

# The IIR-NTU Speaker Diarization Systems for RT 2009

Trung Hieu Nguyen<sup>[1,2]</sup>, Hanwu Sun<sup>[1]</sup>, ShengKui Zhao<sup>[2]</sup>,  
Swe Zin Kalayar Khine<sup>[1]</sup>, Huy Dat Tran<sup>[1]</sup>, Tin Lay Nwe Ma<sup>[1]</sup>,  
Bin Ma<sup>[1]</sup>, Eng Siong Chng<sup>[2]</sup>, Haizhou Li<sup>[1,2]</sup>

<sup>[1]</sup> Institute for Infocomm Research, Singapore  
<sup>[2]</sup> Nanyang Technological University, Singapore



**NANYANG  
TECHNOLOGICAL  
UNIVERSITY**



# Outline

---

- System structures
- Results
- Performance analysis
- Conclusions



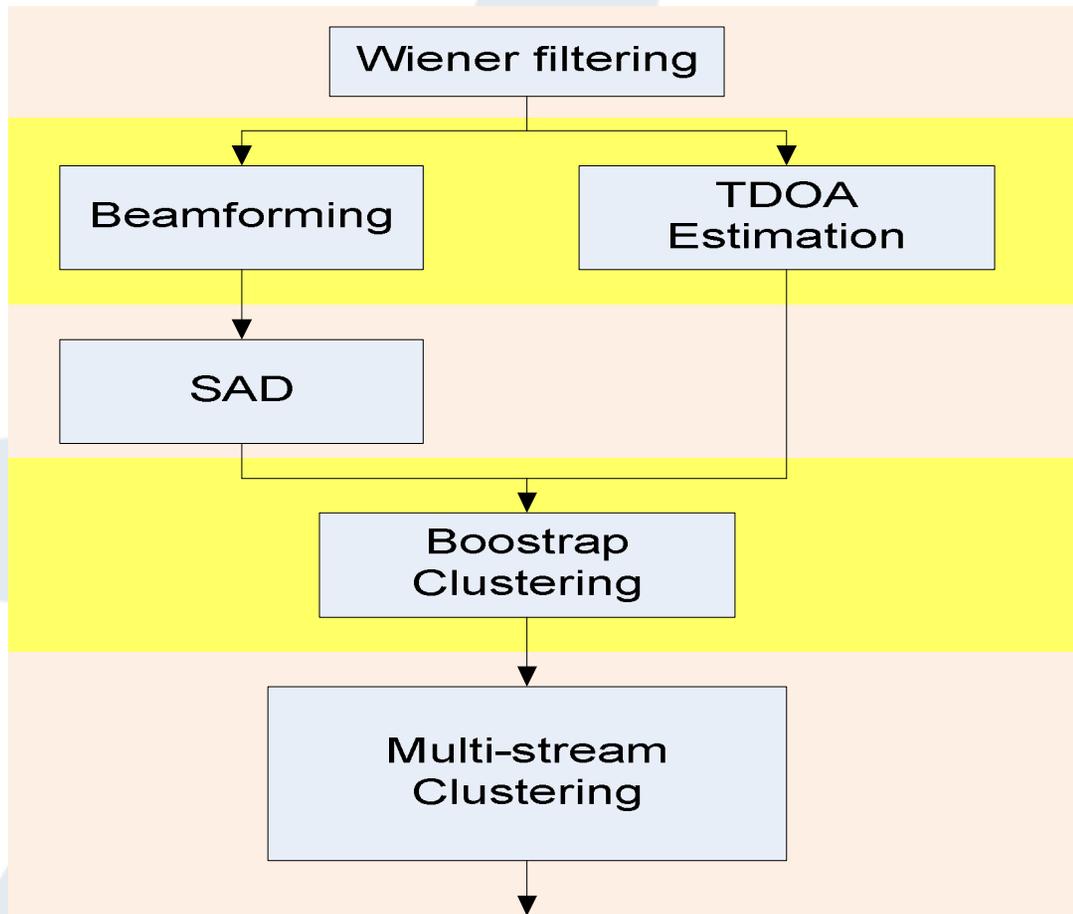
NANYANG  
TECHNOLOGICAL  
UNIVERSITY



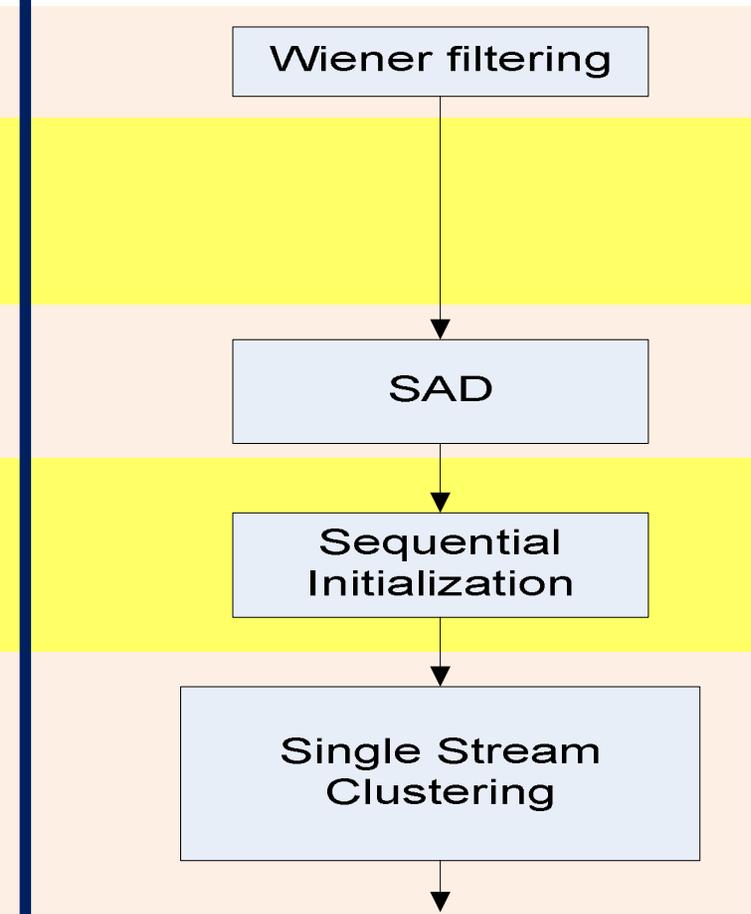
# System Structures



## MDM System



## SDM System





# Wiener Filtering

---

- Wiener filtering is applied to all audio channels for speech enhancement. The implementation of Wiener filter is from Qualcomm-ICSI-OGI front end [1].

