The I²R-NTU submission for NIST RT-07 Conference Room Recording Speaker Diarization

Trung Hieu Nguyen[2], Hanwu Sun[1], Tin Lay Nwe[1], Eugene Koh[2], Bin Ma[1], Eng-Siong Chng[2], Haizhou Li[1], Susanto Rahardja[1]

[1] Institute for Infocomm Research(I²R), Singapore
[2] Nanyang Technological University(NTU), Singapore

Our Task

- Speaker Diarization for Conference Room using MDM

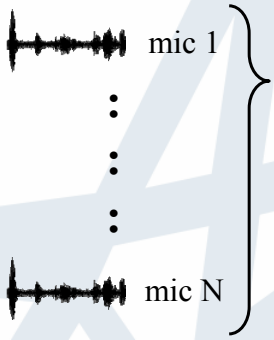
Speech-to-text (STT)	**Speaker Diarization (SPKR)**	Speaker Attributed Speech-To-Text (SASTT)
Conference Room	Lecture Room	Coffee Break
Single Distant Microphone (SDM)	**Multiple Distant Microphone (MDM)**	

Outline

- Our Primary System
  - Direction of Arrival (DOA) Estimation
  - Bootstrap Clustering
  - Cluster Purification
  - Speech Activity Detection
- Results for RT-07
- Conclusions

System Overview

Our Primary System

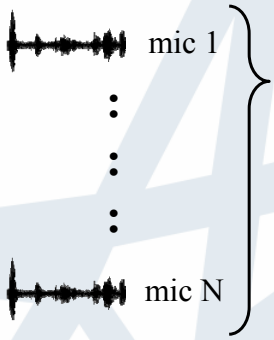


1. DOA estimation
2. Bootstrap clustering
3. Cluster purification
4. Speech activity detection

Meeting diarization

- 4 modules

DOA Estimation

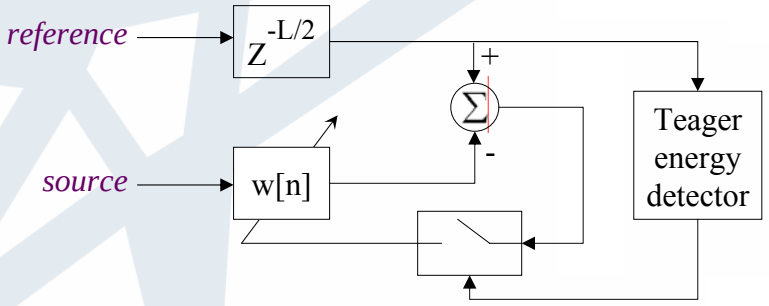


1. DOA estimation
2. Bootstrap clustering
3. Cluster purification
4. Speech activity detection

Meeting diarization

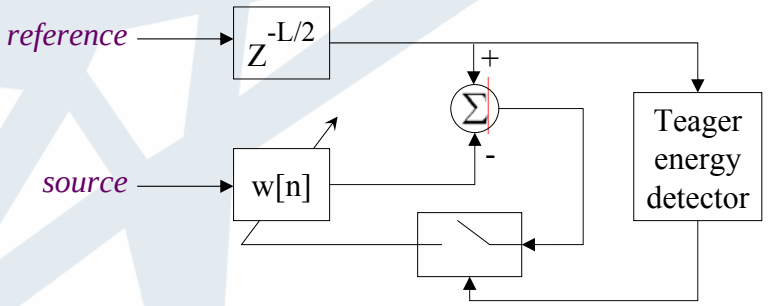
DOA Estimation

- Estimated by adapting a filter using Normalized Least-Mean Square (NLMS)
- For each microphone pair, one designated as *reference*, the other *source*



