Frontier of Frontend for Conversational Speech Processing

Shoko Araki

Signal Processing Research Group
NTT Communication Science Laboratories
NTT Corporation, Japan

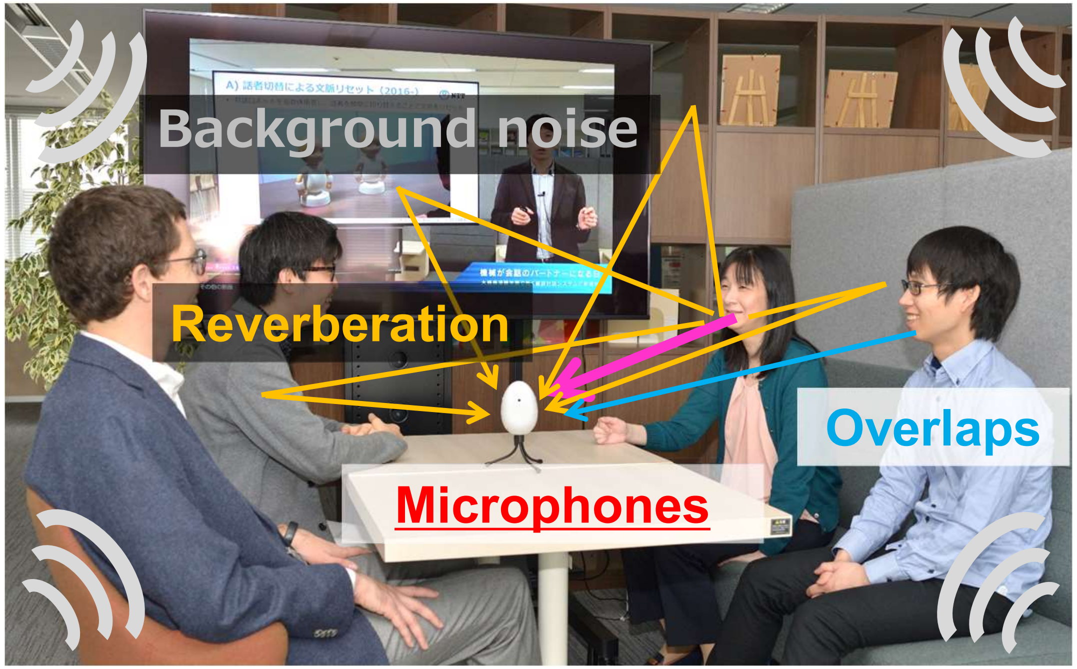
Enriching everyday conversation

NTT

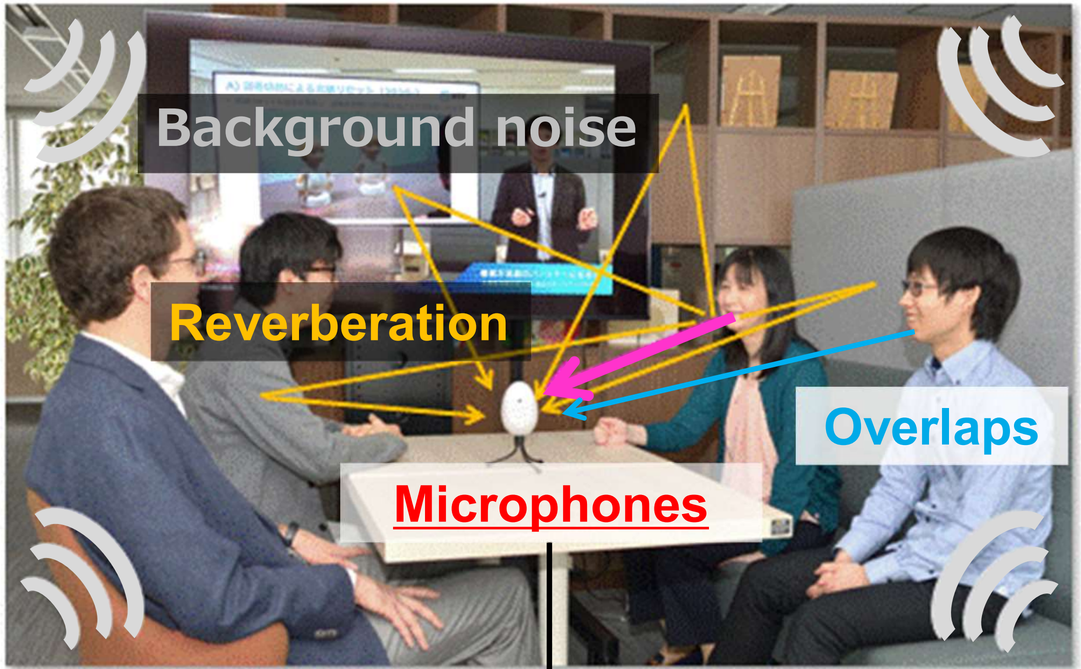
Develop key technologies
for understanding natural human speech conversations
to better support our everyday communication

Conversational speech processing

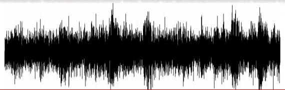


Speech recognition | Speaker recognition | Speaker attribute estimation | Speech summarization | Speech translation | …

