hat{y}_n(t) = \underbrace{\sum_{m=1}^{M} \{f_{mn} * y_m\}(t)}_{\text{microphone array processing}} + \underbrace{\sum_{l=1}^{L} \{q_{ln} * x_l\}(t)}_{\text{echo cancellation}}$$

Topics of Acoustic Signal Processing

- Beamforming
- **Blind Source Separation**
- **Sound Source Localization**
- Multichannel Echo Cancellation
- Dereverberation, Noise Suppression, Speaker Diarization

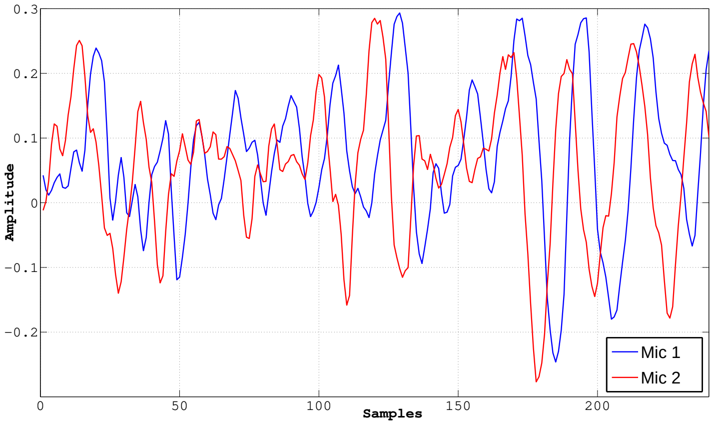
Outline

Sound Sources in an Acoustic Scene



Time-Delay Estimation

Example: Signals at two microphones



Signal Model: Two Microphones

Signals in the discrete time domain

$$\begin{cases} x_1[n] &= s[n] + v_1[n] \\ x_2[n] &= s[n-\Delta] + v_2[n] \end{cases}$$

- $x_1[n], x_2[n]$: Captured signal at each microphone
- $s[n]$: Source signal
- $\Delta = f_s\tau$: Time difference of arrival (TDOA) in samples
- $v_1[n], v_2[n]$: Noise and reverberation at each microphone

Signal Model: Two Microphones

Signals in the discrete time domain

$$\begin{cases} x_1[n] &= s[n] + v_1[n] \\ x_2[n] &= s[n-\Delta] + v_2[n] \end{cases}$$

- $x_1[n], x_2[n]$: Captured signal at each microphone
- $s[n]$: Source signal
- $\Delta = f_s \tau$: Time difference of arrival (TDOA) in samples
- $v_1[n], v_2[n]$: Noise and reverberation at each microphone

Time-Delay Estimation

A problem to find Δ given $x_1[n]$ and $x_2[n]$

Generalized Cross Correlation Method for TDE

Signals in the frequency domain

$$\begin{cases} X_1(\omega) & = S(\omega) + V_1(\omega) \\ X_2(\omega) & = S(\omega)e^{-j\omega\Delta} + V_2(\omega) \end{cases}$$

Generalized Cross Correlation Method for TDE

Signals in the frequency domain

$$\begin{cases} X_1(\omega) & = S(\omega) + V_1(\omega) \\ X_2(\omega) & = S(\omega)e^{-j\omega\Delta} + V_2(\omega) \end{cases}$$

Generalized Cross Correlation (GCC) Method

$$\hat{\Delta} = \arg\max_{\Delta} \int_{\omega} \psi(\omega) X_1(\omega) X_2^*(\omega) e^{j\omega\Delta} d\omega$$

Generalized Cross Correlation Method for TDE

Signals in the frequency domain

$$\begin{cases} X_1(\omega) & = S(\omega) + V_1(\omega) \\ X_2(\omega) & = S(\omega)e^{-j\omega\Delta} + V_2(\omega) \end{cases}$$