DOA Estimation

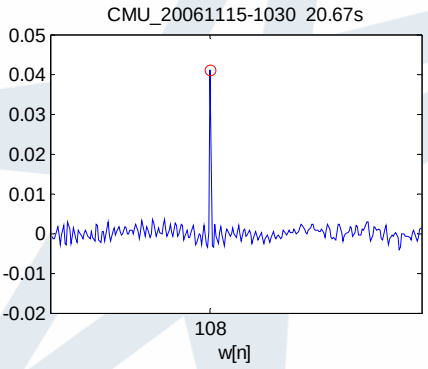
- *reference* delayed by L/2, L = filter length
- w[n] adapted only when *reference* has sufficient energy



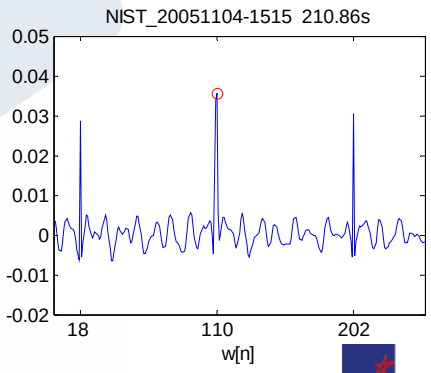
DOA Estimation

- Delay estimated by locating highest peaks in w[n]

Single speaker at 108



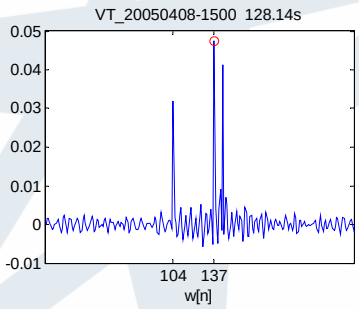
3 speakers at 18, 110 and 202



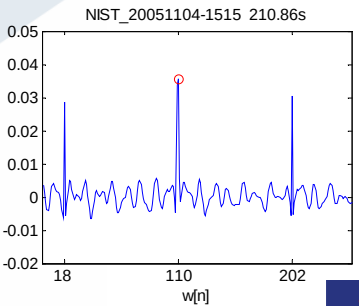
DOA Estimation

- Reverberations / Multiple speakers leads to multiple peaks
  - Choppiness means choosing the maximum is not always correct

Reverberant segment



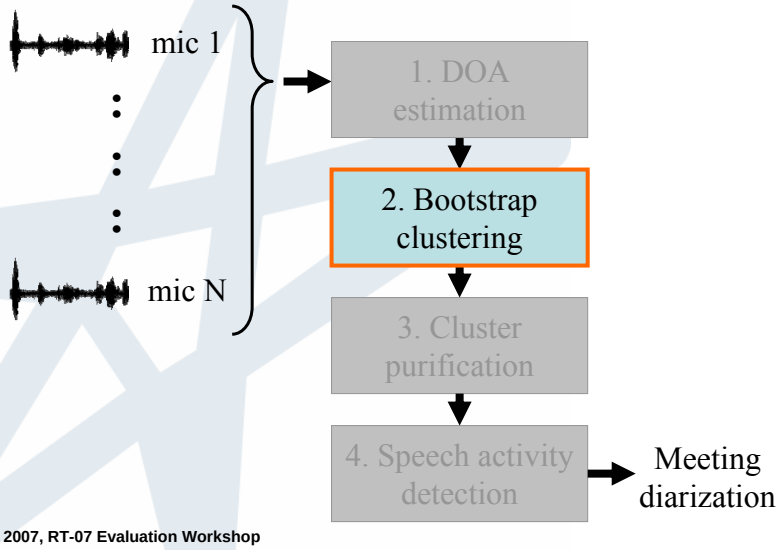
3 simultaneous speakers



DOA Estimation

- Microphone pair selection criteria – retain when
  - 1. Signal has high SNR
  - 2. w[n] has large average highest-peak to next-highest-peak ratio
  - 3. w[n] has large overall dynamic range