Conversational speech processing



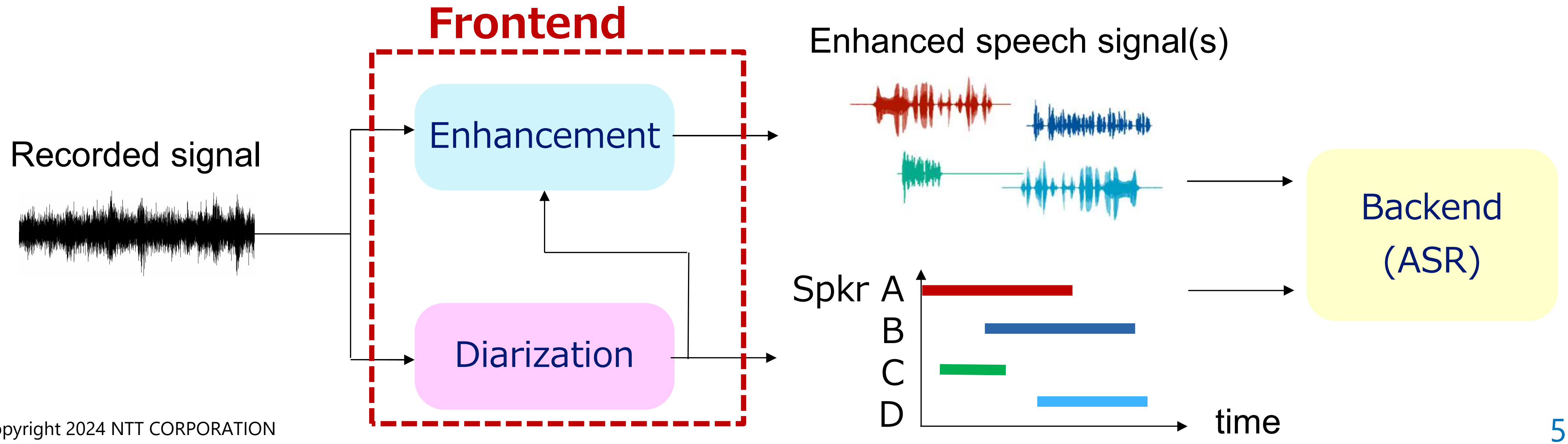
Recorded signal



Frontend

- Speech recognition
- Speaker recognition
- Speaker attribute estimation
- Speech summarization
- Speech translation
- …

Frontend for Conversational Speech Processing



Real-time Meeting Analysis System (demo video)

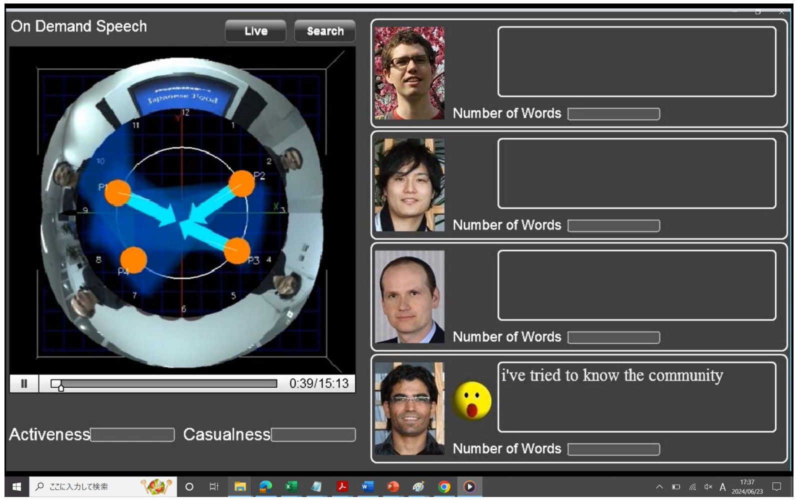
NTT

[Araki+2010 (NTT))][Araki+2011 (NTT)][Hori+2012(NTT)]

Audio-Visual Processing

- 8ch microphone array
- Omni-directional camera

Real-time Meeting Browser



Who is speaking When, What, and to Whom?

T. Hori, et al, "Low-latency realtime meeting recognition and understanding using distant microphones and omni-directional camera," IEEE TASLP, 2012.

Real-time Meeting Analysis System (demo video)