[1] A. Adami, L. Burget, S. Dupont, H. Garudadri, F. Grezl, H. Hermansky, P. Jain, S. Kajarekar, N. Morgan, and S. Sivasdas, "Qualcomm-icsi-ogi features for asr," in Proc. ICSLP, vol. 1, 2002, pp. 4-7.



NANYANG  
TECHNOLOGICAL  
UNIVERSITY





# Beamforming

---

- The enhanced audio channels are then filtered and summed to produce a beamformed audio channel with the BeamformIt tool-kit [2].

[2] BeamformIt acoustic beamformer. [Online]. Available: <http://www.xavieranguera.com/beamformit/>

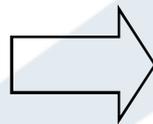
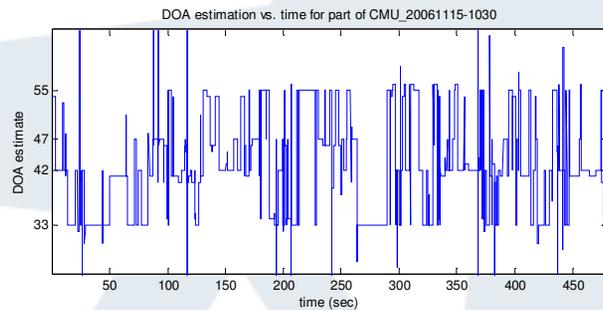