Generalized Cross Correlation (GCC) Method

$$\hat{\Delta} = \arg\max_{\Delta} \int_{\omega} \psi(\omega) X_1(\omega) X_2^*(\omega) e^{j\omega\Delta} d\omega$$

Phase Transform (PHAT) $\Rightarrow \psi_{\text{PHAT}}(\omega) = 1/|X_1(\omega)X_2^*(\omega)|$

$$\hat{\Delta} = \arg\max_{\Delta} \int_{\omega} \frac{X_1(\omega)X_2^*(\omega)}{|X_1(\omega)X_2^*(\omega)|} e^{j\omega\Delta} d\omega$$

Sound Source Localization With a Microphone Array

Multi-microphone extension of the GCC method

$$\hat{\mathbf{q}} = \arg\max_{\mathbf{q}\in\mathcal{Q}} \sum_{m=1}^{M} \sum_{k=1}^{M} \int_{\omega} \psi(\omega) X_m(\omega) X_k^*(\omega) e^{j\omega\Delta_{mk}^{\mathbf{q}}} d\omega,$$

Sound Source Localization With a Microphone Array

Multi-microphone extension of the GCC method

$$\hat{\mathbf{q}} = \arg\max_{\mathbf{q}\in\mathcal{Q}} \sum_{m=1}^{M}\sum_{k=1}^{M}\int_{\omega} \psi(\omega)X_m(\omega)X_k^*(\omega)e^{j\omega\Delta_{mk}^{\mathbf{q}}}d\omega,$$

- where $\Delta_{mk}^{\mathbf{q}} = \Delta_k^{\mathbf{q}} - \Delta_m^{\mathbf{q}}$ and $\mathcal{Q}$ denotes the search space of all potential source locations.

Sound Source Localization With a Microphone Array

Multi-microphone extension of the GCC method

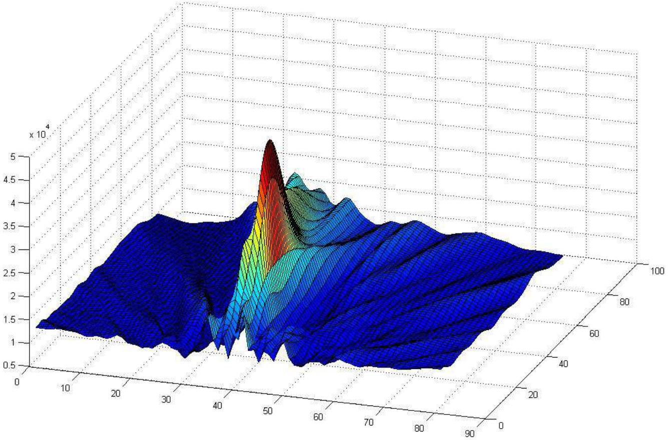
$$\hat{\mathbf{q}} = \arg\max_{\mathbf{q}\in\mathcal{Q}} \sum_{m=1}^{M}\sum_{k=1}^{M}\int_{\omega} \psi(\omega) X_m(\omega) X_k^*(\omega) e^{j\omega\Delta_{mk}^{\mathbf{q}}}\, d\omega,$$

- where $\Delta_{mk}^{\mathbf{q}} = \Delta_k^{\mathbf{q}} - \Delta_m^{\mathbf{q}}$ and $\mathcal{Q}$ denotes the search space of all potential source locations.