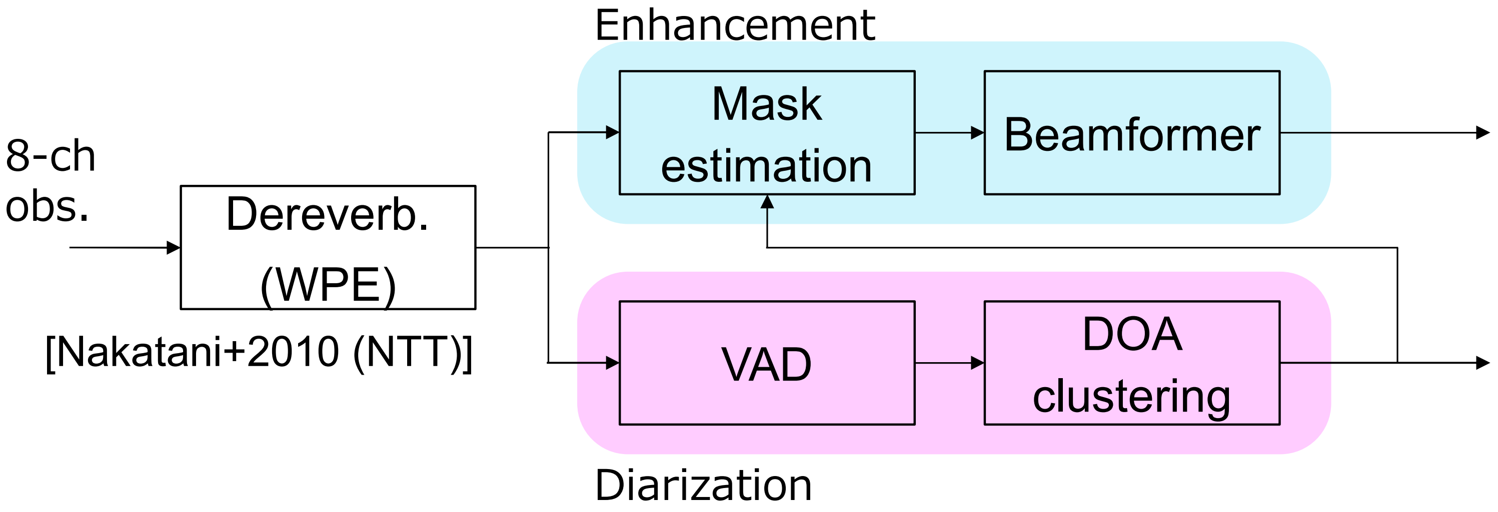
Real-time Audio-visual Meeting Recognition and Understanding Using Distant Microphone Array

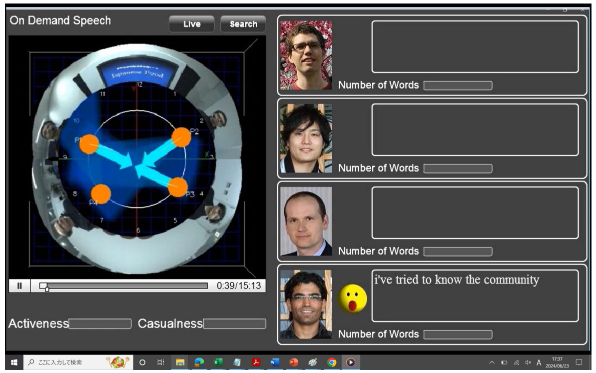
Presented at
NTT CS Labs. Open House 2011 and
ICASSP 2012 Show & Tell

NTT Corporation

Real-time Meeting Analysis System in 2010

NTT





[Araki+2010 (NTT))]
[Araki+2011 (NTT)]
[Hori+2012 (NTT)]

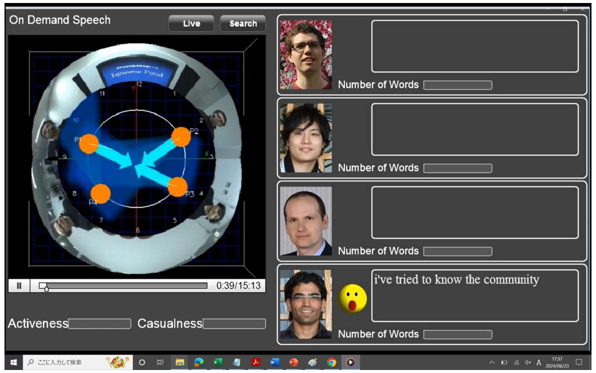
- Worked well
- In "Real-time", "low-latency" in 2010
  - No Neural Network / No training for frontend
  - No GPU for speech processing

T. Hori, et al, "Low-latency realtime meeting recognition and understanding using distant microphones and omni-directional camera," IEEE TASLP, 2012
S. Araki, et al.,, "Online meeting recognizer with multichannel speaker diarization", Asilomar 2010.
T. Nakatani et al., "Speech dereverberation based on variance-normalized delayed linear prediction," IEEE TASLP, 2010.

Towards frontend for various daily scenarios

NTT

PoC (2010)



Limited scenarios
(e.g., small meeting)

- High S/N, low reverb.
- 4 speakers
- Seated

Breakthrough Advance

- **Enhancement**
- **Diarization**
- **ASR**

Real world (2024)

Various daily scenarios
(e.g., CHiME-7/8 challenges)

