Microsoft Research

Microsoft Research Asia Faculty Summit 2012

Multi-user Source Localization using Audio-Visual Information of KINECT

Hong-Goo Kang
DSP Lab, Dept. of EE, Yonsei University

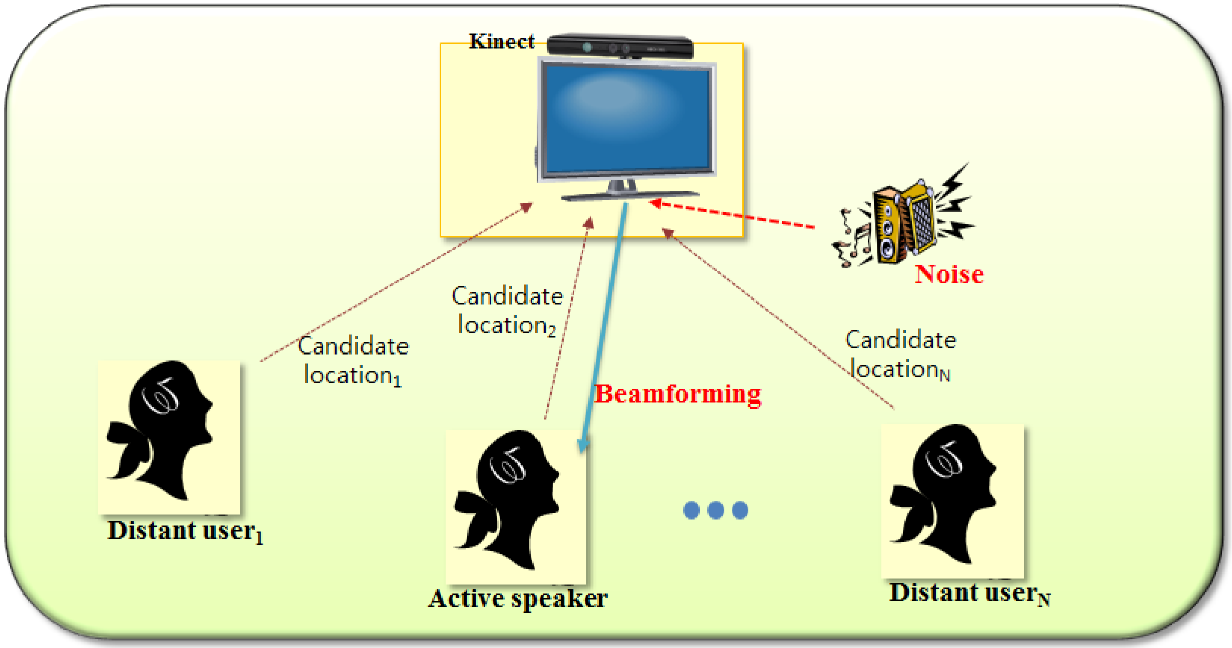
Contents

Yonsei University DSP Lab.

- Overview: Research Overview
- Group 1 : Multiple User Localization
- Group 2 : Active Speaker Detection and Beamforming
- Group 3 : Speaker Identification
- Discussion

Research Overview (1)

- Multiple user source tracking and localization
- Microphone array processing: beamforming
- Active speaker detection & real time speaker identification

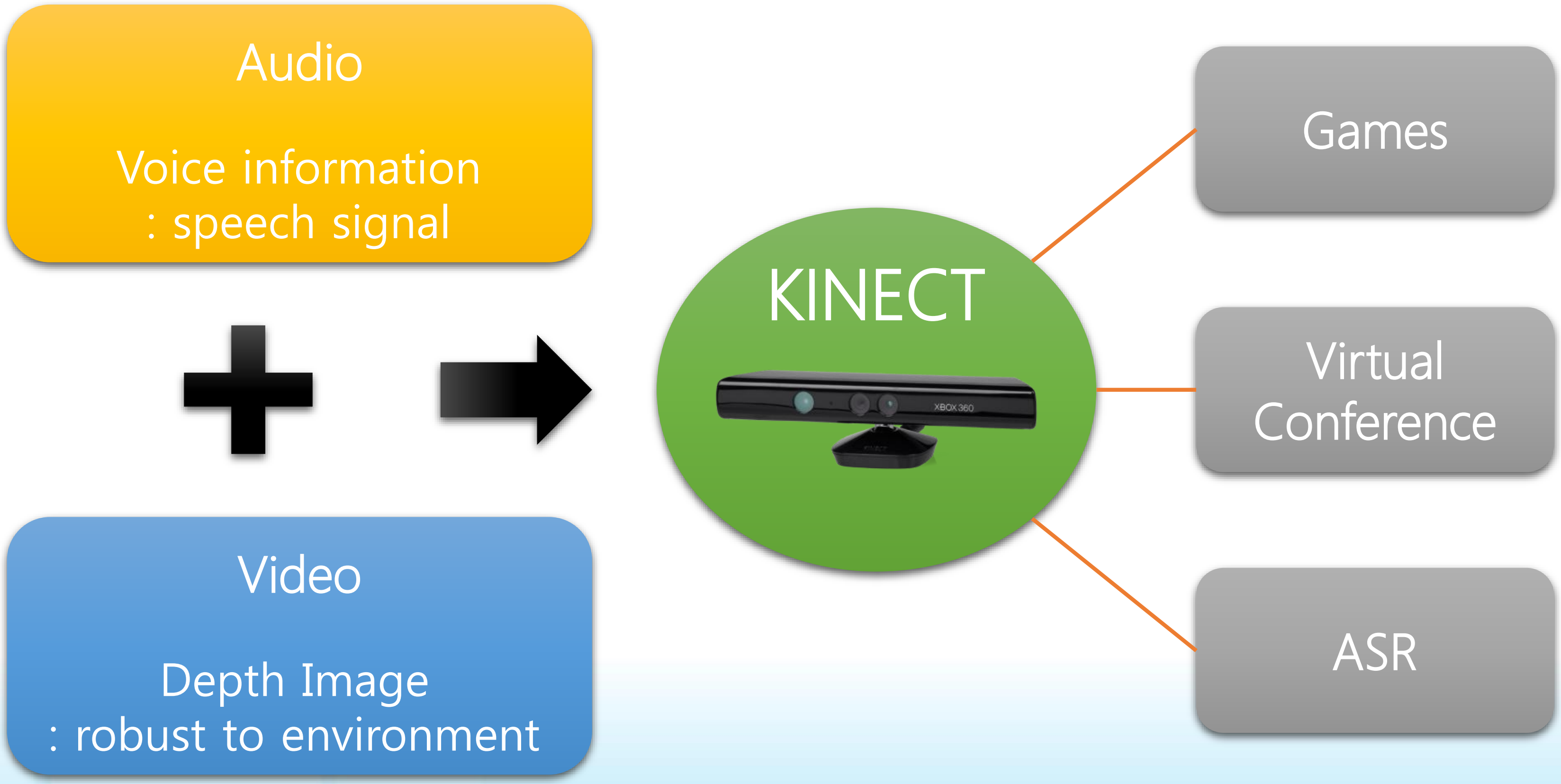


Research Overview (2)

- Speech signal processing for multi-user environment
  - Requires a user dependent processing → user location/direction
  - Degrades performance in harsh acoustical environment, e.g. noise, reverberation, interference, and so on → beamforming

- Solutions/Approaches
  - Introduces video signal processing approaches (depth image)
    - Head/face detection and tracking
  - Improves the performance of microphone array based speech enhancement algorithms (beamforming) using the head location information
  - Apply the enhanced signal to speech signal processing applications, i.e. speaker recognition/identification