Bootstrap Clustering

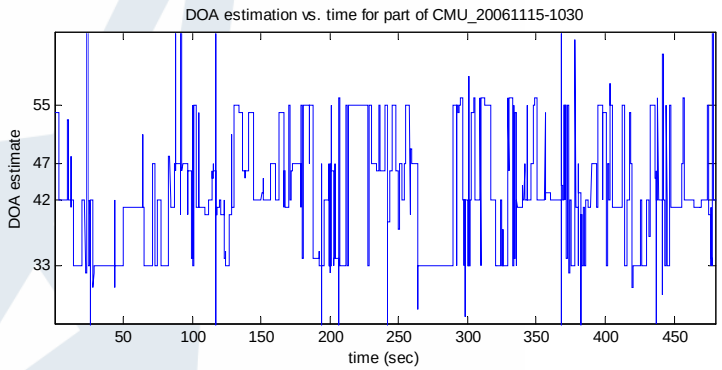


Bootstrap Clustering

- 2 steps
  - Within-pair quantization
  - Inter-pair quantization

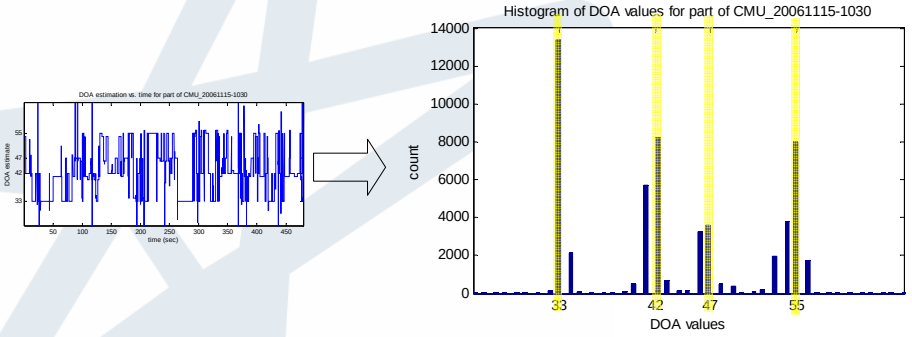
Within-pair Quantization

- Using DOA estimates



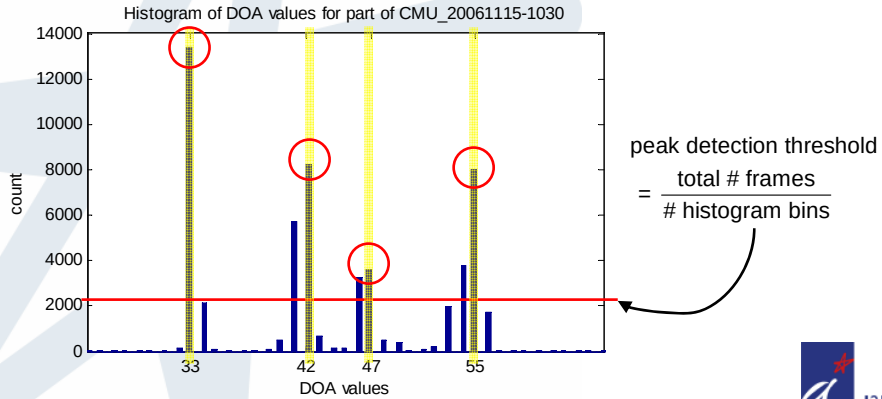
Within-pair Quantization

- Using DOA estimates
  - To construct a histogram of DOA values



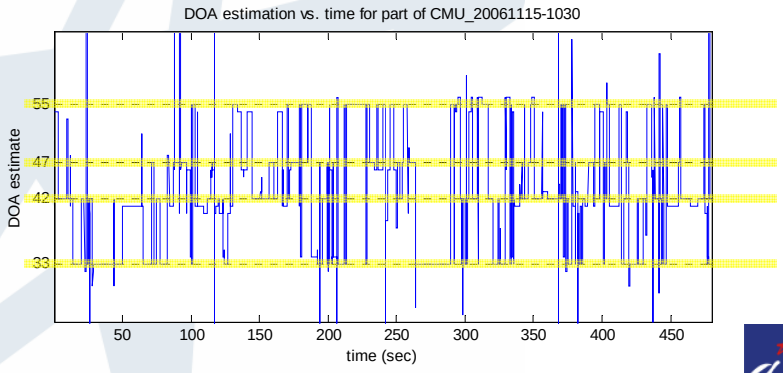
Within-pair Quantization

- Using DOA estimates
  - To construct a histogram of DOA values
  - Identify centroids in histogram