NANYANG  
TECHNOLOGICAL  
UNIVERSITY

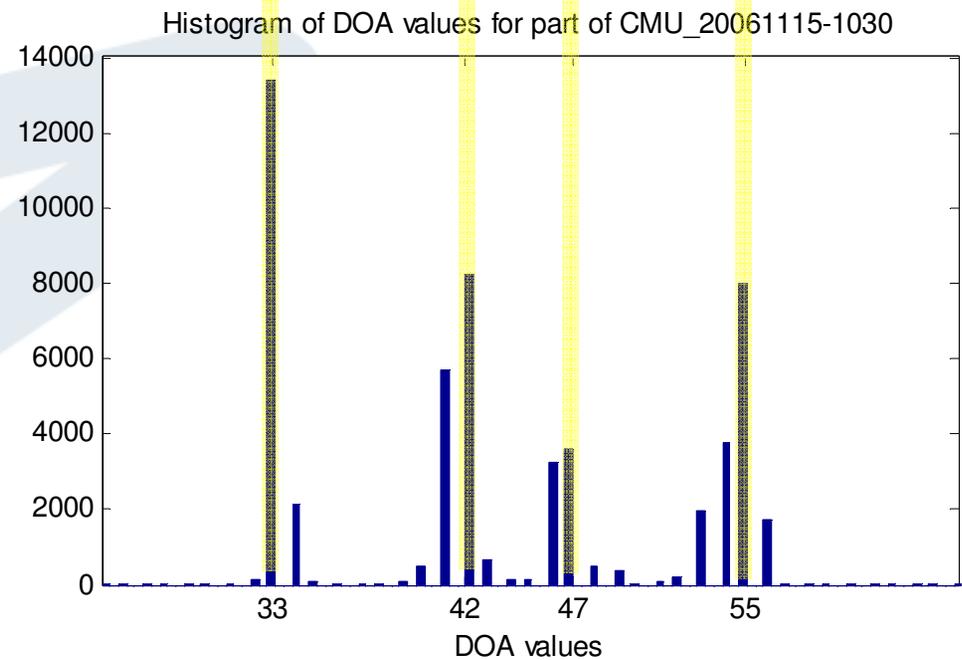


# TDOA Estimation (1/2) – Microphone Pairs Selection

- Compute the histogram of all microphone pairs



count



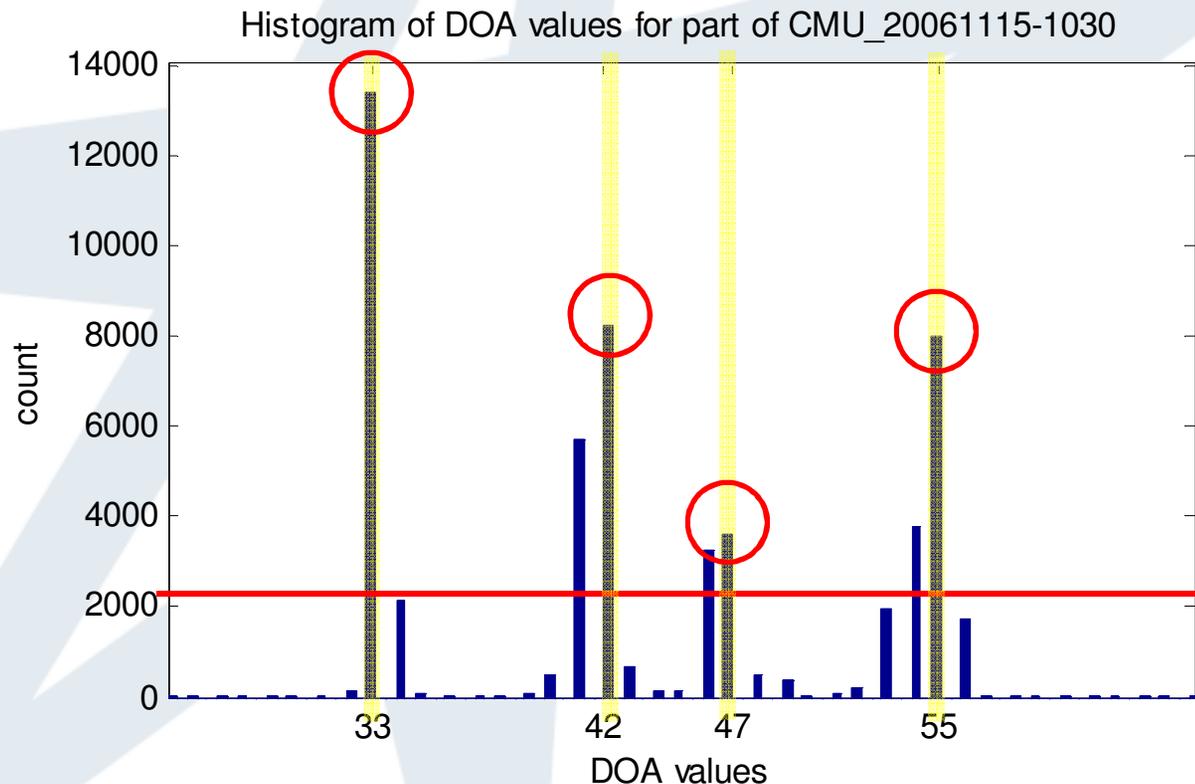
NANYANG  
TECHNOLOGICAL  
UNIVERSITY





# TDOA Estimation (2/2) – Microphone Pairs Selection

- Find the number of peaks in the histogram for the given threshold below
- Find top 6 microphone pairs with highest detected number peaks if the possible microphone pairs are greater than 6
- Use these 6 microphone pairs TDOA values for initial clustering.



peak detection threshold

$$= \frac{\text{total \# frames}}{\text{\# histogram bins}}$$



NANYANG  
TECHNOLOGICAL  
UNIVERSITY



# Speech Activity Detection

---

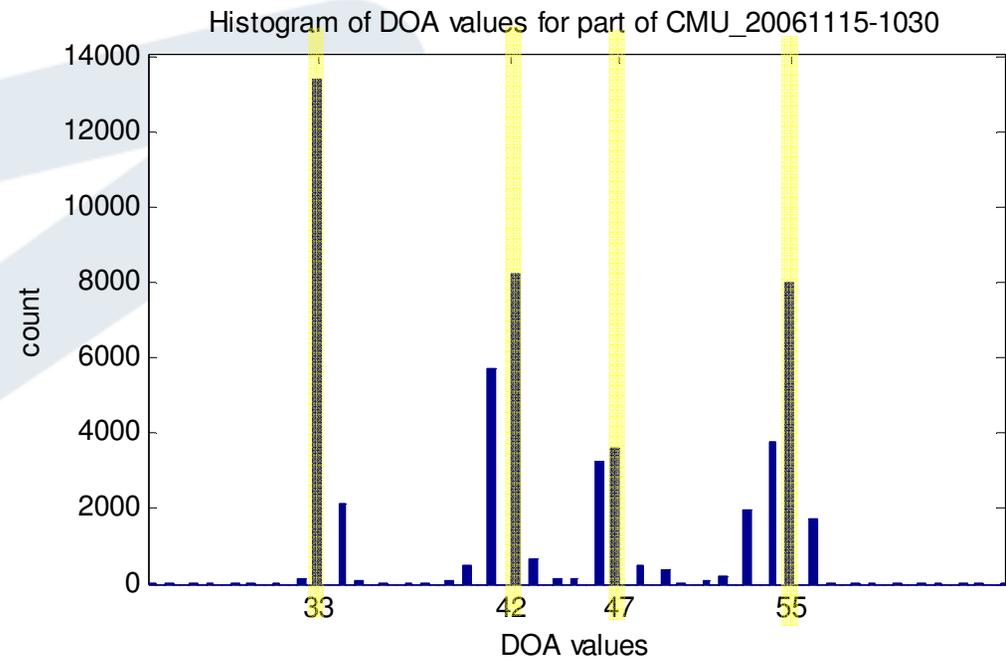
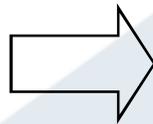
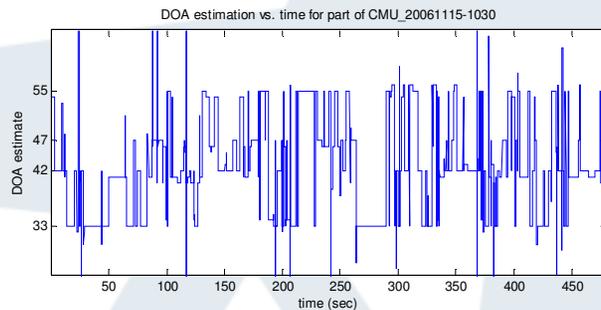


- a. Generate the 36 MFCC features (12 MFCC plus their first and second order derivatives) 30ms window with 20ms overlap
- b. Select 10% highest energy with relative high zero cross rates frames as the initial speech guessing.
- c. Select 20% lowest energy with relative low zero cross rates frames as the initial noise guessing
- d. Train initial speech model  $\lambda_S$  and Non-speech model  $\lambda_{NS}$  via EM, the speech and noise model sizes are set to be 16 and 4, respectively.
- e. Frame-wise maximum likelihood evaluation against  $\lambda_S$  and  $\lambda_{NS}$
- f. Retrain speech model  $\lambda_S$  and non-speech model  $\lambda_{NS}$  via MAP.
- g. Compute the speech/non-speech frame ratio, if the percentage change of the ratio (compared to its previous ratio) is less than 1%, stop, otherwise go to step e.