With the PHAT weighting, it can be simplified as

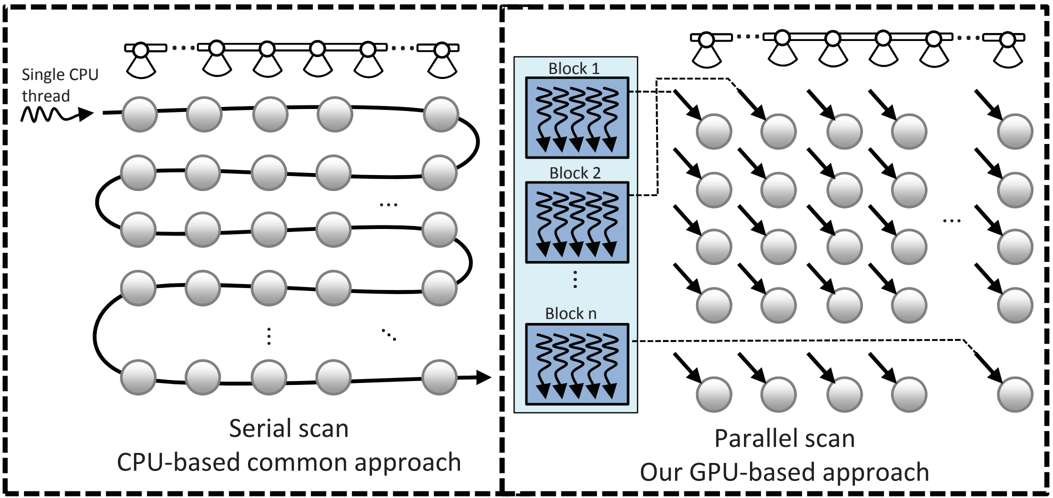
$$\hat{\mathbf{q}} = \arg\max_{\mathbf{q}\in\mathcal{Q}} \int_{\omega} \left| \sum_{m=1}^{M} \frac{X_m(\omega)}{|X_m(\omega)|} e^{j\omega\Delta_m^{\mathbf{q}}} \right|^2 d\omega,$$

Steered Response Power With Phase-Transform

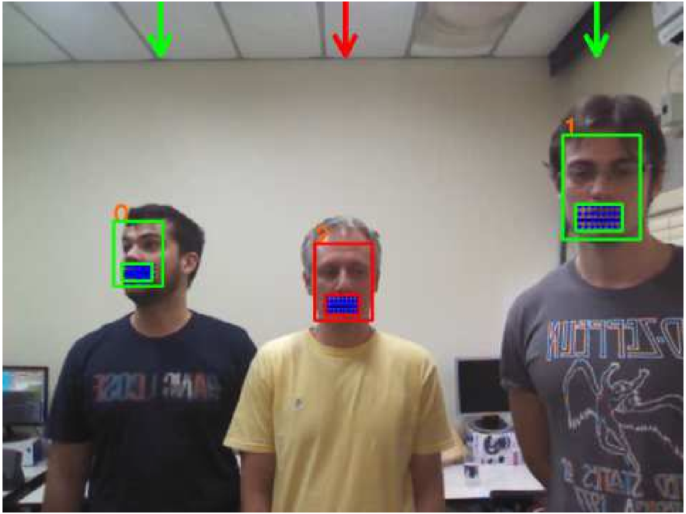


SRP-PHAT “Power Plot”