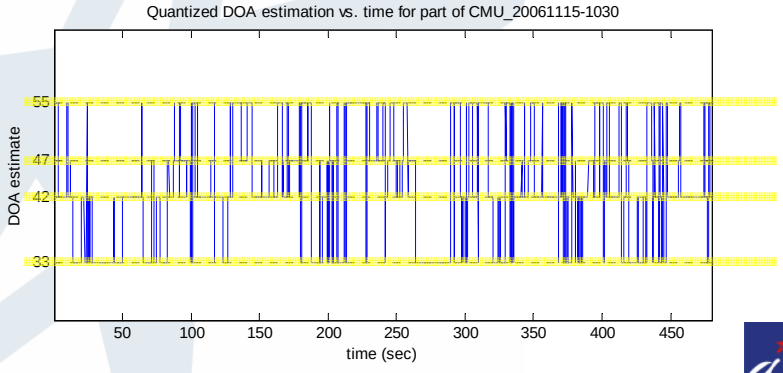
Within-pair Quantization

- Using DOA estimates
  - To construct a histogram of DOA values
  - Identify centroids in histogram



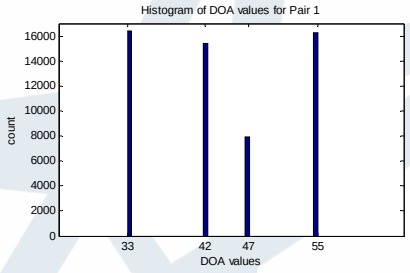
Within-pair Quantization

- Using DOA estimates
  - To construct a histogram of DOA values
  - Identify centroids in histogram
  - Map other values to nearest centroid

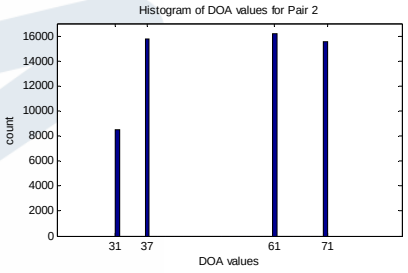


Inter-pair Quantization

- Using quantized DOA from multiple microphone pairs



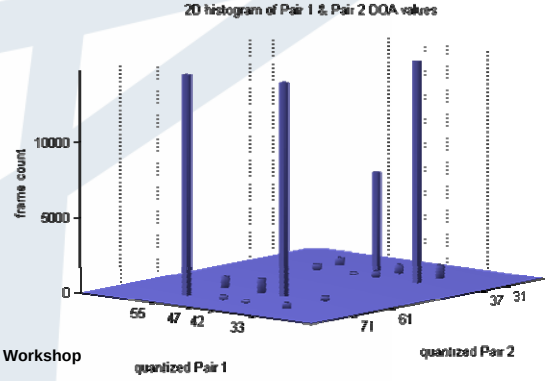
Pair A: Mic 1 & 2



Pair B: Mic 1 & 3

Inter-pair Quantization

- Using quantized DOA from multiple microphone pairs
  - Construct a multi-dimensional histogram
  - Identify centroids with high bin count