Proposed Framework



Research Groups

Group 1

Multi-user Localization

- Video-based head detection and tracking
- Coordinate translation 2D -> 3D

Group 2

Beamforming

- Active speaker detection
- Beamforming

Group 3

Speaker Identification

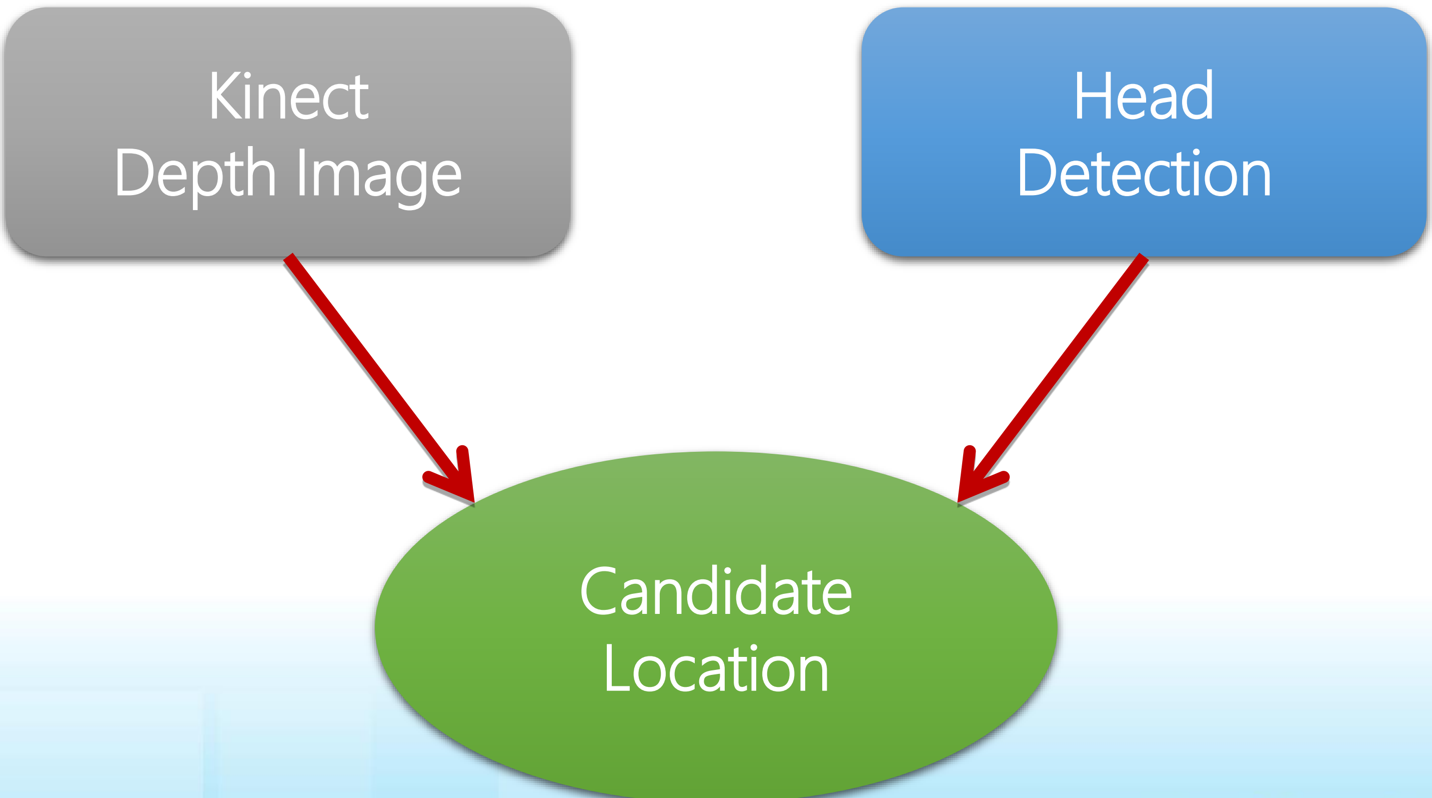
- Feature extraction
- User identification

Contents

Yonsei University DSP Lab.

Multiple User Localization

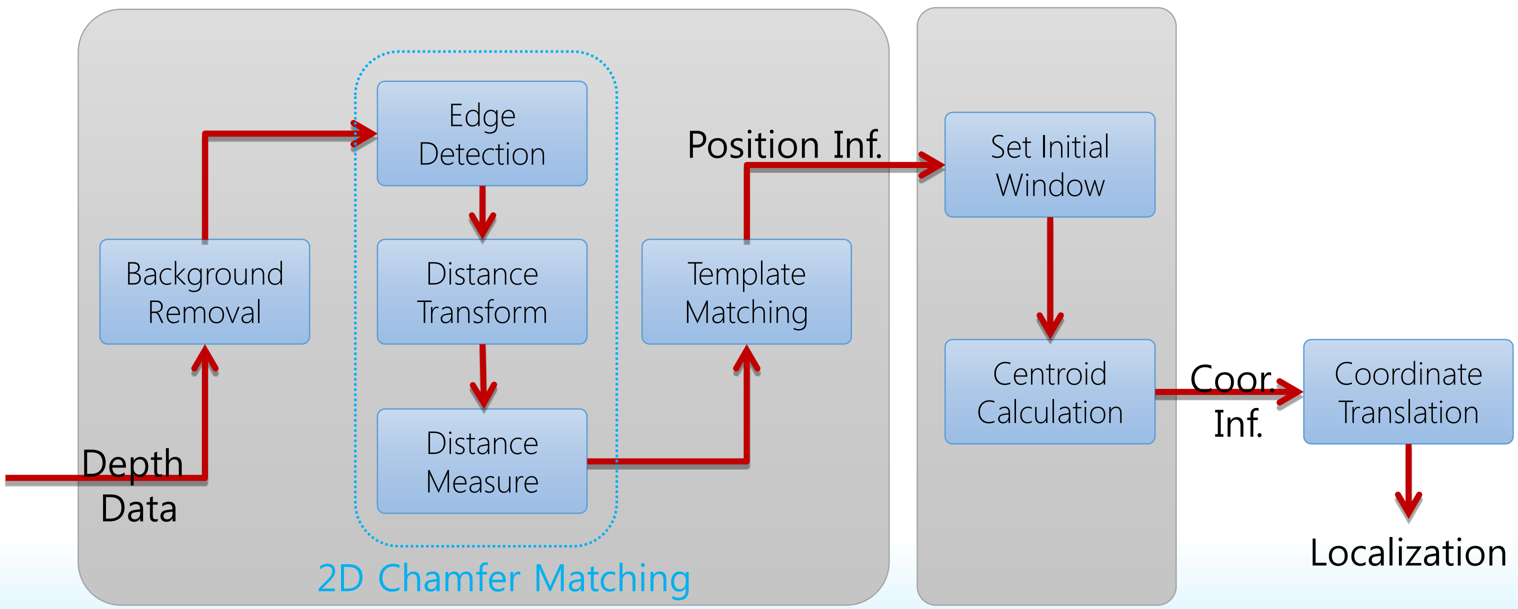
- Objective

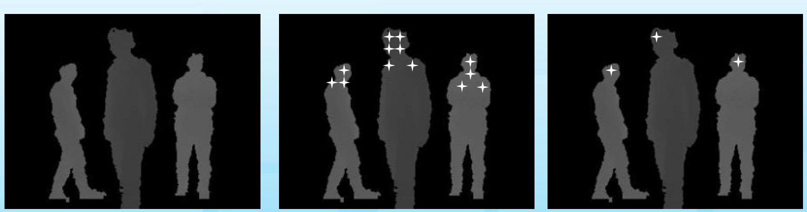


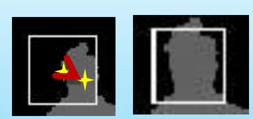
Depth Image based Head Tracking

- Overview of head detection and tracking

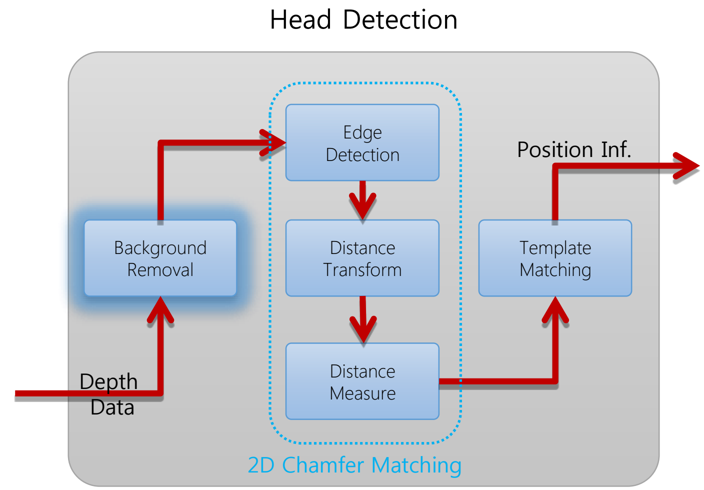
Head Detection   Head Tracking



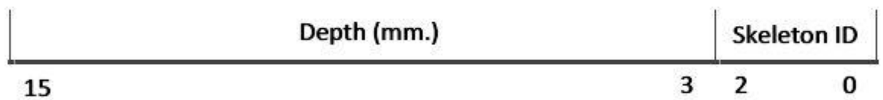




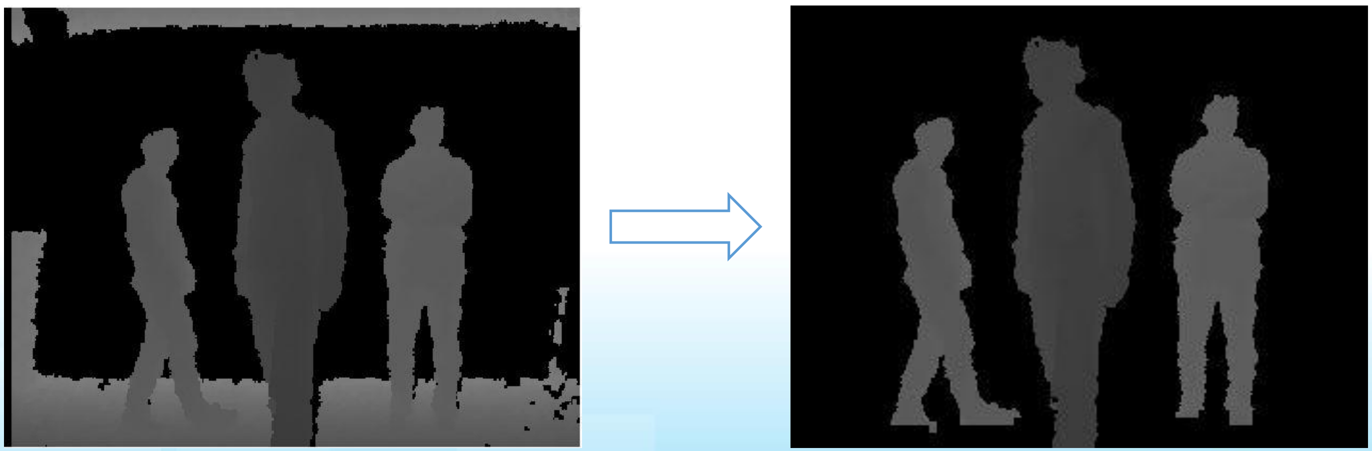
Head Detection (1)