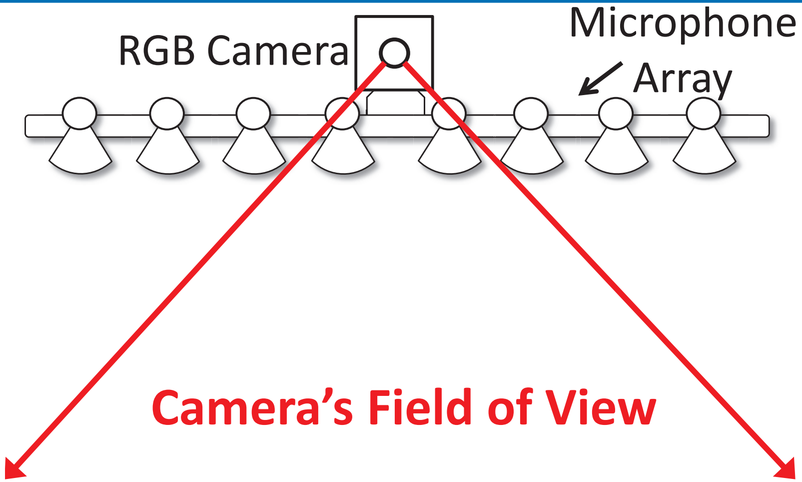
SRP-PHAT Computation on GPU



Simultaneous Speaker Detection and Localization



System Setup

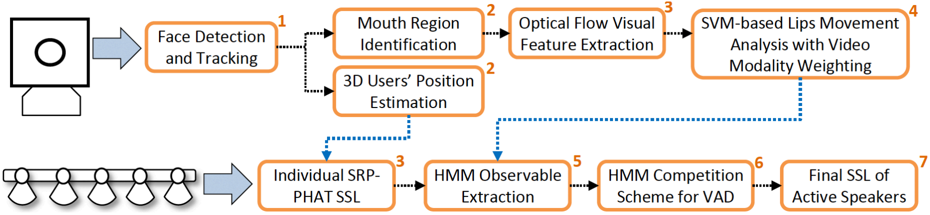


Eight microphones and one RGB camera

Multimodal Speaker Detection and Localization

Use the HMM competition scheme

- Run it for each one of the detected faces
- The SRP-PHAT plot has many local maxima for multiple simultaneous speaker cases
- Mid-Fusion of the parameters before HMM competition[1]



Mid-Fusion Algorithm for Speaker Detection and Localization

[1] V. P.Minotto, C. R. Jung, and B. Lee, "Simultaneous-Speaker Voice Activity Detection and Localization Using Mid-Fusion of SVM and HMMs," IEEE TMM 2014

Sound Source Localization

Note that:

- Sound source localization methods work reasonably well
  For long (200 ms or longer) signals,
  With relatively low reverberation and background noise, or
  With the help of multimodal sensor fusion
- Otherwise, they usually fail.

Sound Source Localization

Note that:

- Sound source localization methods work reasonably well

  For long (200 ms or longer) signals,

  With relatively low reverberation and background noise, or

  With the help of multimodal sensor fusion
- Otherwise, they usually fail.

It is generally believed that:

- Humans do much better than any algorithmic methods.

Sound Source Localization

Note that:

- Sound source localization methods work reasonably well

  For long (200 ms or longer) signals,

  With relatively low reverberation and background noise, or

  With the help of multimodal sensor fusion
- Otherwise, they usually fail.

It is generally believed that:

- Humans do much better than any algorithmic methods.