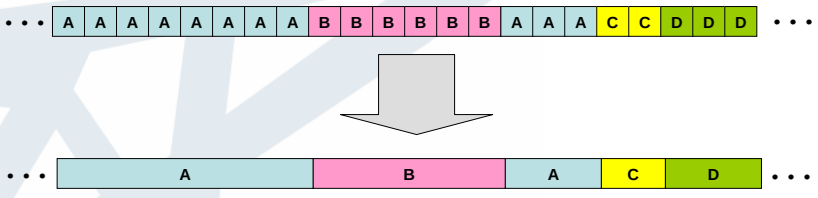


Inter-pair Quantization

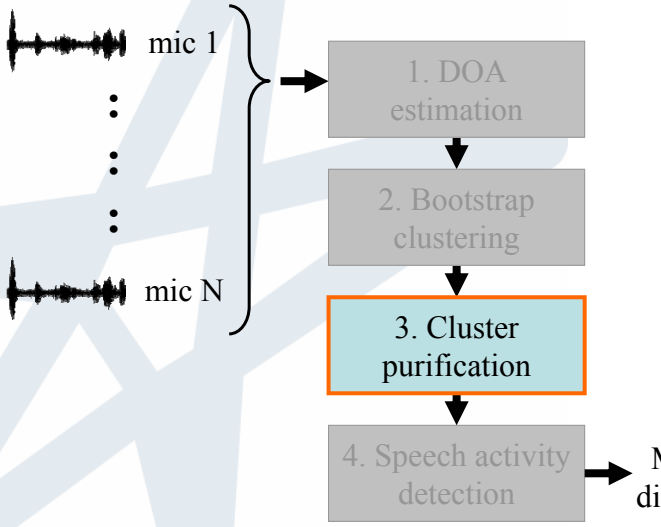
- Using quantized DOA from multiple microphone pairs
  - Construct a multi-dimensional histogram
  - Identify centroids with high bin count
  - Merge all other bins to nearest centroid
- Centroids after merging are given unique labels

Segment Formation

- Using labels resulting from both quantization steps
  - Contiguous frames with same labels are grouped to form our segments



Cluster Purification

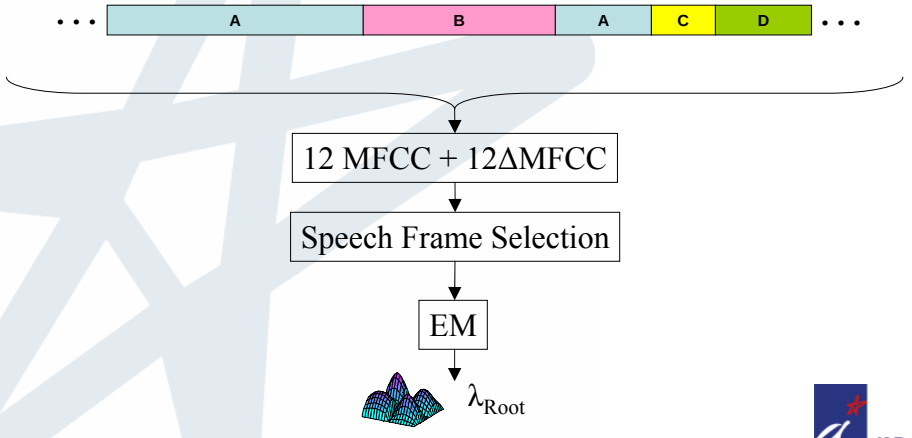


Cluster Purification

- Beamforming using GCC-PHAT
  - Delay & Sum of distant microphones
  - For EDI, only 1st array used
  - Yields single enhanced signal for feature extraction