# Bootstrap Clustering (1/7) – Within Pair Quantization

- Constructs a histogram of TDOA values for each selected microphone pairs

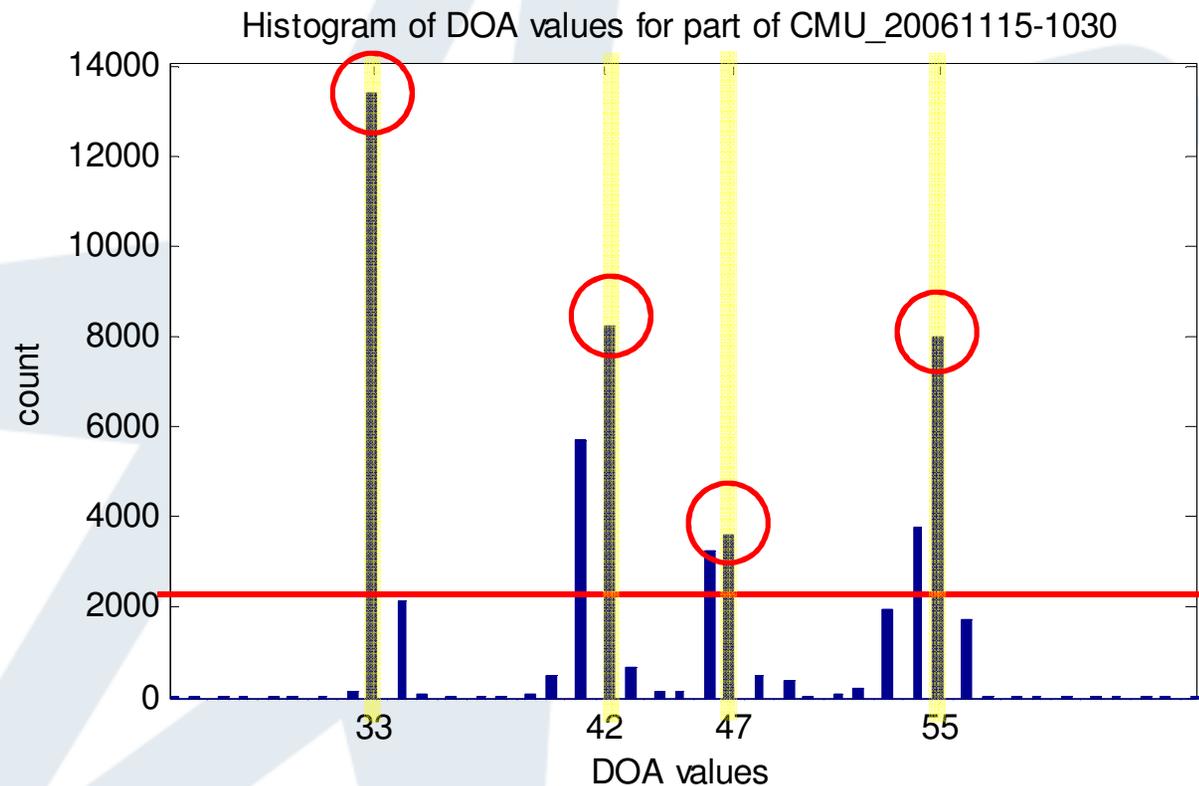


NANYANG  
TECHNOLOGICAL  
UNIVERSITY



# Bootstrap Clustering (2/7) – Within Pair Quantization

- Find the peaks in the histograms



peak detection threshold  
 $= \frac{\text{total \# frames}}{\text{\# histogram bins}}$