- Low S/N, more reverb.
- Arbitrary number of speakers
- Dynamic, moving

Contents

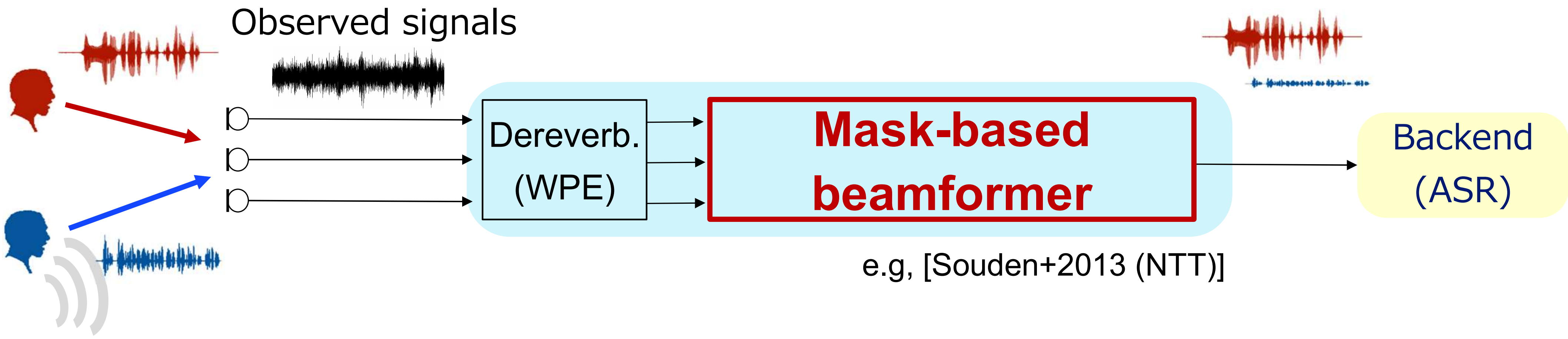
1. Frontend for conversational speech processing
   - Mask-based beamformer
2. Key technologies for handling various recording conditions
   - Blind mask estimation: Spatial feature clustering
   - Arbitrary number of speakers:
     - Speaker Diarization
     - Target speech extraction
   - Dynamic conditions: Beamformer for moving speakers
3. Remaining challenges & Closing remarks

Speech enhancement: Requirements

Reduce mismatch between observed speech and backend

- Reduce noise and interference
  while maintaining target speech (distortionless)
  → so that the frontend does not adversely affect the backend

Speech enhancement: Requirements



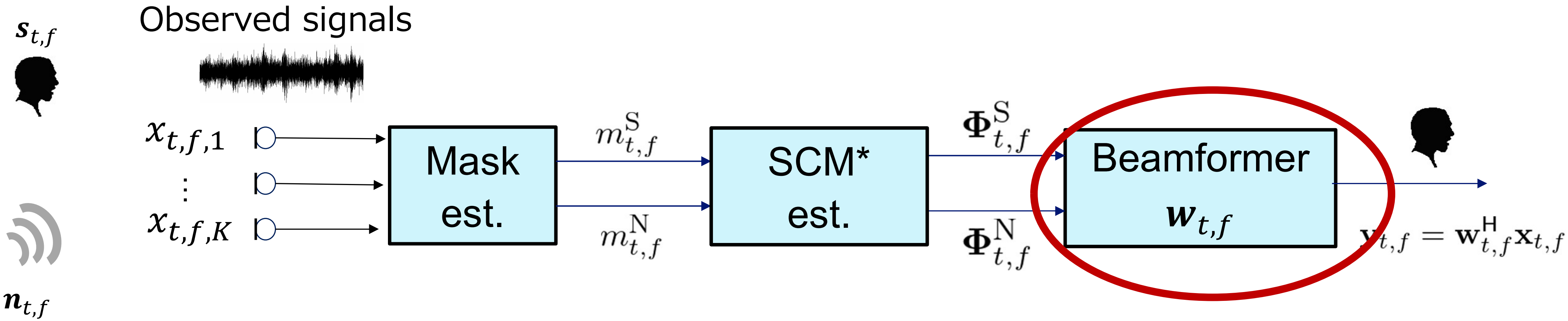
Reduce mismatch between observed speech and backend

- Reduce noise and interference
  while maintaining target speech (distortionless)
  → so that the frontend does not adversely affect the backend

M. Souden et al., "A Multichannel MMSE-Based Framework for Speech Source Separation and Noise Reduction," IEEE TASLP, 2013.

Mask-based beamformer

[Higuchi+ 2016 (NTT)], [Heymann+ 2016], …

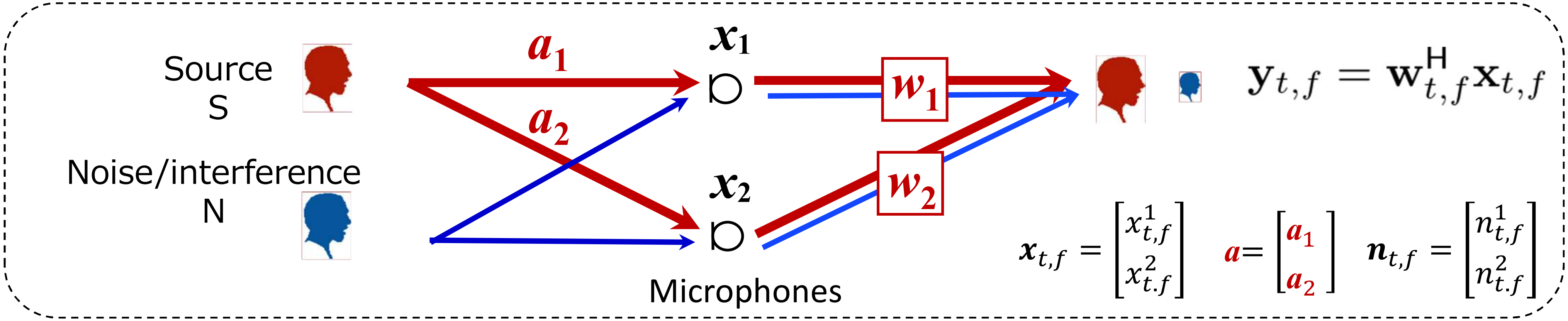


*SCM: spatial covariance matrix

T. Higuchi, et al., "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," ICASSP2016.
J. Heymann, et al., "Neural network based spectral mask estimation for acoustic beamforming,"ICASSP2016.