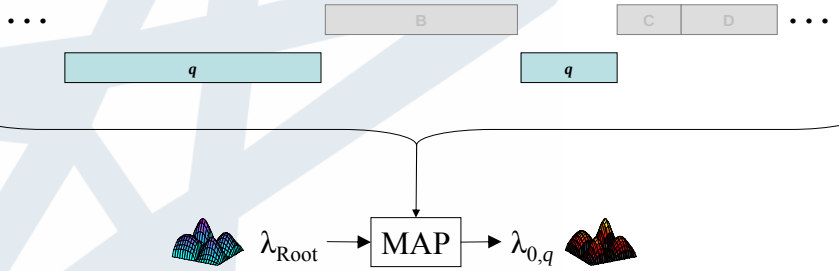
Cluster Purification

Step 1: All segments are used to train λ_{Root}



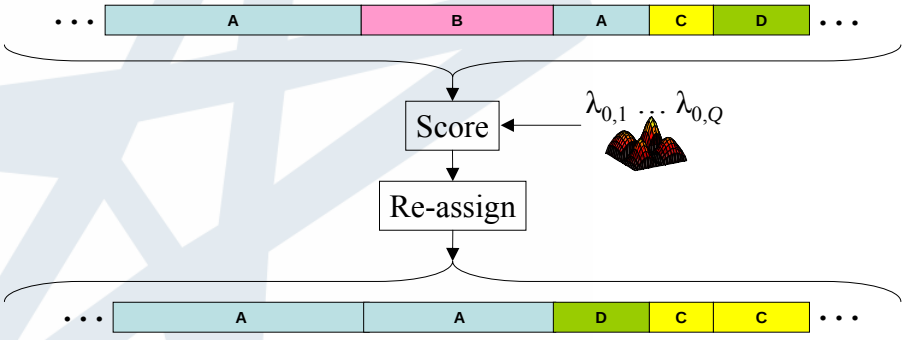
Cluster Purification

Step 2: For each cluster $q = 1..Q$,Q = number of clusters, GMM $\lambda_{0,q}$ is adapted from λ_{Root}



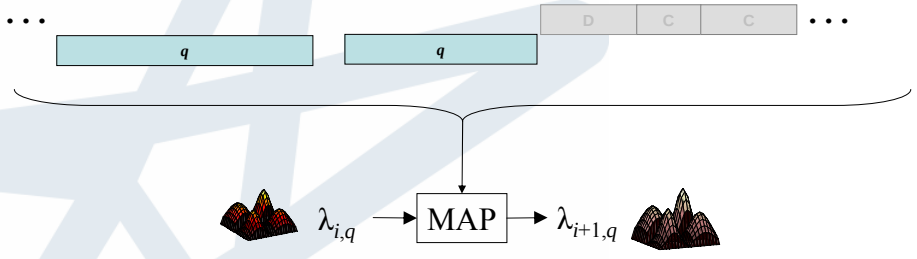
Cluster Purification

Step 3: All segments are scored against $\lambda_{i,1} \ldots \lambda_{i,Q}$, $i = 0$ for initial iteration. Segments are then re-assigned to the cluster in which it scored highest.



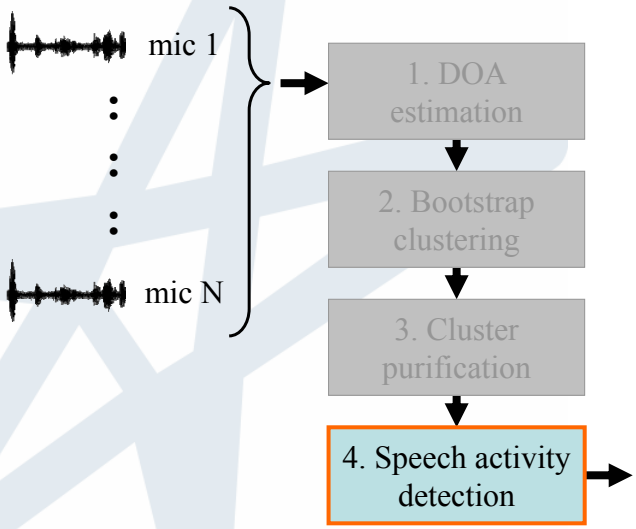
Cluster Purification

Step 4: For $q = 1..Q$, re-assigned clusters are used to adapt $\lambda_{i+1,q}$ from $\lambda_{i,q}$.