NANYANG  
TECHNOLOGICAL  
UNIVERSITY

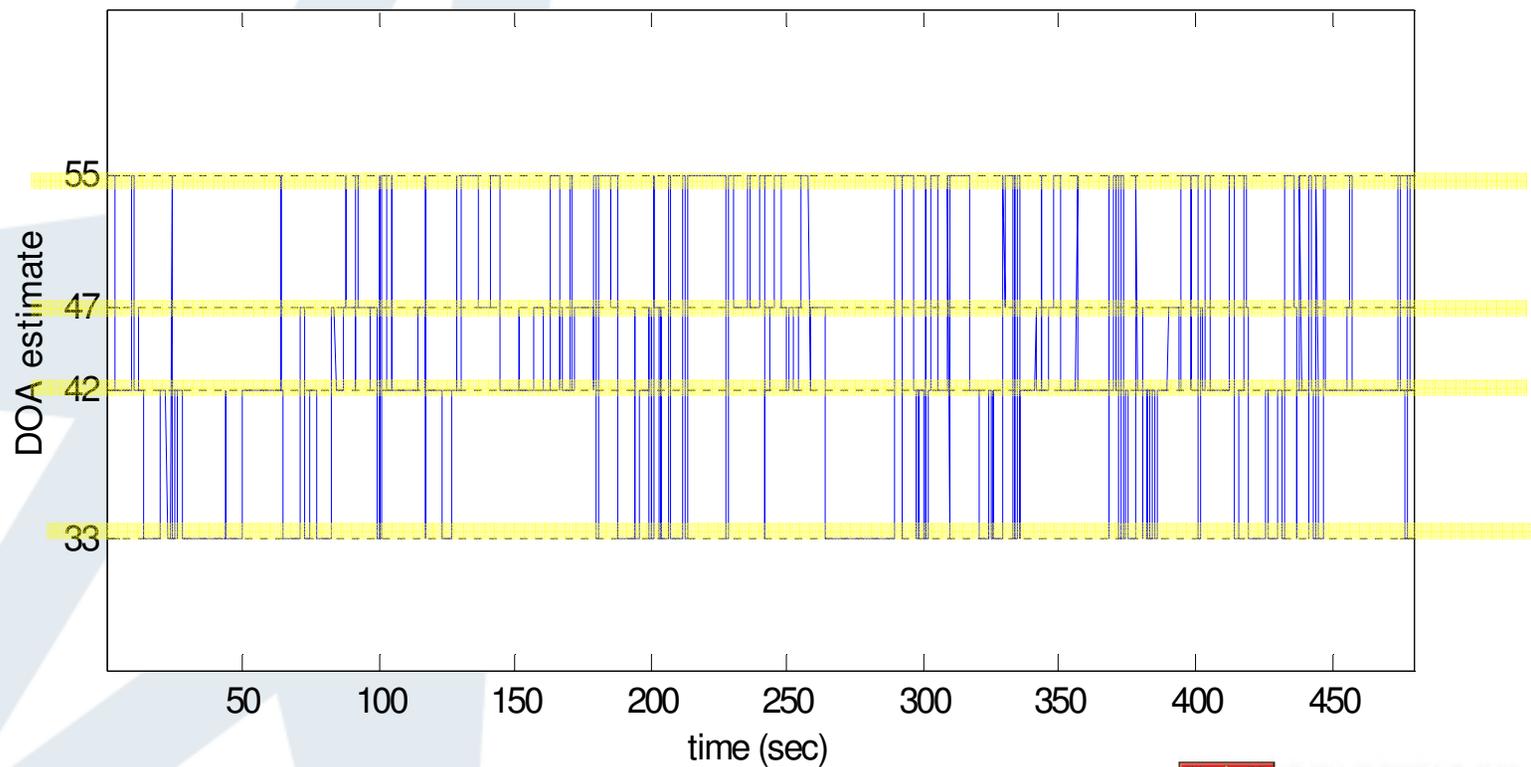


# Bootstrap Clustering (3/7) – Within Pair Quantization

---

- Identifies centroids in histogram

Quantized DOA estimation vs. time for part of CMU\_20061115-1030



NANYANG  
TECHNOLOGICAL  
UNIVERSITY

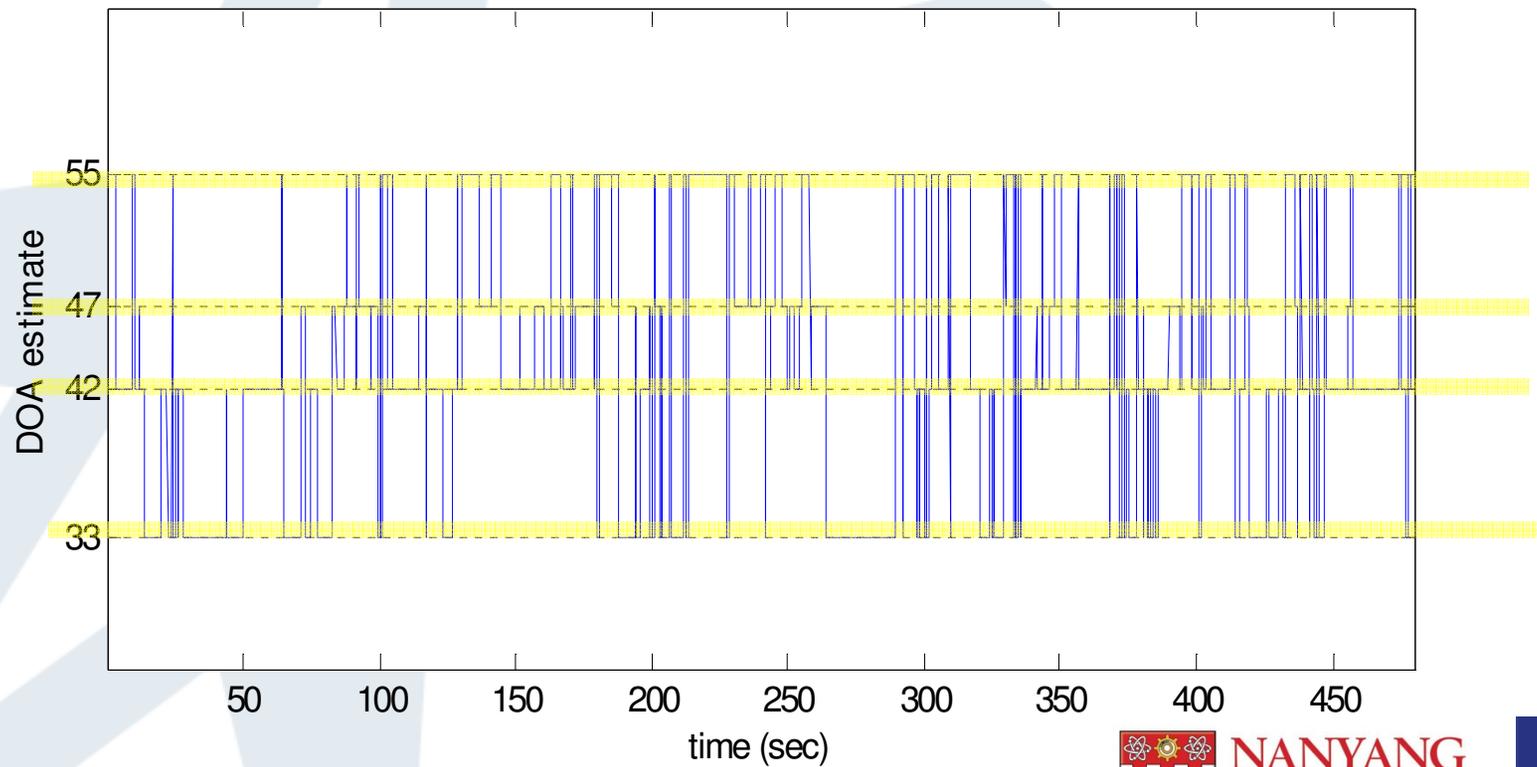


# Bootstrap Clustering (4/7) – Within Pair Quantization

---

- Maps other values to nearest centroid

Quantized DOA estimation vs. time for part of CMU\_20061115-1030

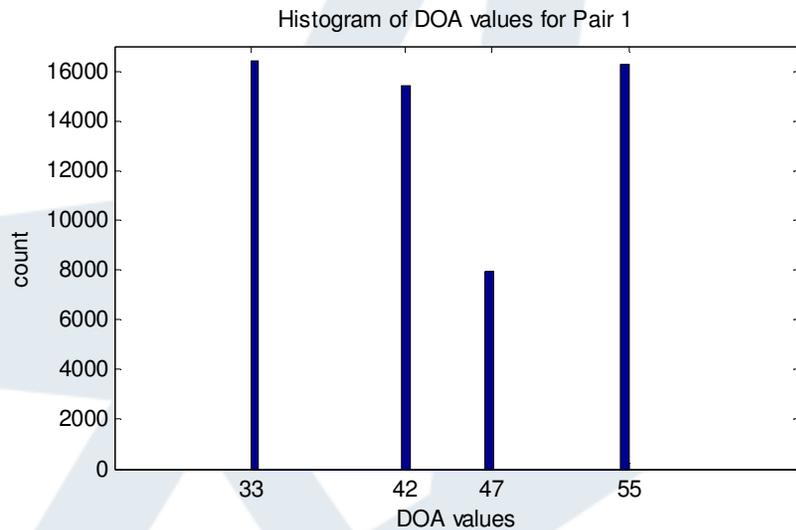


NANYANG  
TECHNOLOGICAL  
UNIVERSITY

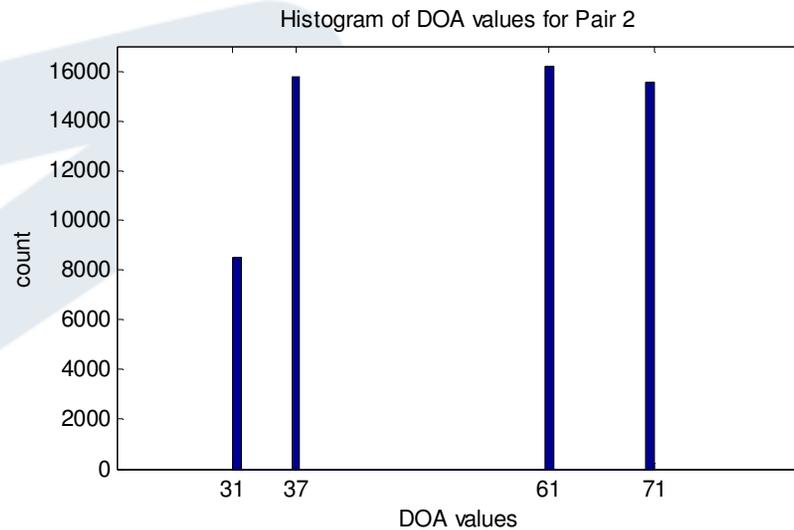


# Bootstrap Clustering (5/7) – Inter-Pair Quantization

- Using quantized TDOA from multiple microphone pairs