- Background removal
  - Use player information

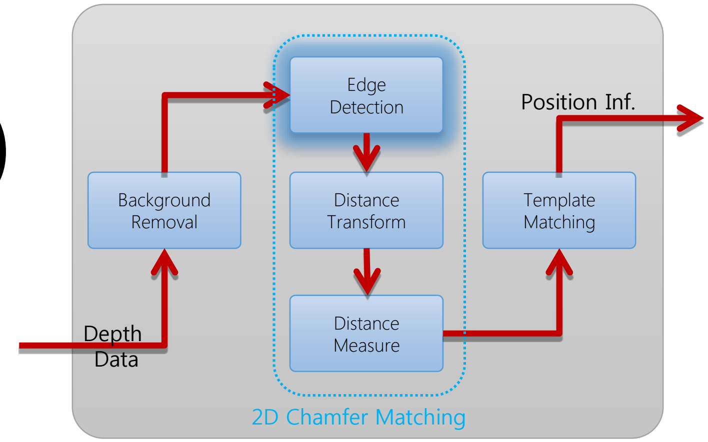


  - Player index = 0 ⟹ Background!
  - Result



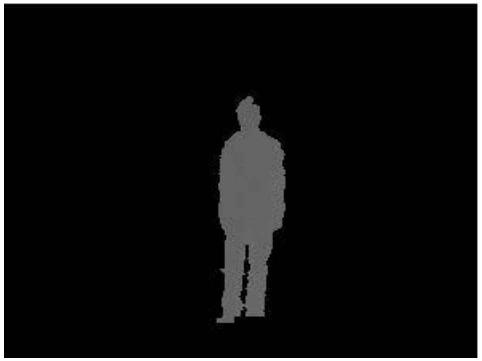
Head Detection (2)

Head Detection

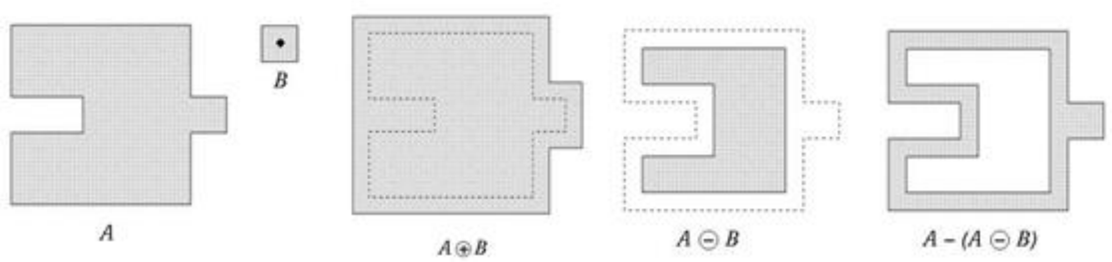


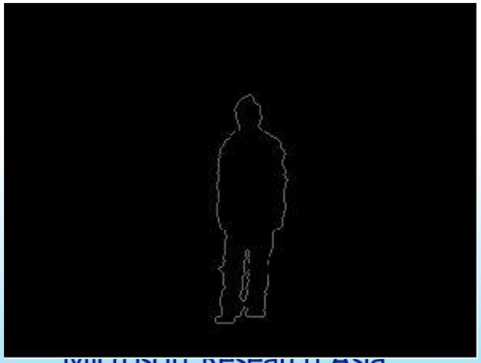
- Chamfer matching
  - Shape detection algorithm

- Morphological edge detection
  - Uses dilation and erosion images
    - Dilation : grow image region
    - Erosion : shrink image region

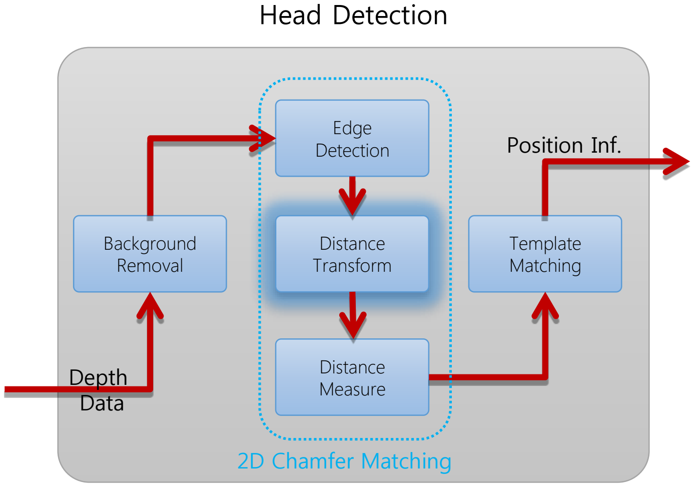


Original



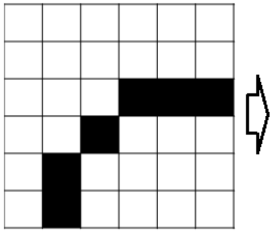


Edge image



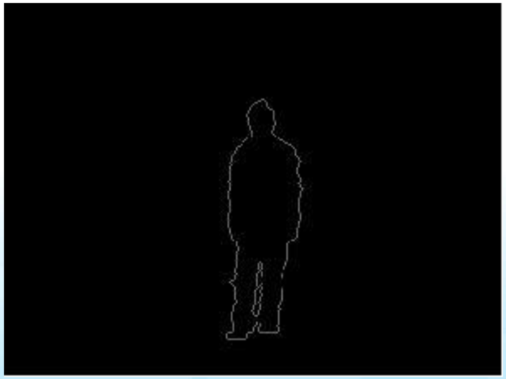
Head Detection (3)

- Distance Transform
  - Converts binary image into distance image
  - Results in distance from the closest edge pixel

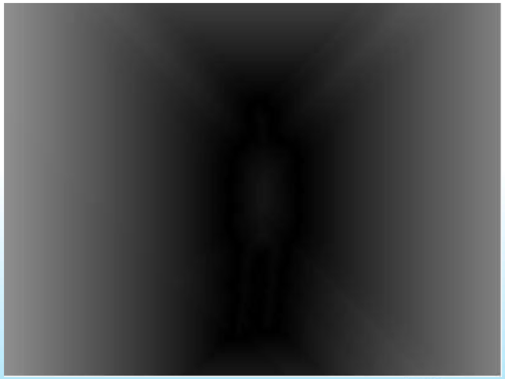


5	4	3	2	2	2
4	3	2	1	1	1
3	2	1	0	0	0
2	1	0	1	1	1
1	0	1	2	2	2
1	0	1	2	3	3

  - Result

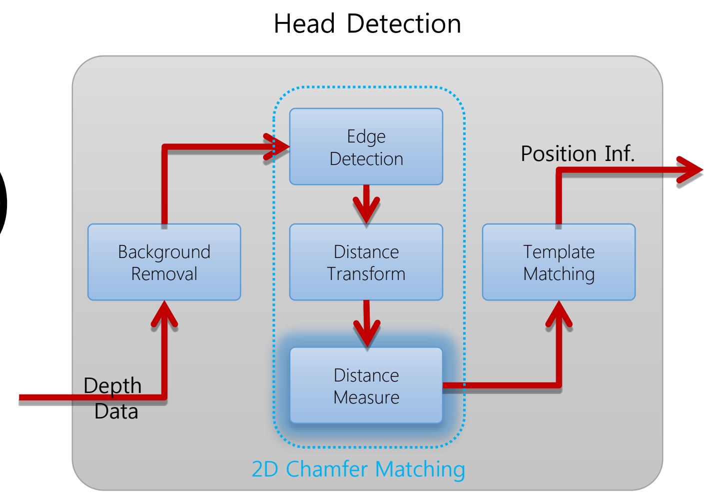


Edge image

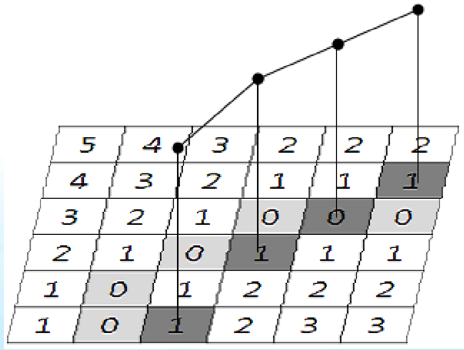


Distance image

Head Detection (4)

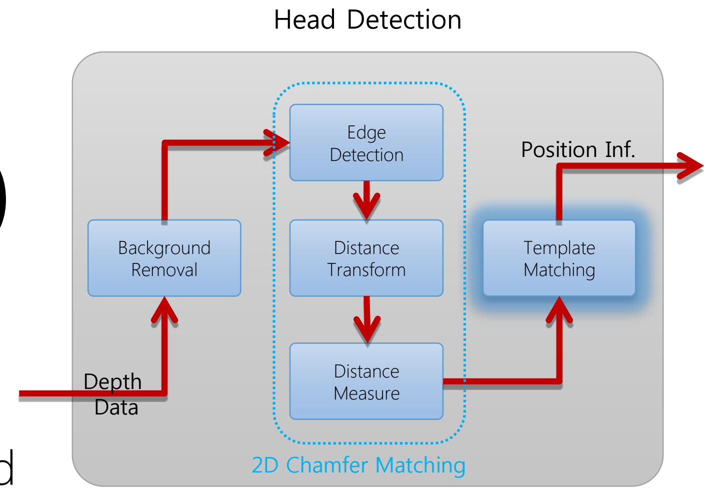


- Distance measure
  - Between edge and distance image
    - Edge image : template head image
    - Distance image : depth image
  - RMS chamfer distance
    - $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(e_i d_i)^2}$