MVDR beamformer

MVDR: minimum variance distortionless response [Frost, 1972]



Minimize noise and interference while maintaining target speech

MVDR (Minimum Variance Distortionless Response) beamformer

$$\min_{\mathbf{w}_{t,f}} |\mathbf{w}_f^H \mathbf{n}_{t,f}|^2 \quad \text{subject to} \quad \mathbf{w}_{t,f}^H \mathbf{a}_{t,f} = 1 \quad \text{(Distortionless)}$$

Effective when accurate a is given, but it is unavailable in a real conversation ☹

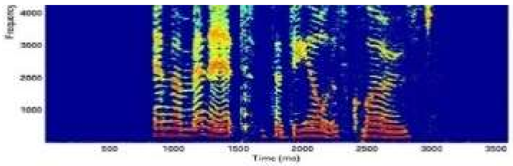
MVDR beamformer ← SCM ← Mask

- $\mathbf{a}_f$ can be estimated using $\boldsymbol{\Phi}^{\mathsf{S}}_{t,f}, \boldsymbol{\Phi}^{\mathsf{N}}_{t,f}$
- MMSE-based MVDR beamformer (avoid estimating $\mathbf{a}_f$)

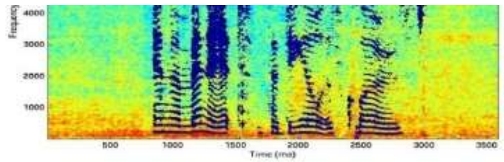
$$\mathbf{w}_{t,f} = \frac{(\boldsymbol{\Phi}^{\mathsf{N}}_{t,f})^{-1}\boldsymbol{\Phi}^{\mathsf{S}}_{t,f}}{\mathrm{Tr}((\boldsymbol{\Phi}^{\mathsf{N}}_{t,f})^{-1}\boldsymbol{\Phi}^{\mathsf{S}}_{t,f})}\mathbf{u},$$

[Souden+2010]

$$\boldsymbol{\Phi}^{\mathsf{S}}_{t,f} = \frac{1}{\sum_{t=1}^{T} m^{\mathsf{S}}_{t,f}} \sum_{t=1}^{T} \underline{m^{\mathsf{S}}_{t,f}} \mathbf{x}_{t,f}\mathbf{x}^{\mathsf{H}}_{t,f}$$



$$\boldsymbol{\Phi}^{\mathsf{N}}_{t,f} = \frac{1}{\sum_{t=1}^{T} m^{\mathsf{N}}_{t,f}} \sum_{t=1}^{T} \underline{m^{\mathsf{N}}_{t,f}} \mathbf{x}_{t,f}\mathbf{x}^{\mathsf{H}}_{t,f}$$



$\boldsymbol{\Phi}^{\mathsf{S}}_{t,f}, \boldsymbol{\Phi}^{\mathsf{N}}_{t,f}$: Spatial covariance matrices (SCMs) for source & noise

$m^{\mathsf{S}}_{t,f}, m^{\mathsf{N}}_{t,f}$: Time-frequency masks for source & noise

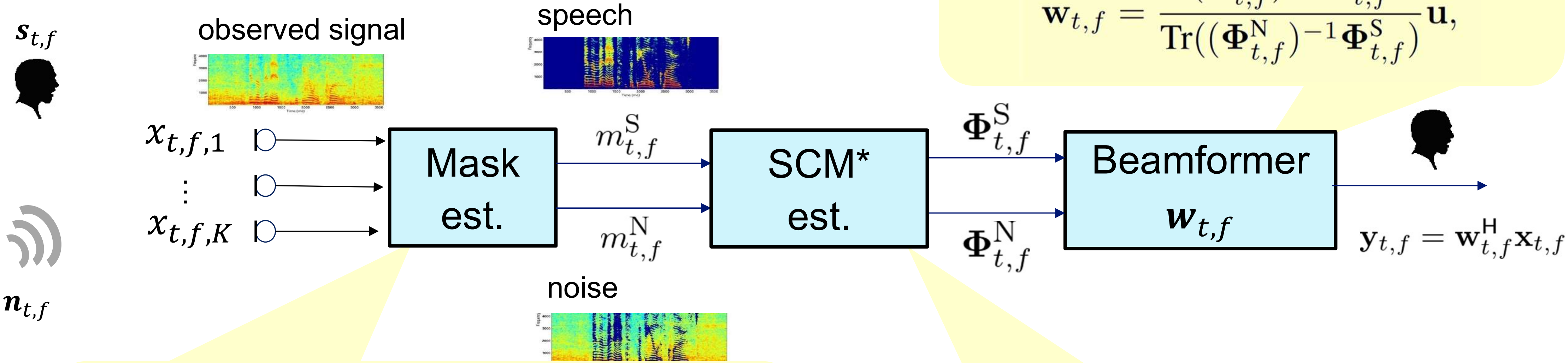
M. Souden, et al., "On Optimal Frequency-Domain Multichannel Linear Filtering for Noise Reduction," IEEE TASLP, 2010,