Pair 1: Mic 1 & 2



Pair 2: Mic 1 & 3

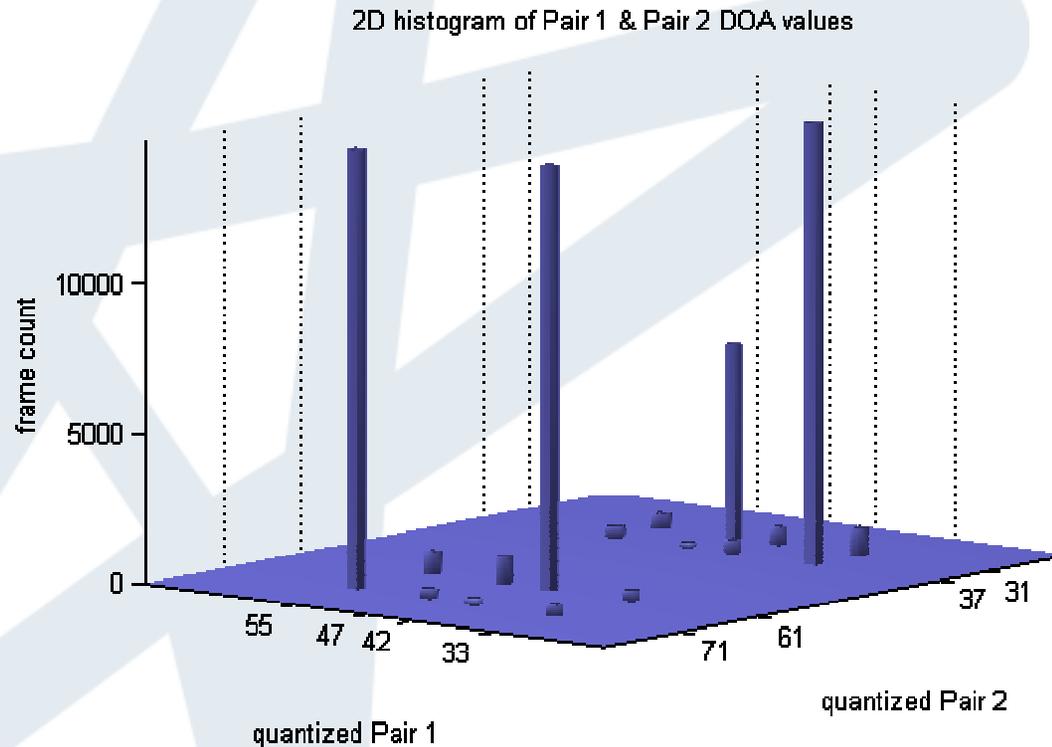


NANYANG  
TECHNOLOGICAL  
UNIVERSITY



# Bootstrap Clustering (6/7) – Inter-Pair Quantization

- Using quantized DOA from multiple microphone pairs
  - Constructs a multi-dimensional histogram
  - Identifies 9 centroids with highest bin count



# Bootstrap Clustering (7/7) – Inter-Pair Quantization

---



- Using quantized TDOA from multiple microphone pairs
  - Constructs a multi-dimensional histogram
  - Identifies 9 centroids with highest bin count (initial 9 clusters)
  - Merges all other bins to the nearest centroids
- Centroids after merging are given unique labels