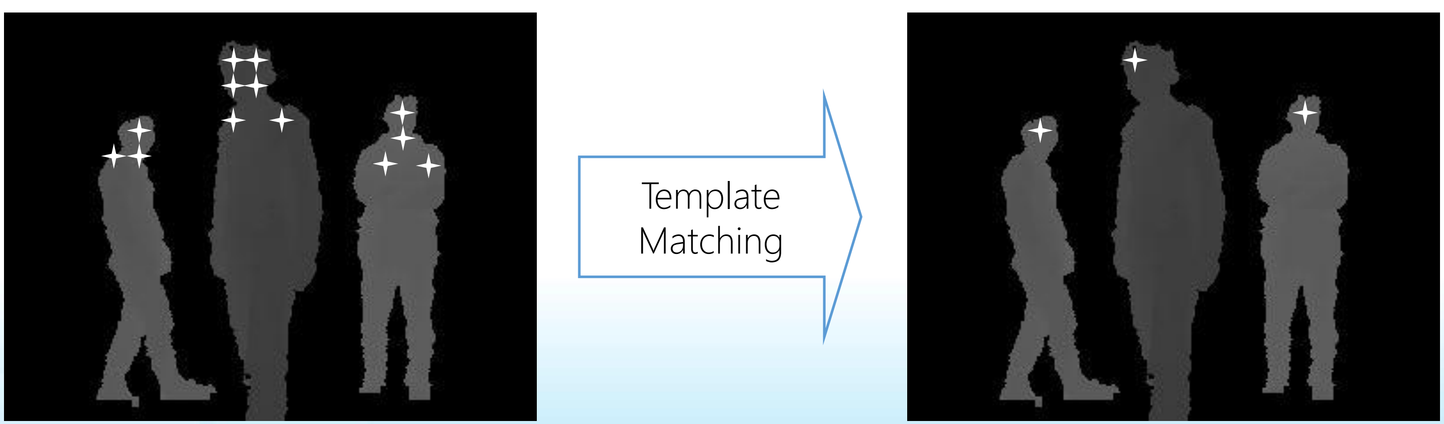


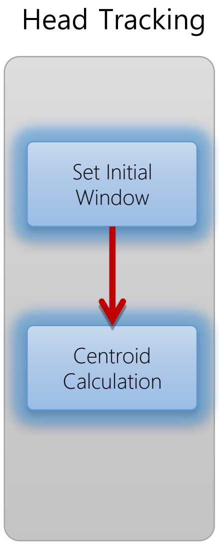
Black dots : edge image
Numbers : distance from edge

Head Detection (5)



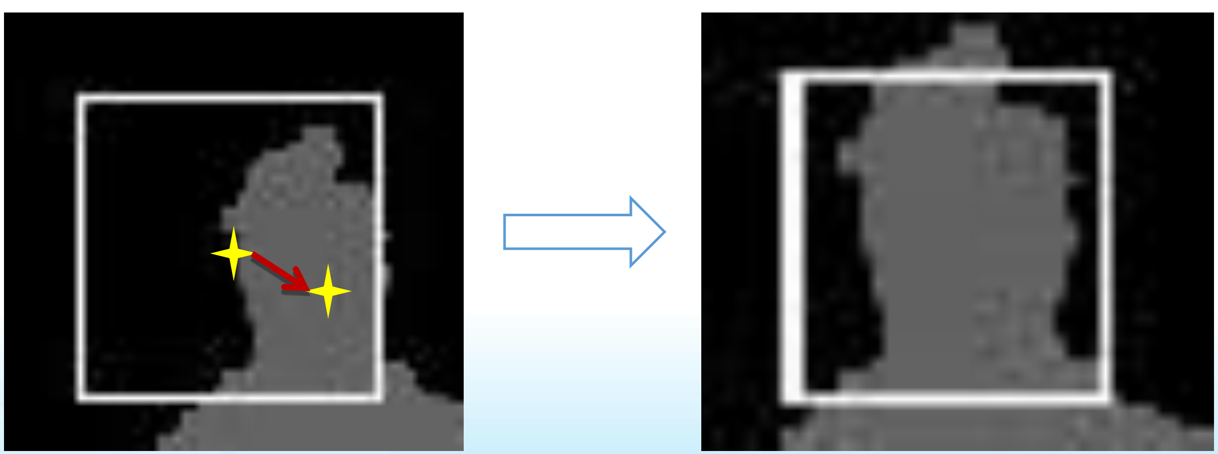
- Template matching
  - Detects candidate of head locations obtained from the chamfer matching
  - Uses nine head template images
  - Obtains fine head location





Head Tracking

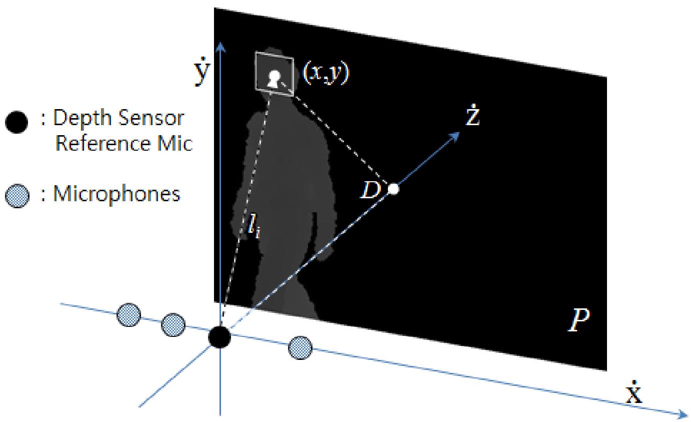
- Initial window
  - Sets the coordinates from head detection
- Real-time head tracking
  - Shifts the center of window center to the centroid of head
  - Adjusts the size of window



Coordinate Translation

- Sound source localization
  - Uses the center pixel of head location to relative location from KINECT
    - $(\dot{x}, \dot{y}, \dot{z}) = f(x, y, D)$

(x, y): Pixel indices
D: Distance
$(\dot{x}, \dot{y}, \dot{z})$: Relative source location in meters

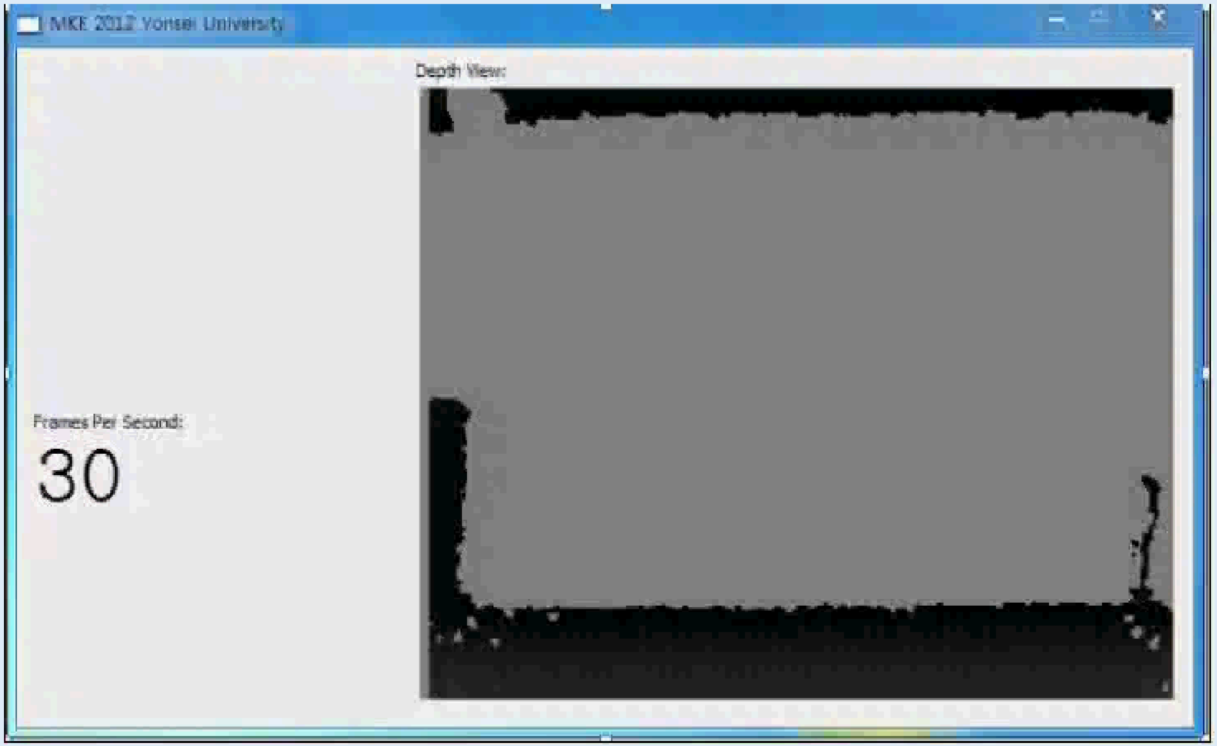


- Coordinate translation equation
  - Linear expressions using trigonometry

$$\dot{x} = \frac{13D(160 - x)}{4000}$$
$$\dot{y} = \frac{13D(120 - y)}{4000}$$
$$\dot{z} = D$$

Head Detection and Tracking

- Multi-user tracking demo



Contents

Yonsei University DSP Lab.