Mask-based beamformer

[Higuchi+ 2016 (NTT)], [Heymann+ 2016], …

MVDR** beamformer [Souden+2010]

$$\mathbf{w}_{t,f} = \frac{(\mathbf{\Phi}^{\mathrm{N}}_{t,f})^{-1}\mathbf{\Phi}^{\mathrm{S}}_{t,f}}{\mathrm{Tr}((\mathbf{\Phi}^{\mathrm{N}}_{t,f})^{-1}\mathbf{\Phi}^{\mathrm{S}}_{t,f})}\mathbf{u},$$



- Spatial feature clustering based
  e.g.,) cACGMM [Ito+2016(NTT)]
- Spectro-temporal info-based
  e.g.) Continuous source separation (CSS)
  Target speaker extraction
- Hybrid e.g.) [Nakatani+2017(NTT)][Drude+2019]

Estimate spatial covariance matrix (SCM) of target speech and noise

*SCM: spatial covariance matrix

** Minimum Variance Distortionless Response

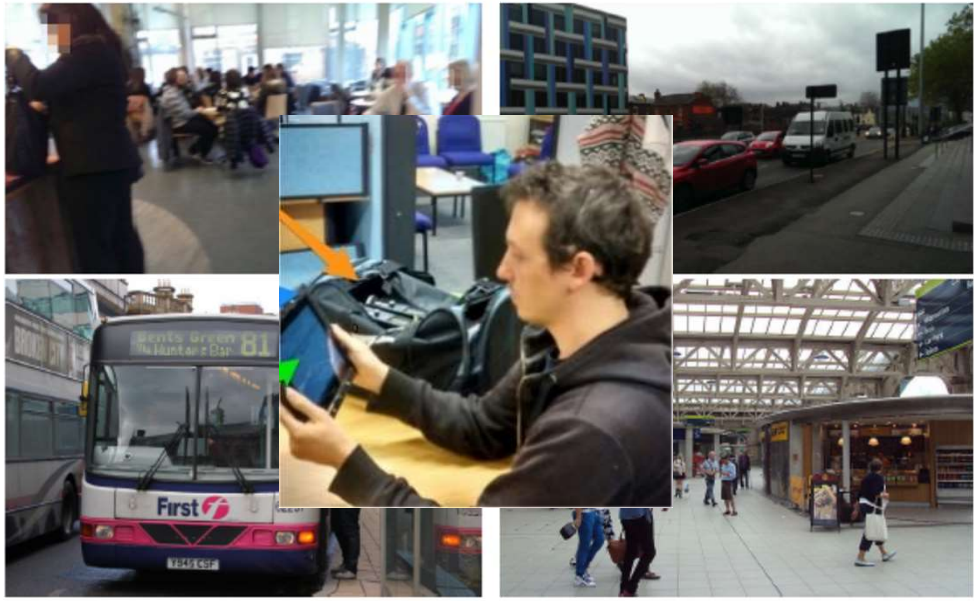
T. Higuchi, et al., "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," ICASSP2016.
J. Heymann, et al., "Neural network based spectral mask estimation for acoustic beamforming,“ICASSP2016.
M. Souden, et al., "On Optimal Frequency-Domain Multichannel Linear Filtering for Noise Reduction," IEEE TASLP, 2010,

Mask-based beamformer proved effective

[Yoshioka+2015 (NTT)]