NANYANG  
TECHNOLOGICAL  
UNIVERSITY



# Sequential Initialization

---



1. Extract LPCC with 19 coefficients using HTK Tool-kit. The audio data is then uniformly split to form 30 clusters.
2. Each cluster is modeled by a GMM with 4 mixtures.
3. Split each cluster into segments of 500 ms and only keep the top 25% of segments that best fit the cluster GMM and built the new cluster models using these segments. 75% of the segments are marked as unlabelled.
4. Classified the unlabelled segments into one of the clusters. However, only  $K$  segments are marked as classified. The classified segments and labeled segments are used to update the cluster GMMs. This step is repeated until all the unlabeled segments are marked as classified.
5. Run Viterbi decode to re-assign the cluster labels.
6. Step 2-5 are repeated 10 times.



# Multi Stream Clustering (1/3)

---

- Two feature streams for clustering: TDOA and LPCC 19
- TDOA features: GMM with 2 mixtures. LPCCs: GMM with 16 mixtures. Standard agglomerative hierarchical clustering algorithm with Viterbi re-segmentation as explained in [1].
- Modified version of CLR as the distance measure between clusters.
- The weighting algorithm is based on the Fisher distance.
- Stopping criterion: Ts criterion [2]

[1] C. Wooters and M. Huijbregts, "The ICSI RT07s speaker diarization system," Lecture Notes in Computer Science, vol. 4625, pp. 509-519, 2008.

[2] Trung Hieu Nguyen, Eng Siong Chng, Haizhou Li, "T-Test Distance and Clustering Criterion for Speaker Diarization", Interspeech, 2008.



NANYANG  
TECHNOLOGICAL  
UNIVERSITY



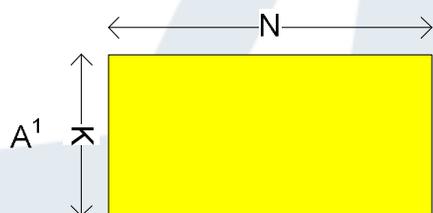
# Multi Stream Clustering (2/3) – Automatic Stream Weight Selection

Suppose that there are  $K$  clusters and  $N$  frames. For each feature stream, train a GMM for each cluster.

$$\lambda_{11}, \lambda_{12}, \dots, \lambda_{1K}$$

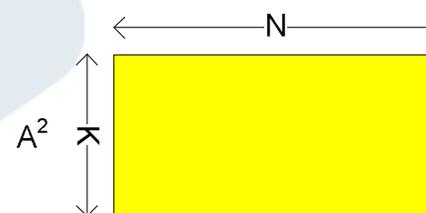
$$\lambda_{21}, \lambda_{22}, \dots, \lambda_{2K}$$

Compute the likelihood scores for each frame given each model



$$1 \leq i \leq K$$

$$1 \leq j \leq N$$



$$A_{ij}^1 = \log(P(u_{1j} | \lambda_{1i})) - \log(P(u_{1j} | \lambda_{1UBM}))$$

$$A_{ij}^2 = \log(P(u_{2j} | \lambda_{2i})) - \log(P(u_{2j} | \lambda_{2UBM}))$$

$$A = w_1 \times A^1 + w_2 \times A^2$$

$$S_1 = \{A_{ij} | \forall i, j: \text{frame } j \in \text{cluster } i\}$$

$$S_2 = \{A_{ij} | \forall i, j: \text{frame } j \notin \text{cluster } i\}$$

$w_1$  and  $w_2$  are tuned to maximize the Fisher distance between  $S_1$  and  $S_2$

# Multi Stream Clustering (3/3) – Cluster Distance Measure

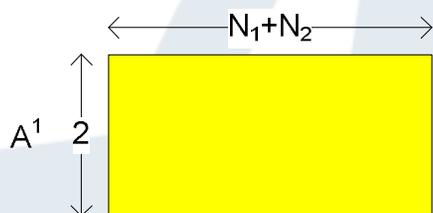


Given 2 clusters  $C_1, C_2$  with the number of frames are  $N_1$  and  $N_2$  respectively. For each feature stream, train a GMM for each cluster.

$$\lambda_{11}, \lambda_{12}$$

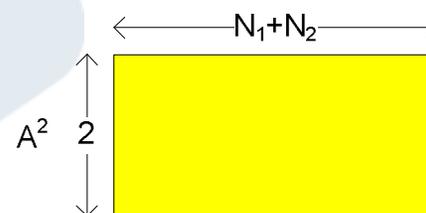
$$\lambda_{21}, \lambda_{22}$$

Compute the likelihood scores for each frame given each model



$$1 \leq i \leq 2$$

$$1 \leq j \leq N_1 + N_2$$



$$A_{ij}^1 = \log(P(u_{1j} | \lambda_{1i})) - \log(P(u_{1j} | \lambda_{1UBM}))$$

$$A_{ij}^2 = \log(P(u_{2j} | \lambda_{2i})) - \log(P(u_{2j} | \lambda_{2UBM}))$$

$$A = w_1 \times A^1 + w_2 \times A^2$$

$$S_1 = \{A_{ij} | \forall i, j: \text{frame } j \in \text{cluster } i\}$$

$$S_2 = \{A_{ij} | \forall i, j: \text{frame } j \notin \text{cluster } i\}$$

$$d(C_1, C_2) = \frac{|\text{mean}(S_1) - \text{mean}(S_2)| \sqrt{N_1 + N_2}}{\sqrt{\text{var}(S_1) + \text{var}(S_2)}}$$

# Single Stream Clustering (1/2)

---

- Feature: LPCC 19
- Each cluster is modeled by a GMM with 16 mixtures.
- Standard agglomerative hierarchical clustering algorithm with Viterbi re-segmentation as explained in [1].
- Modified version of CLR as the distance measure between clusters.
- Stopping criterion: Ts criterion [2]

[1] C. Wooters and M. Huijbregts, "The ICSI RT07s speaker diarization system," Lecture Notes in Computer Science, vol. 4625, pp. 509-519, 2008.