Beamforming

- Beamforming
  - Takes a spatial filtering
  - Enhances target speech by
    - suppressing noise and interference
- Beamforming process

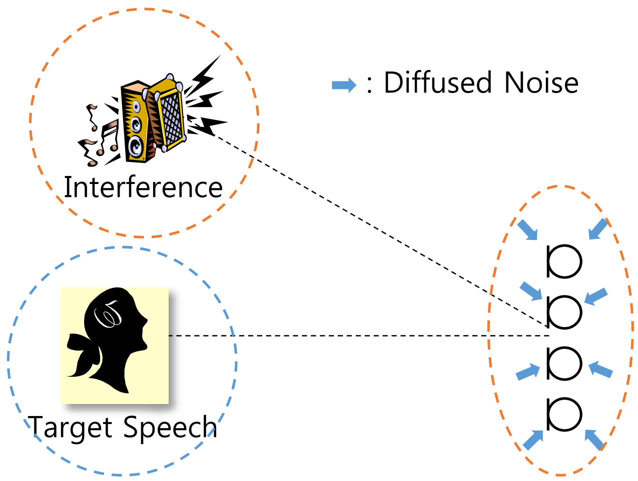
Source Location Information → Spatial Filtering

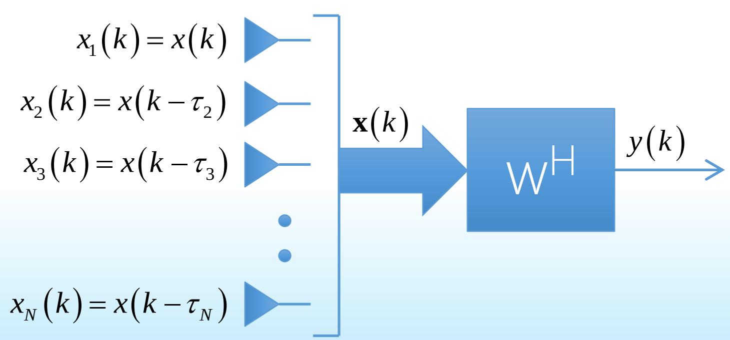
- Spatial Filtering

x: input signal
w: weights
y: output signal

$$y(k) = \mathbf{w}^{\mathrm{H}}\mathbf{x}$$

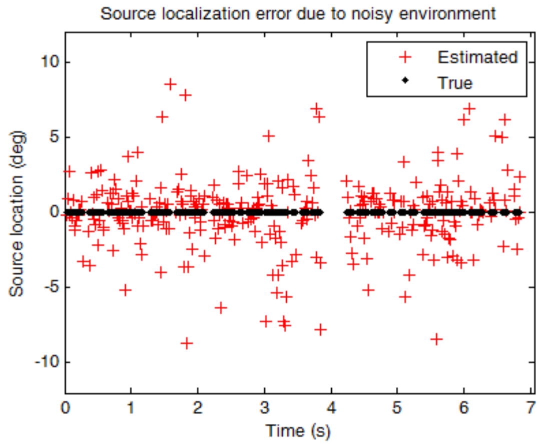
$$\mathbf{Y}(\omega) = \sum_{n=1}^{N} \mathbf{W}_n^H(\omega)\mathbf{X}_n(\omega)$$





Sound Source Localization
Simulation

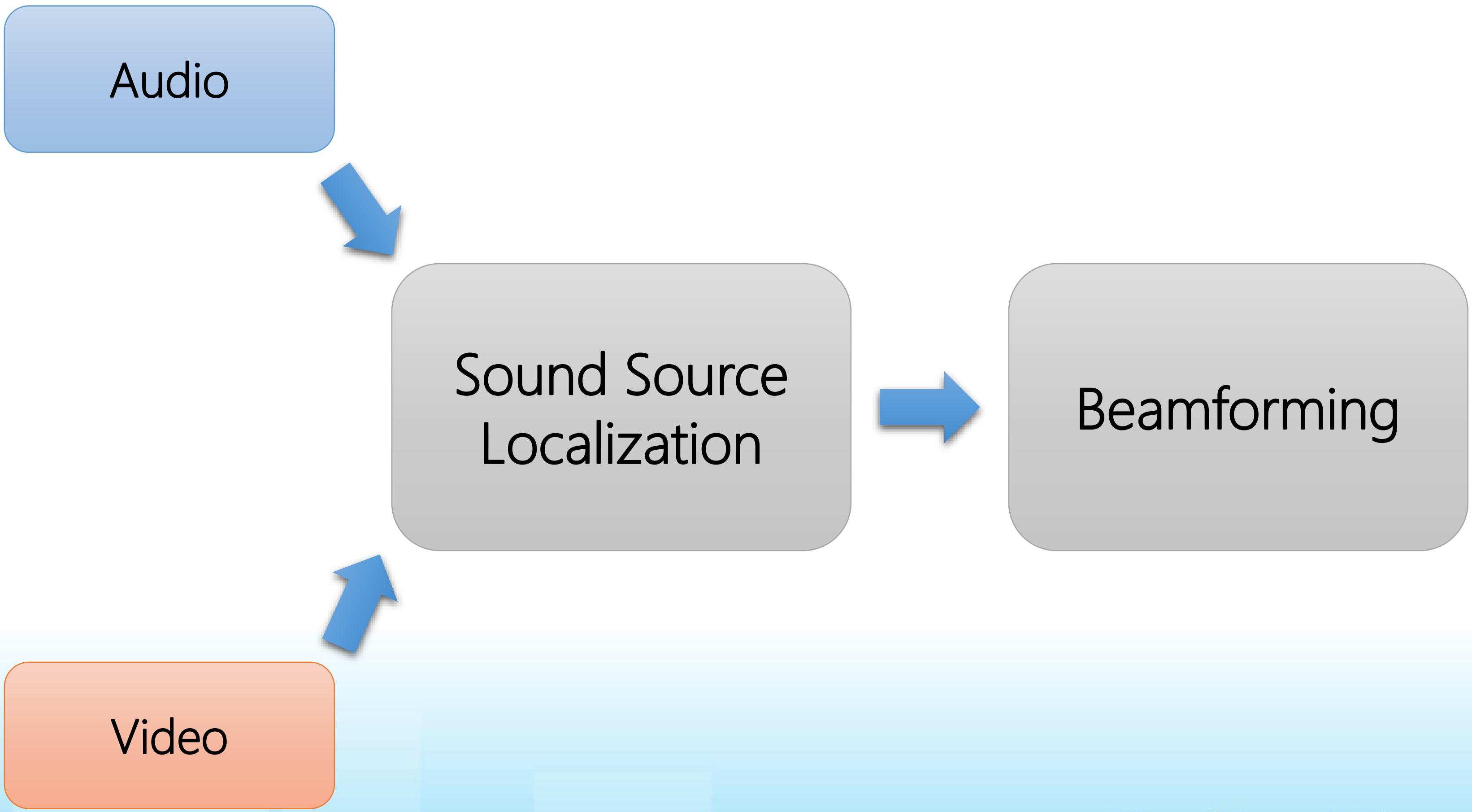
- Conventional DoA estimation
  - Male speech, 16kHz
  - 32ms Hanning window, 16ms overlap
  - Diffused white noise, 20dB SNR



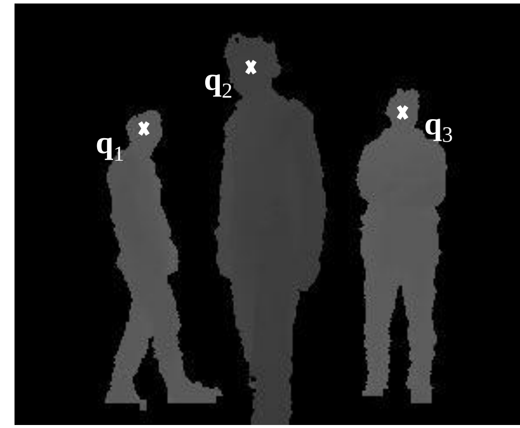
- Limitations of conventional approach
  - Susceptible to acoustical effects
    - Noise, Reverberation, Interference
  - Error in localization

    → Performance degradation in the beamforming processing
- High computational complexity
  - Need to estimate source location for every frame

Proposed Method - Overview

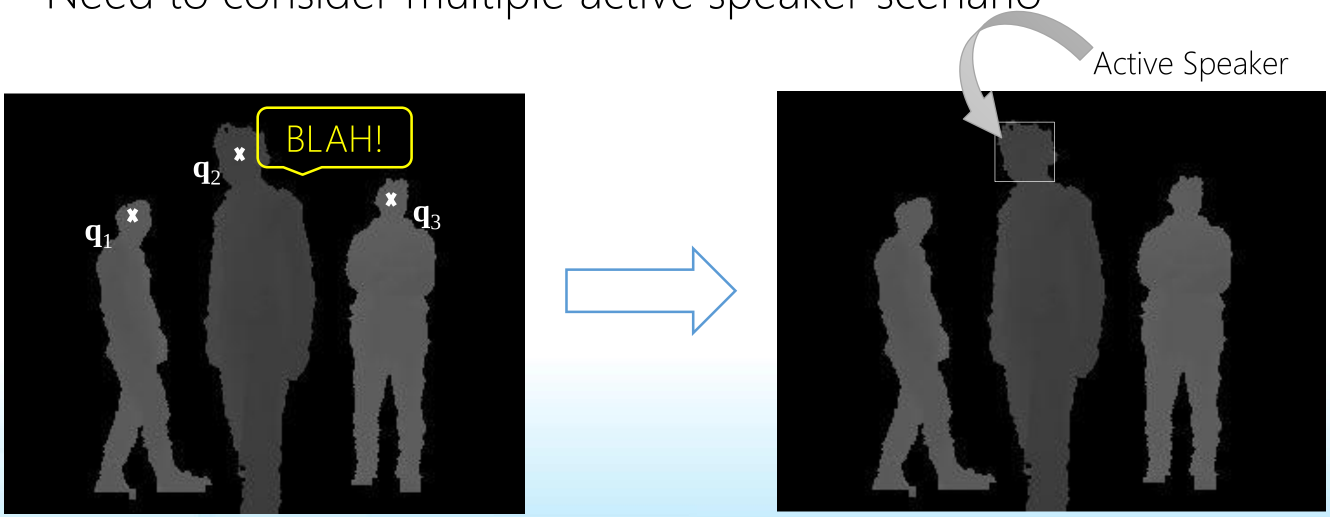


Beamforming for Multiple Candidates



- Q: multiple # of candidate speakers ?
  - How to find active speaker?