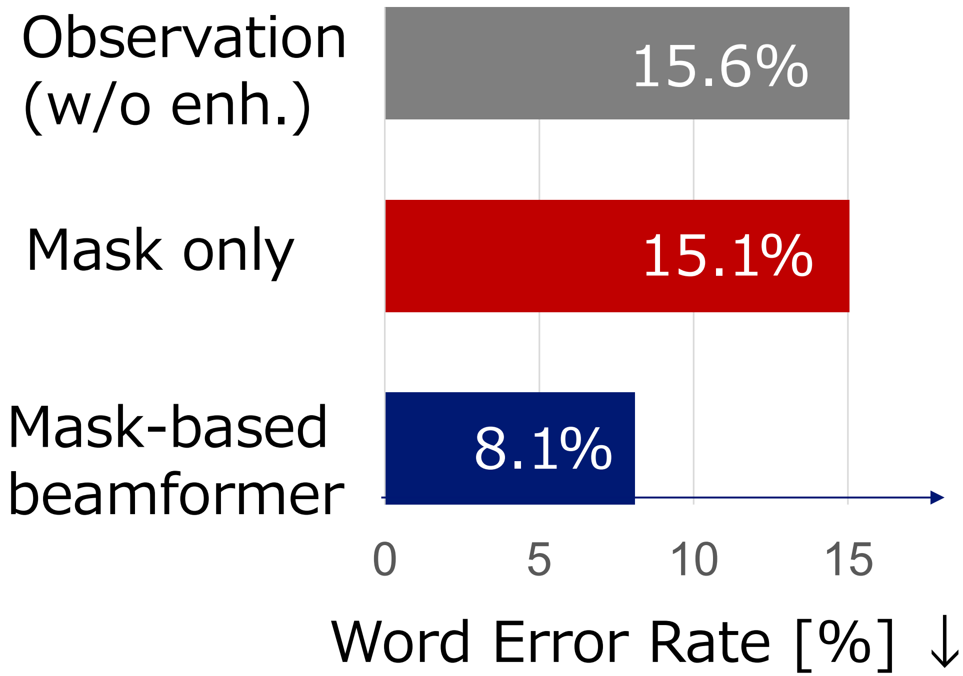
for DNN-based ASR backend

CHiME-3/4: ASR in public area



https://spandh.dcs.shef.ac.uk/chime_challenge/CHiME4/index.html

cGMM-based mask + MVDR beamformer

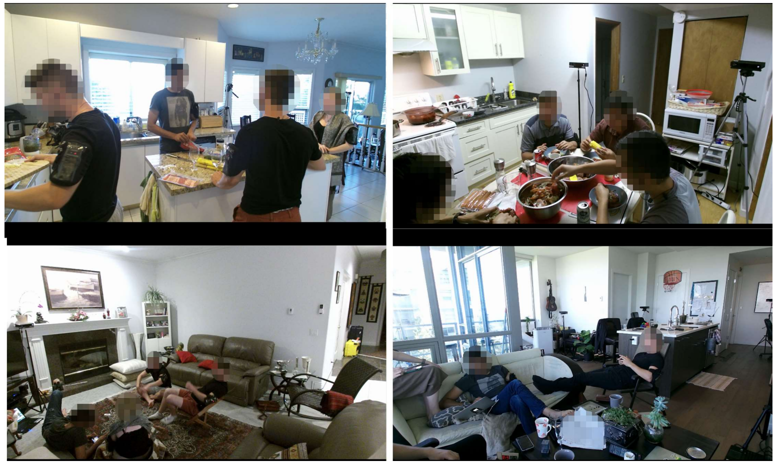


T. Yoshioka et al., "The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," ASRU2015.

Mask-based beamformer proved effective

for real-world conversation

◆ **CHiME-5/6: ASR/diarization**
in dinner party

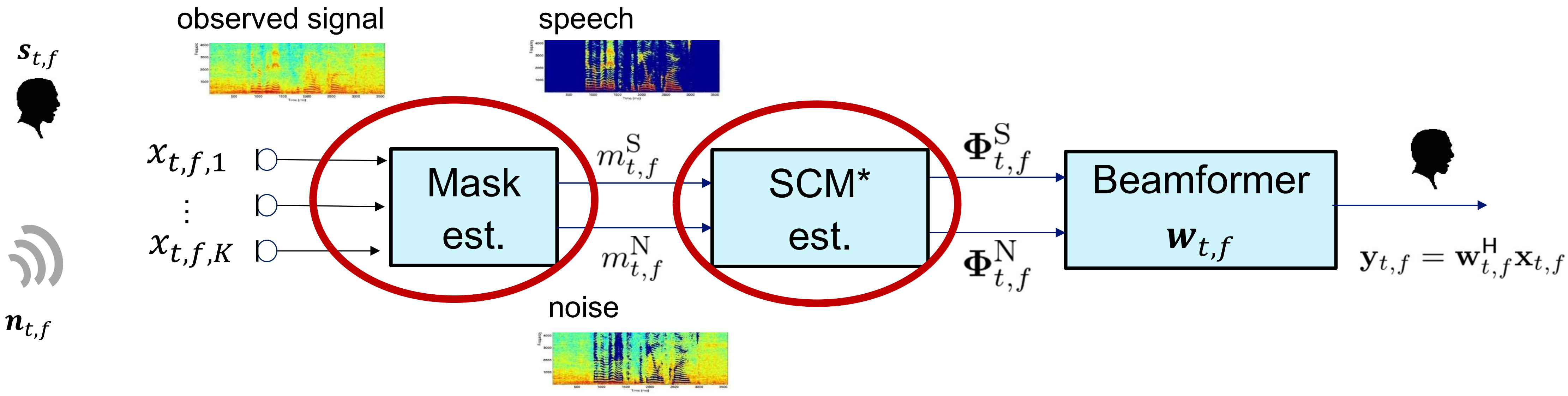


◆ **NTT's PoC: ASR**
in exhibition noise

◆ **CHiME-7: ASR/diarization**
in multiple scenarios

Three real datasets
- CHiME-6: Dinner party (4 participants)
- DiPCO: Dinner party (4 participants)
- Mixer: Interview (2 speakers)

Mask-based Beamformer in real conversations



- Blind / unsupervised approach for unseen conditions
  → Spatial feature clustering
  Arbitrary number of speakers
  → Target Speaker Extraction
- Dynamic conditions:
  → Time-varying SCM estimation