[2] Trung Hieu Nguyen, Eng Siong Chng, Haizhou Li, "T-Test Distance and Clustering Criterion for Speaker Diarization", Interspeech, 2008.



NANYANG  
TECHNOLOGICAL  
UNIVERSITY



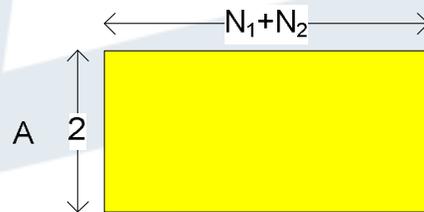
# Single Stream Clustering (2/2) – Cluster Distance Measure



Given 2 clusters  $C_1, C_2$  with the number of frames are  $N_1$  and  $N_2$  respectively. Train a GMM for each cluster

$$\lambda_1, \lambda_2$$

Compute the likelihood scores for each frame given each model



$$A_{ij} = \log \left( P(u_j | \lambda_i) \right) - \log \left( P(u_j | \lambda_{UBM}) \right)$$

$$1 \leq i \leq 2$$
$$1 \leq j \leq N_1 + N_2$$

$$S_1 = \{A_{ij} | \forall i, j: \text{frame } j \in \text{cluster } i\}$$

$$S_2 = \{A_{ij} | \forall i, j: \text{frame } j \notin \text{cluster } i\}$$

$$d(C_1, C_2) = \frac{|mean(S_1) - mean(S_2)| \sqrt{N_1 + N_2}}{\sqrt{var(S_1) + var(S_2)}}$$

# Results – RT 2009 Evaluation

---



Mic. Condition	Overlap SPKR Err.	Non-Overlap SPKR Err.	SAD
MDM	9.21	3.83	2.74
SDM	16.04	10.72	2.57



NANYANG  
TECHNOLOGICAL  
UNIVERSITY



# Performance Analysis (1/6) – SDM Systems

---

	System A	System B	System C	System D	System E
Feature	MFCC 19	MFCC 19	LPCC 19	LPCC 19	LPCC 19
Duration Constraint	2s	0.5s with word insertion penalty of -150			
Cluster Distance Measure	GLR	GLR	GLR	Proposed	Proposed
Initialization	Uniform	Uniform	Uniform	Uniform	Sequential



**NANYANG**  
TECHNOLOGICAL  
UNIVERSITY



# Performance Analysis (2/6) – SDM Systems

---

	RT 2005	RT 2006	RT 2007
SAD (%)	3.00	4.50	3.01

	System A	System B	System C	System D	System E
RT 2005	15.57	13.00	11.95	10.88	12.08
RT 2006	27.08	20.68	17.04	18.78	19.63
RT 2007	16.96	13.54	13.25	11.94	10.92

DERs of various SDM systems with optimal stopping criterion



**NANYANG  
TECHNOLOGICAL  
UNIVERSITY**



# Performance Analysis (3/6) – Stopping Criteria

---

	<b>RT 2005</b>	<b>RT 2006</b>	<b>RT 2007</b>
DER(%)	9.94	15.21	11.78
With optimal stopping criterion			

	<b>RT 2005</b>	<b>RT 2006</b>	<b>RT 2007</b>
<b>RT 2005</b>	13.35	18.75	18.24
<b>RT 2006</b>	13.35	18.75	18.24
<b>RT 2007</b>	14.62	21.56	13.45

**With GLR threshold as stopping criterion (DER)**



**NANYANG  
TECHNOLOGICAL  
UNIVERSITY**



# Performance Analysis (4/6) – Stopping Criteria

---

	RT 2005	RT 2006	RT 2007
RT 2005	12.47	28.73	25.22
RT 2006	22.78	22.94	28.23
RT 2007	17.58	35.57	20.64

With ICR threshold as stopping criterion (DER)

	RT 2005	RT 2006	RT 2007
RT 2005	16.58	30.19	24.93
RT 2006	17.03	20.4	24.00
RT 2007	18.88	26.94	17.73

With BIC threshold as stopping criterion (DER)



**NANYANG  
TECHNOLOGICAL  
UNIVERSITY**



# Performance Analysis (5/6) – Stopping Criteria

---

RT 2005	RT 2006	RT 2007
13.90	19.21	21.80
With $T_s$ [1] threshold as stopping criterion (DER)		

RT 2005	RT 2006	RT 2007
12.56	19.91	16.67
With modified $T_s$ threshold as stopping criterion (DER)		

[1] Trung Hieu Nguyen , Eng Siong Chng, Haizhou Li, “T-Test Distance and Clustering Criterion for Speaker Diarization”, Interspeech, 2008.



NANYANG  
TECHNOLOGICAL  
UNIVERSITY



# Performance Analysis (6/6) – MDM Systems



	RT 2005	RT 2006	RT 2007
Stream Weight Selection using [1]	9.08	14.37	9.59
Proposed Stream Weight Selection	8.93	12.94	9.30

**DERs for MDM systems with different weighting schemes**

[1] Anguera, X., Wooters, C., Pardo, J. M. and Hernando, J., *Automatic Weighting for the Combination of TDOA and Acoustic Features in Speaker Diarization for Meetings*, in: Proc. ICASSP, Honolulu, 2007.



**NANYANG  
TECHNOLOGICAL  
UNIVERSITY**





# Conclusions

---

- Shorter duration constraint seems to improve the performance.
- It is not conclusive which acoustic features are good for speaker diarization.
- MDM system is much better than SDM system, largely due to very good initialization and correct stopping point.
- Performance of sequential initialization is not good. Need more investigations.



NANYANG  
TECHNOLOGICAL  
UNIVERSITY





**THANK YOU !**



**NANYANG  
TECHNOLOGICAL  
UNIVERSITY**