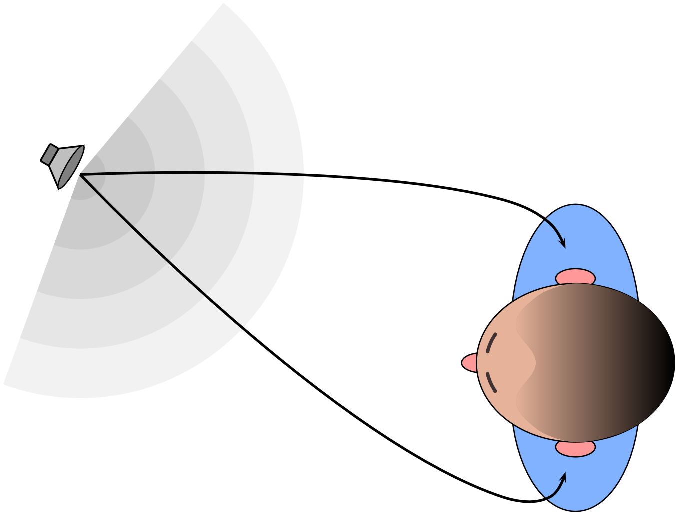
Cocktail party effect

- The ultimate goal of acoustic sensing

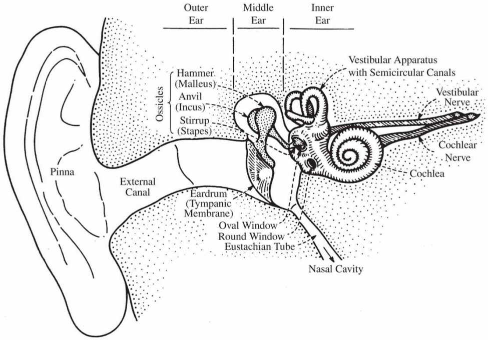
Cocktail Party

Outline

Human Hearing



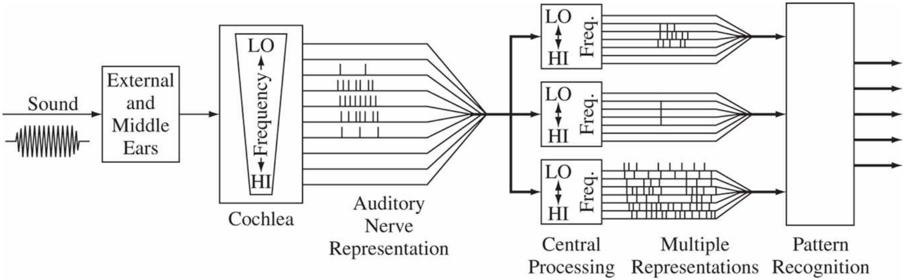
Human Ear



(source: Theory and Applications of Digital Speech Processing, Rabiner and Schafer, Pearson, 2011)

- Acoustic wavefront hits the eardrum (outer ear)
- Movement at the eardrum is transmitted via ossicles (middle ear)
- Stapes is attached to the oval window of the cochlea (inner ear)

Sound Representation in the Auditory System

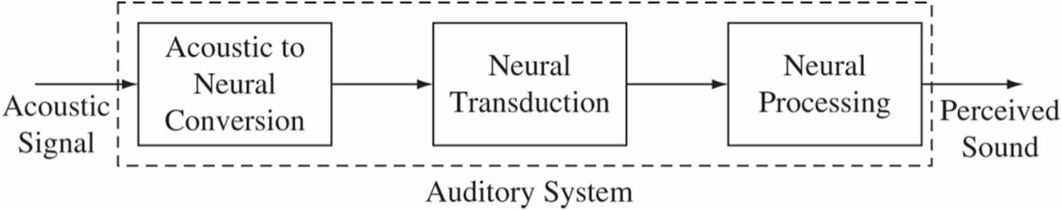


(source: Theory and Applications of Digital Speech Processing, Rabiner and Schafer, Pearson, 2011)

Cochlear processing

- Vibration at the oval window is converted to fluid motion in the cochlea
- Fluid motion causes mechanical vibration on the basilar membrane
- Mechanical vibration causes action potential in the inner hair cell
- These action potentials trigger neural firings to be aggregated in the central auditory system

Auditory System: A Neural Network



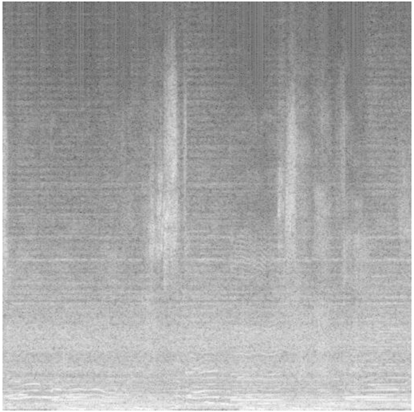
(source: Theory and Applications of Digital Speech Processing, Rabiner and Schafer, Pearson, 2011)

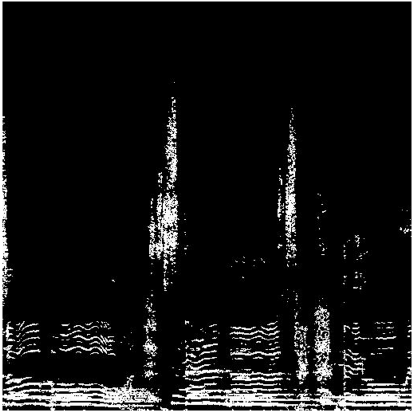
Auditory System

- The auditory cortex serves as the central processing unit
- It creates multiple representations of the neural firings
- Then the human brain perceives the sound