- A: Simultaneously form multiple beams
  - Need to consider multiple active speaker scenario

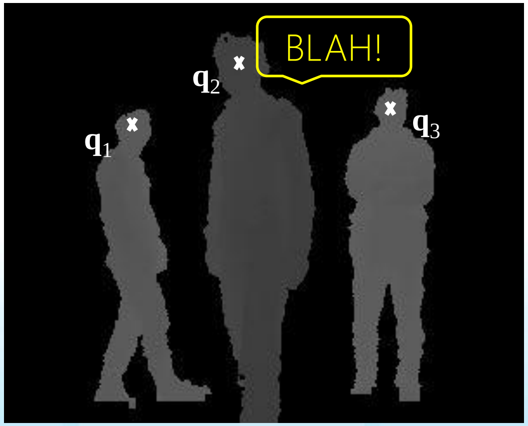


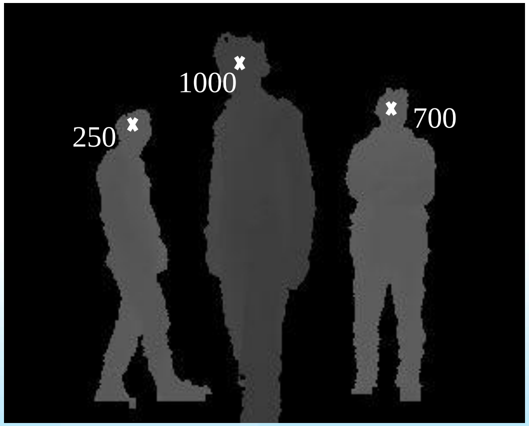
Active Speaker Detection (1)

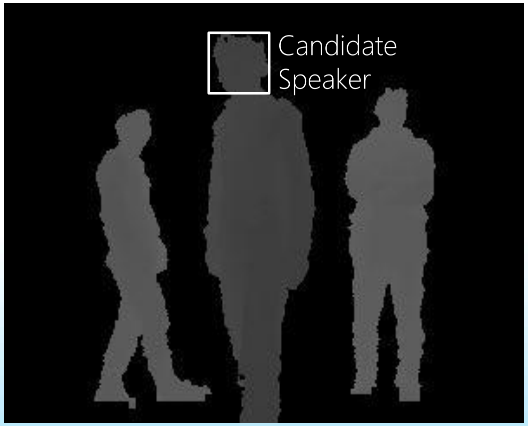
- Candidate speaker detection
  - Apply SRP-PHAT algorithm to all the candidates' locations
  - Find the location $\mathbf{q}$ that has the maximum sound level

$$\hat{\mathbf{q}} = \arg\max_{\mathbf{q}} P(\mathbf{q}) \ \forall\ \mathbf{q} = \{\mathbf{q}_i \mid i = 1, 2, \cdots, 6\}$$

$$P(\mathbf{q}) = \int_{-\infty}^{\infty} \left| \sum_{n=1}^{4} \frac{X_n(\omega) e^{j\omega\Delta(\mathbf{q})}}{|X_n(\omega)|} \right|^2 d\omega$$





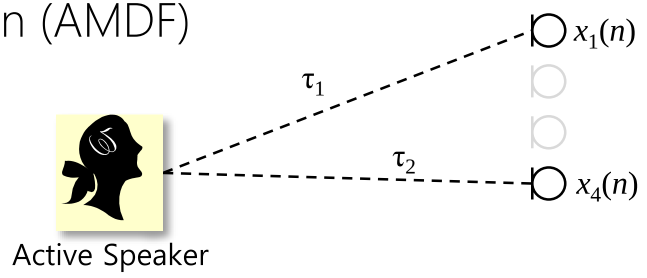


Active Speaker Detection (2)

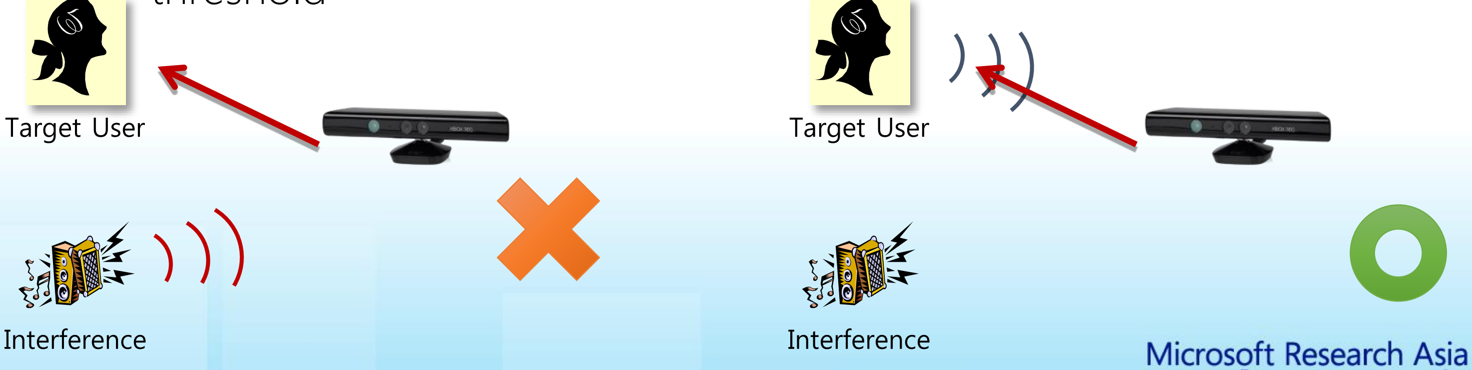
- A/V integrated active user localizer
  - Uses average magnitude difference function (AMDF)
    - Find k that best compensates τ_1-τ_2

$$\hat{k} = \arg\max_{k} \Psi_{AMDF}(k)$$

$$\Psi_{AMDF}(k) = \frac{1}{N}\sum_{n=0}^{N-1} \left| x_1(n) - x_4(n+k) \right|$$

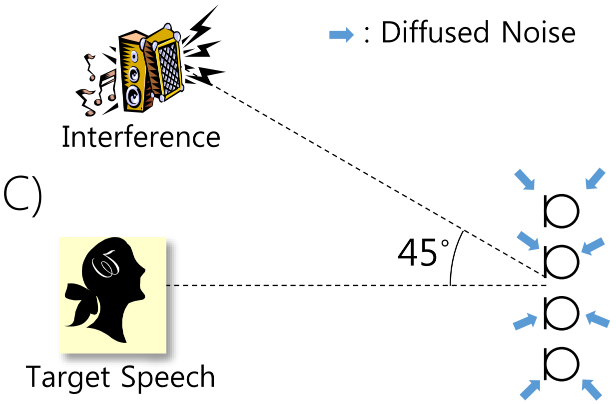


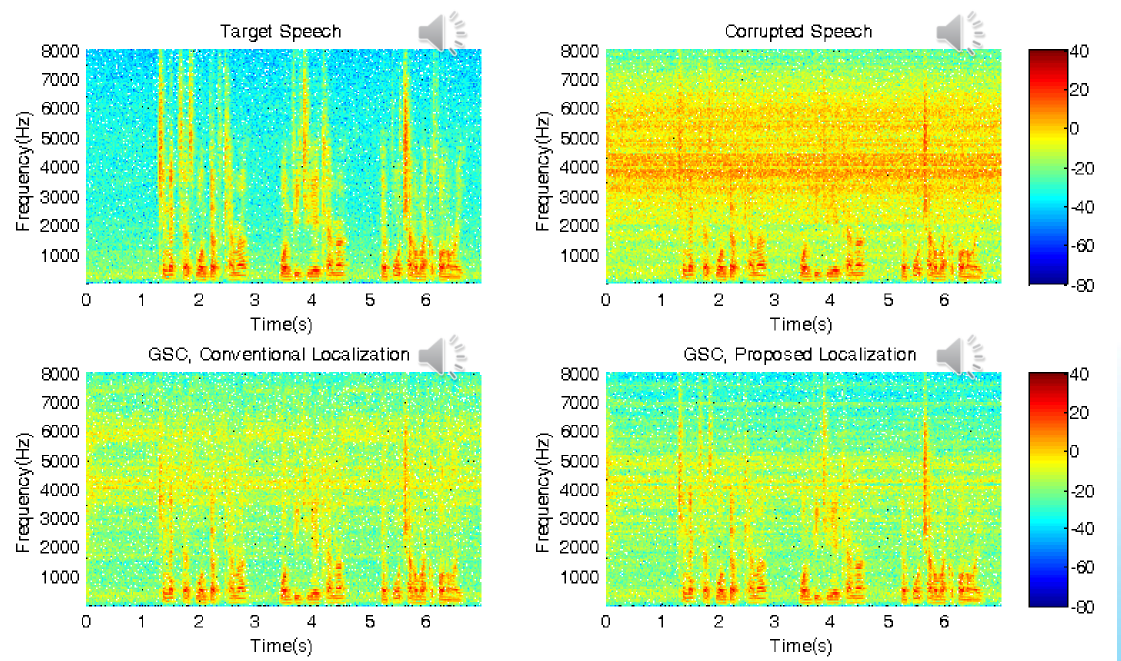
  - Compares $\hat{k}$ with actual τ_1 and τ_2 obtained from video signal
    - Speaker is active if the difference of direction is within the pre-defined threshold