Step 5: $i = i + 1$. Repeat from Step 3. Stop when assignment stabilizes

Speech Activity Detection

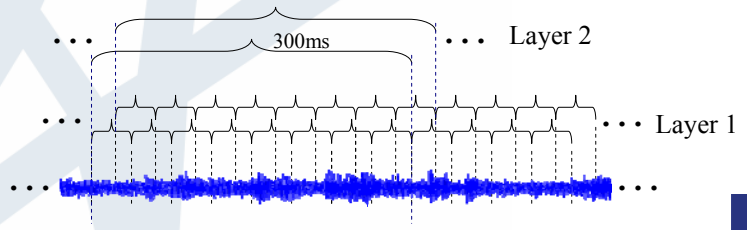


Speech Activity Detection: Non-speech Removal

- 10 Log Frequency Power Coefficients (LFPC) for 20ms window with 10ms overlap
- Speech model λ_S and Non-speech model λ_{NS} trained on ICSI meeting room corpus
- Segment-wise maximum likelihood evaluation against λ_S and λ_{NS}

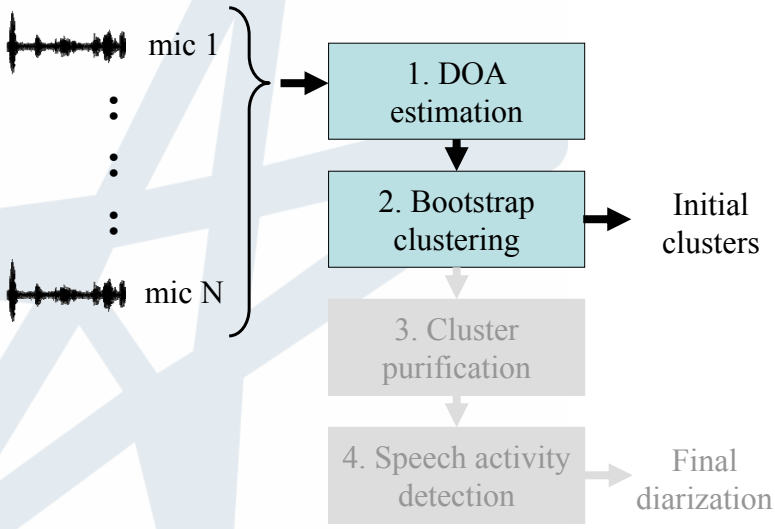
Speech Activity Detection: Silence Removal

- Double Layer Windowing (DLW) method
- 1st Layer
  - Energy calculated for 20ms frames with 10ms overlap
- 2nd Layer
  - Energy summed across 300ms windows at 10ms steps
  - Catches silences > 300ms tolerance



Evaluation Results

Results for Initial Clustering

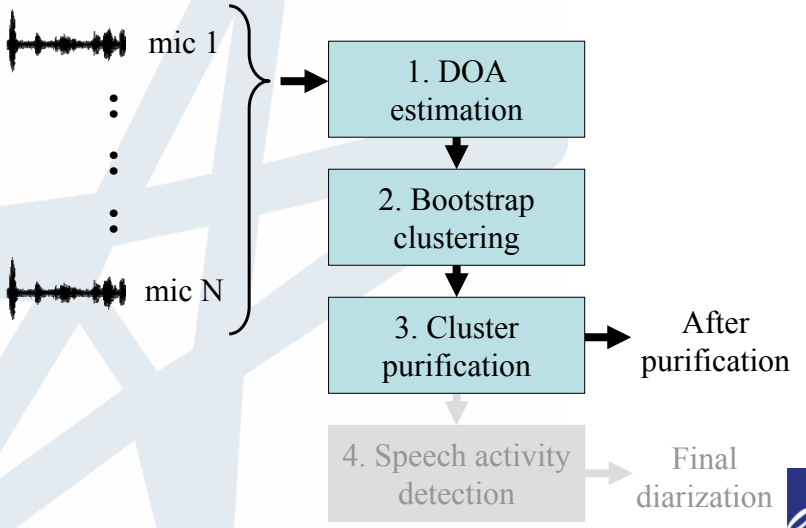


Results for Initial Clustering

	DER(%)
CMU_20061115-1030	22.75
CMU_20061115-1530	17.81
EDI_20061113-1500	24.29
EDI_20061114-1500	30.59
NIST_20051104-1515	23.34
NIST_20060216-1347	22.13
VT_20050408-1500	46.38
VT_20050425-1000	27.36
OVERALL	27.02