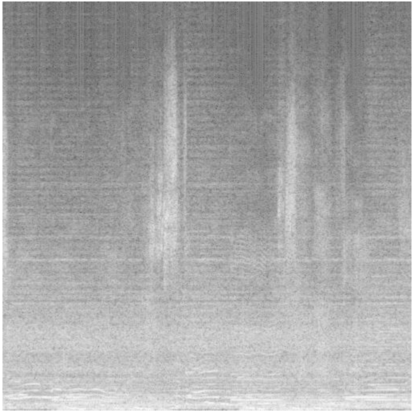
Outline

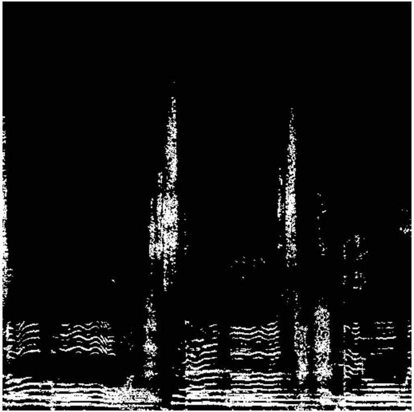
Spectrogram





Spectrogram

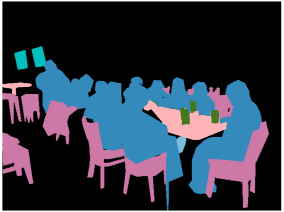




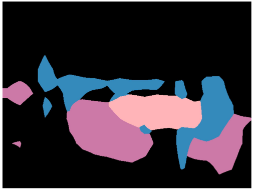
- Time-Frequency representation of audio
- Can be treated as a 2D image

Semantic Segmentation: Fully Convolutional Network

<Image>



<GT>



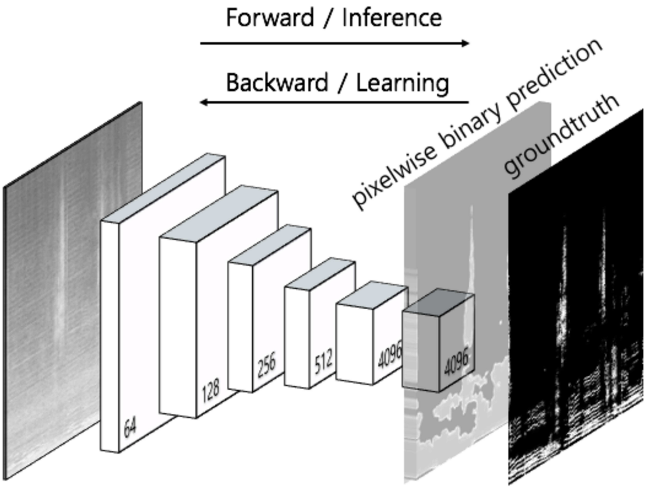
<32s>

<16s>

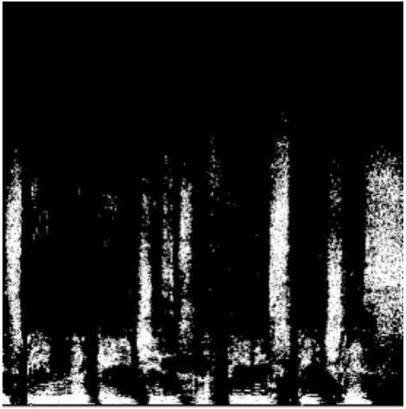
<8s>

Semantic segmentation: Pixel-wise image segmentation

Blind Source Separation: Fully Convolutaional Network



Blind Source Separation: Some Results



Groundtruth

Binary Mask

The FCN created the binary mask from a mixture signal

Summary

Acoustic Sensing

Summary

Acoustic Sensing

+

Summary

Acoustic Sensing

+

Artificial Intelligence

Conclusion

Past

Current

Future

Questions?