Beamforming - Simulation

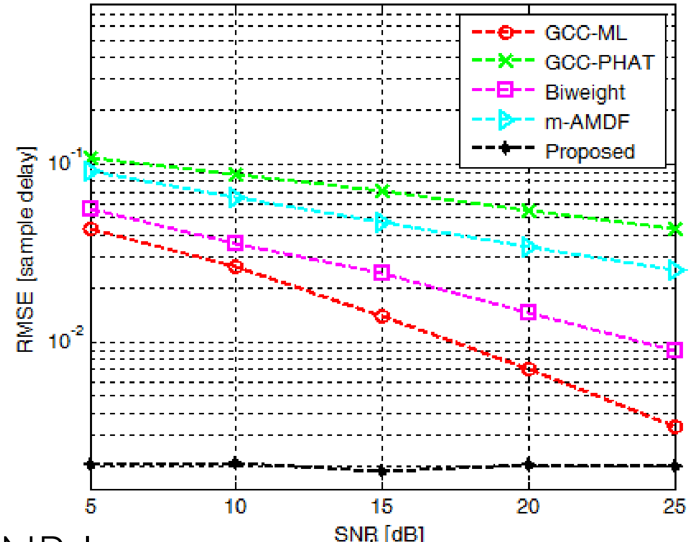
- Simulation set up
  - Uses Generalized Sidelobe Canceller (GSC)
  - Interference: white noise
- Results
  - Conventional vs proposed localization



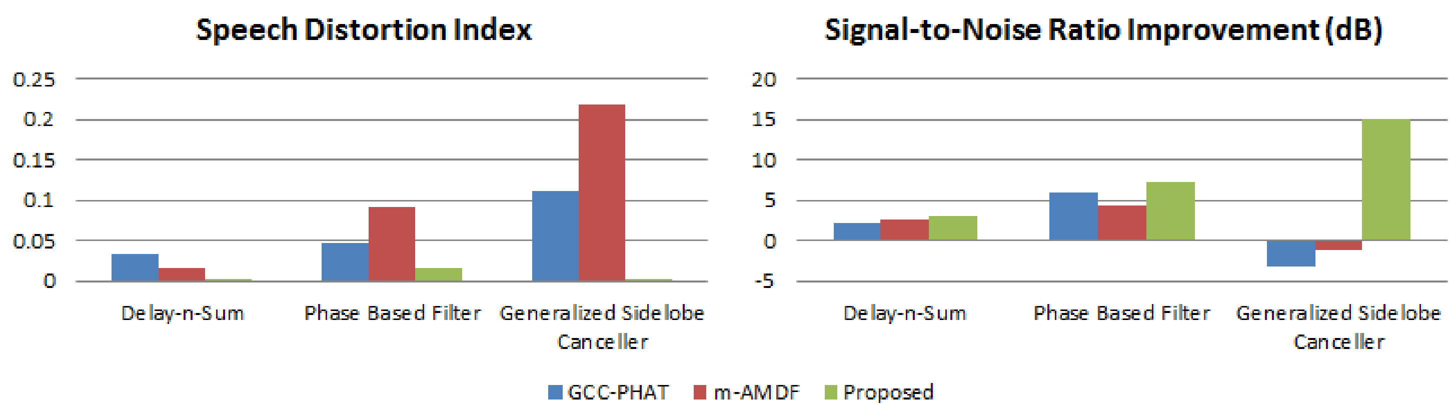


Performance Evaluation

- Localization error

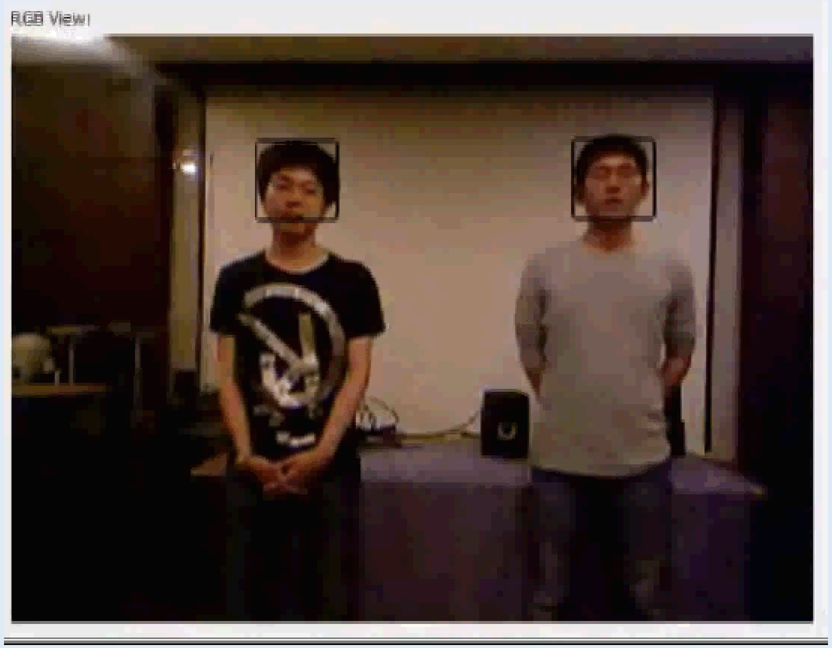


- Spectral distortion and SNR Improvement



Demonstration

- Beamforming
  - Noise reduction and target speech preservation
- Active speaker detection

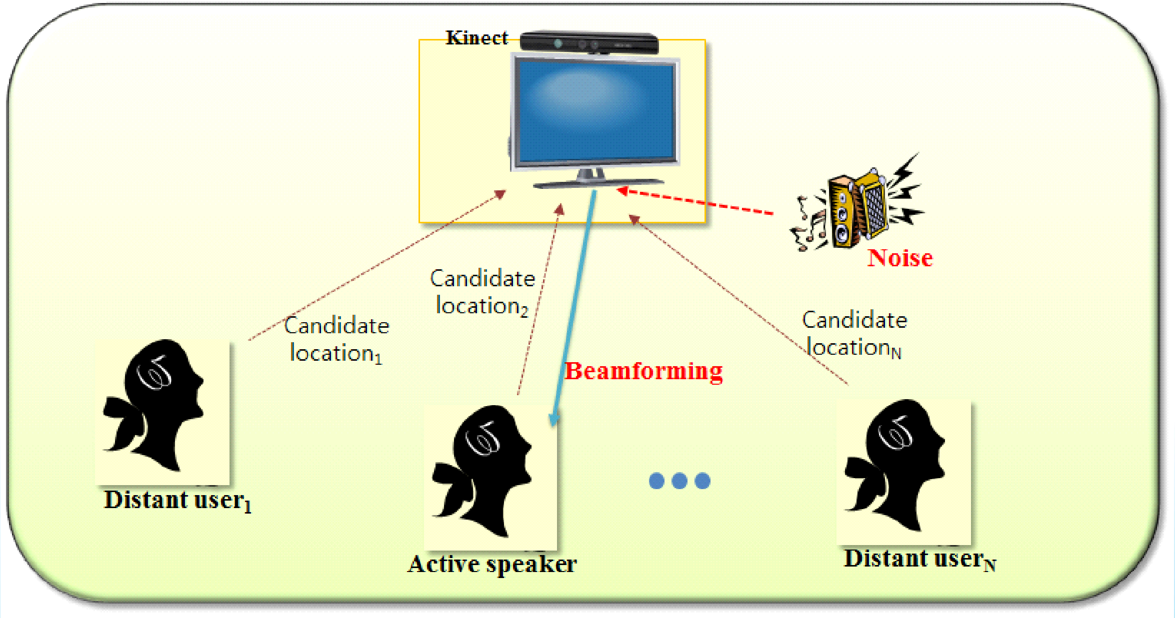


Contents

Yonsei University DSP Lab.

Application

- Online-meeting
  - Speaker identification
  - Speech recognition
  - Gesture recognition



Introduction

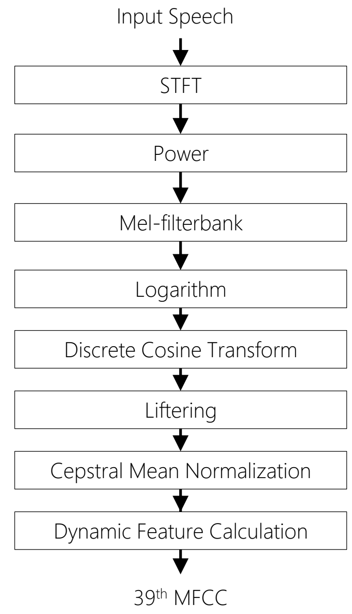
- Active speaker ID is displayed during speech active region