- DOA is insufficient for some tasks, especially VT_20050408-1500

After Cluster Purification

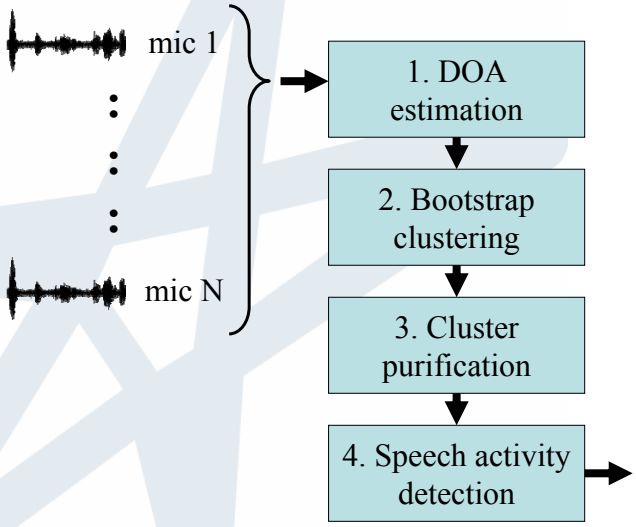


After Cluster Purification

	DER (%)	Absolute Δ DER (%)	Δ Speaker Error Time (s)
CMU_20061115-1030	22.74	-0.01	0
CMU_20061115-1530	17.57	-0.24	-1
EDI_20061113-1500	24.29	0.00	0
EDI_20061114-1500	30.18	-0.41	-3
NIST_20051104-1515	22.80	-0.54	-4
NIST_20060216-1347	18.49	-3.64	-27
VT_20050408-1500	19.77	-26.61	-206
VT_20050425-1000	27.12	-0.24	-2
OVERALL	22.76	-4.26	-243

- DER shows overall improvements
- Errors in VT_20050408-1500 are redeemed

Final Results



Final Results – After Speech Activity Detection

	DER (%)	Absolute Δ DER (%)	Δ F. Alarm Speaker Time (s)
CMU_20061115-1030	19.36	-3.38	-37.94
CMU_20061115-1530	12.46	-5.11	-55.23
EDI_20061113-1500	20.69	-3.60	-28.37
EDI_20061114-1500	15.00	-15.18	-141.47
NIST_20051104-1515	12.66	-10.14	-114.77
NIST_20060216-1347	13.36	-5.13	-52.97
VT_20050408-1500	11.32	-8.45	-89.13
VT_20050425-1000	18.45	-8.67	-62.32
OVERALL	15.32	-7.44	-582.19

- Speech Activity Detection further improves DER
- Improvements due to big lowering of False Alarm Speaker time

Final Results – After Speech Activity Detection

	SAD SPKR Error (%) Before	SAD SPKR Error (%) After	Δ SAD SPKR Error (%)
CMU_20061115-1030	9.91	7.62	-2.29
CMU_20061115-1530	14.11	9.91	-4.20
EDI_20061113-1500	9.59	7.82	-1.77
EDI_20061114-1500	26.79	10.57	-16.22
NIST_20051104-1515	14.31	7.32	-6.99
NIST_20060216-1347	11.46	7.50	-3.96
VT_20050408-1500	15.05	7.60	-7.45
VT_20050425-1000	16.67	11.37	-5.30
OVERALL	14.45	8.65	-5.80

However:

- SAD overall error is still high at 8.65%
- There's still non-speech /silence to be removed
- If SAD SPKR error could be reduced, overall DER will also reduce

Conclusions

Conclusions

- Using DOA to do segmentation + clustering is effective
- Clustering purification helps, especially where DOA alone is not effective
- Silence/Non-speech removal still has room for improvement
- Potential remains for DOA estimation of reverberant/overlapping speech

I²R A*STAR

Thank You

I²R A*STAR