Problems

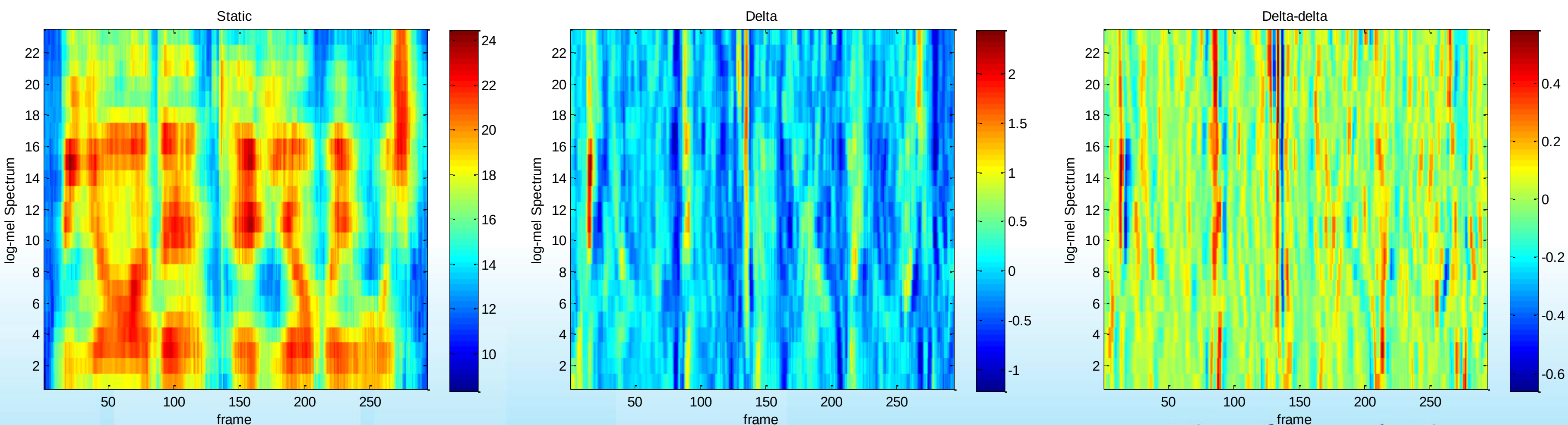
- Implementation issues
  - Real-time online system
    - Frame based decision instead of utterance based decision
  - Complexity
    - Can not use high order of Gaussian mixtures
    - Use the fact that a few of the mixtures of a GMM contributes significantly to the likelihood value for a speech feature vector
- Multi-speaker identification
  - Cannot use the conventional pruning algorithm

Feature Extraction (1)

- Mel-Frequency Cepstral Coefficients (MFCC)
  : followed by ETSI configuration
  - 39th-order MFCC include 0th order coefficient with delta and delta-delta
  - 25ms hamming window / 10ms shift
  - 24th-order mel-filterbank without 1st order (~64Hz)



Log-mel spectral features

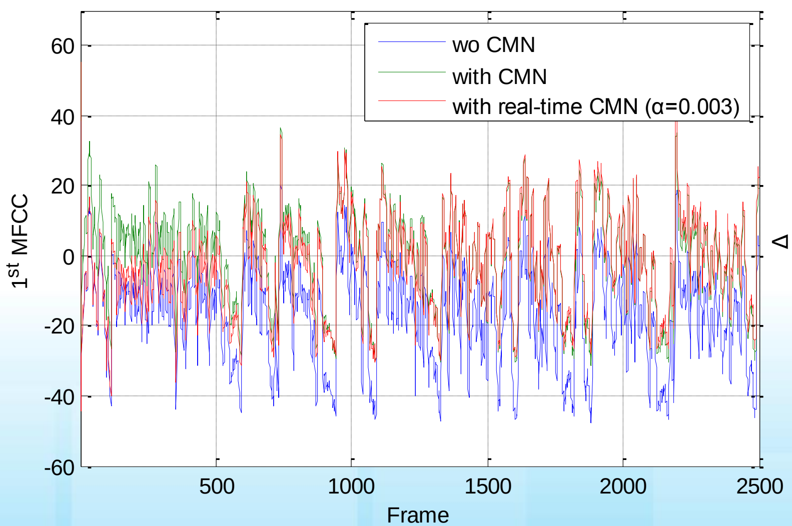


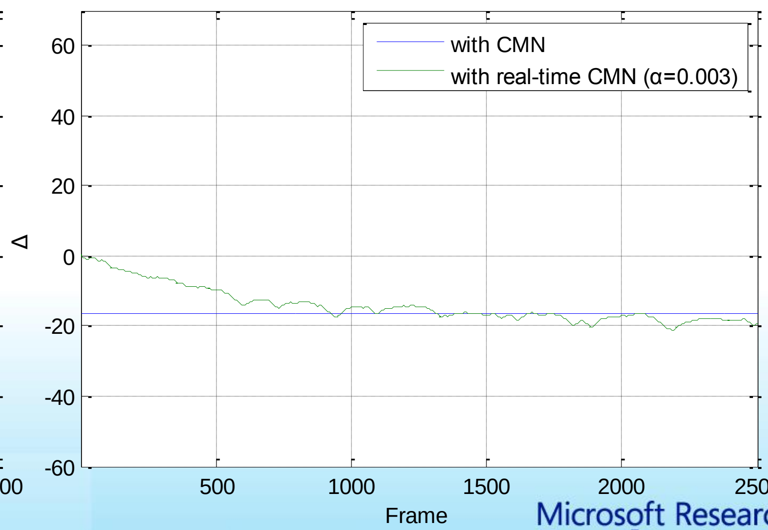
Feature Extraction (2)

- Real-time Cepstral Mean Normalization (CMN)
  - CMN cannot be applied to real-time applications directly.
    - Mean vector can be calculated after receiving the input signal completely
  - Uses an approximated on-line technique

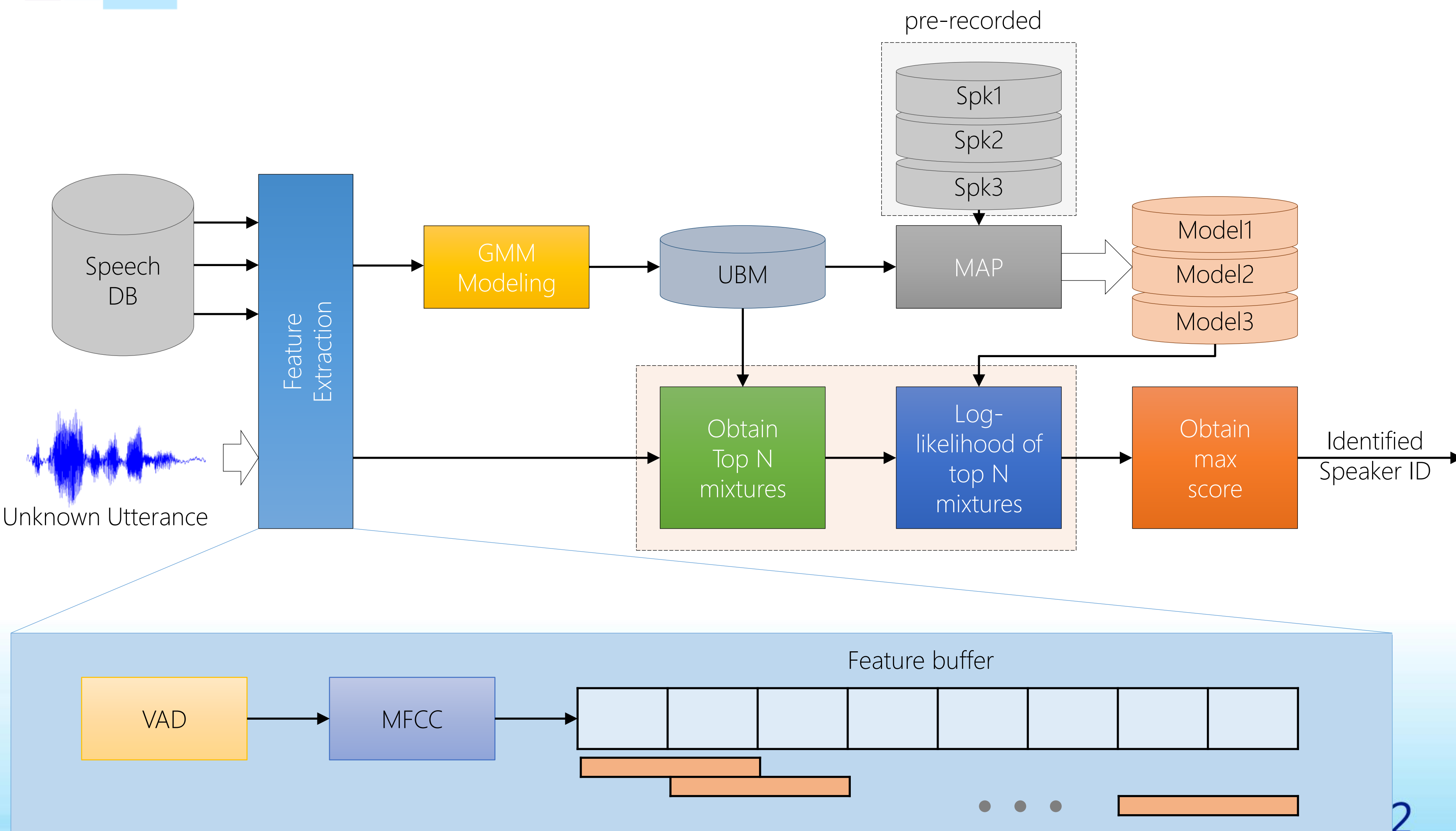
$$\tilde{c}_n = c_n - c^n_{cmn}$$

$$c^n_{cmn} = (1-\alpha)\cdot c^{n-1}_{cmn} + \alpha \cdot \tilde{c}_{n-1}$$





Training – GMM-UBM



Final Demonstration

Contents

Yonsei University DSP Lab.

- Overview : Research Overview
- Group 1 : Multiple User Localization
- Group 2 : Active Speaker Detection and Beamforming
- Group 3 : Speaker Identification
- Discussion

Discussion

- A/V localization and speaker identification: who is where?
  - Utilizes depth video stream
  - Is robust to environmental effects
  - Can be useful for multi-user applications
  - User friendly

- Possible future works
  - System level optimization
    - Audio and Video signal processing
  - Applications
    - Speech recognition, A/V communications

Microsoft Research

Thank you!

Microsoft Research Asia Faculty Summit 2012