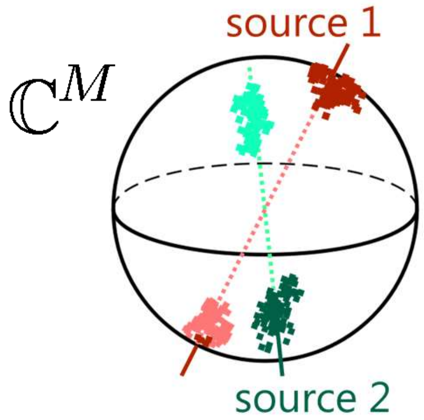
*SCM: spatial covariance matrix

Contents

1. Frontend for conversational speech processing
    - Mask-based Beamformer
2. Key technologies for handling various recording conditions
    - Blind mask estimation: Spatial feature clustering
    - Arbitrary number of speakers:
        - Speaker Diarization
        - Target speech extraction
    - Dynamic conditions: Beamformer for moving speakers
3. Remaining challenges & Closing remarks

Spatial feature clustering-based mask estimation

- Blind/unsupervised method
- Spatial features (with arbitrary num. of mics.):
  Normalized observation vector [Sawada+2010 (NTT)]

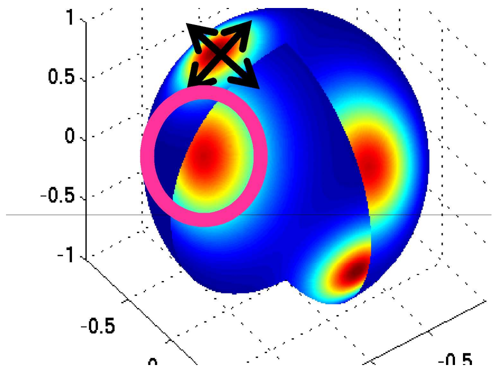


$$\boldsymbol{z}_{tf} = \frac{\boldsymbol{x}_{tf}}{\left\|\boldsymbol{x}_{tf}\right\|_2} \qquad \text{where } \boldsymbol{x}_{tf} = \left[x_{tf}^{(1)} \ \ldots \ x_{tf}^{(M)}\right]^{\mathrm{T}} \in \mathbb{C}^M$$

: Observation vector

  - Unit norm → Unit hyper sphere $\mathbb{C}^M$
  - Each cluster = Each source

- **Complex Watson Mixture Model (cWMM)**
  - [D. H. Tran Vu & Haeb-Umbach 2010]



Complex Watson distribution: Isotropic distribution

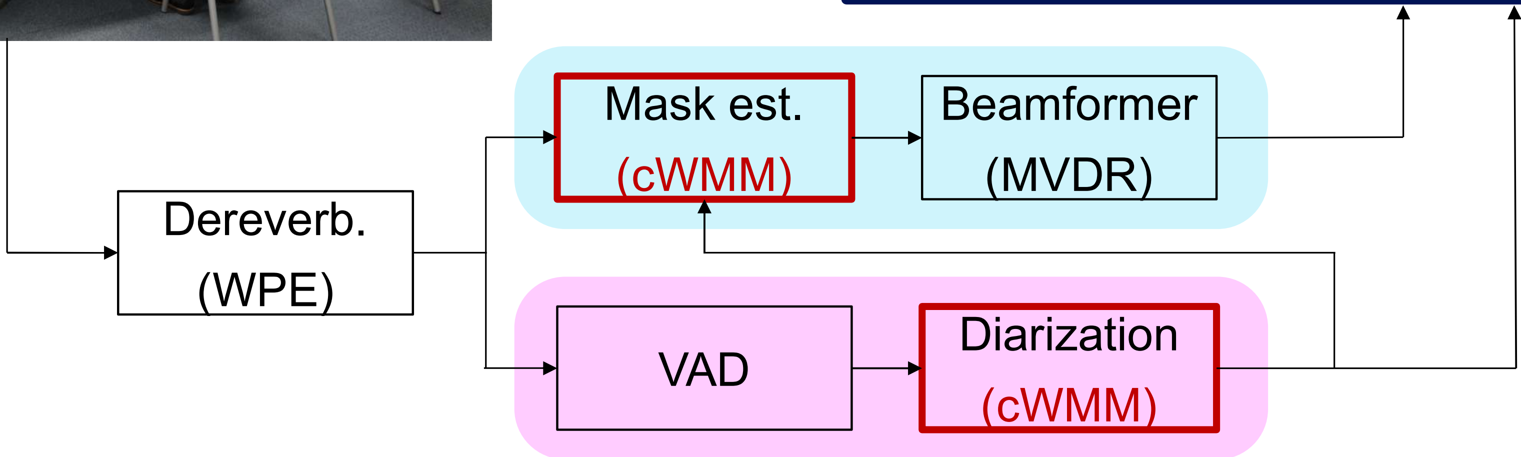
$$\mathcal{W}(\mathbf{z}; \mathbf{a}, \kappa) \propto e^{\kappa|\mathbf{a}^{\mathrm{H}}\mathbf{z}|}$$

H. Sawada et al,, "Underdetermined Convolutive Blind Source Separation via Frequency Bin-Wise Clustering and Permutation Alignment," *IEEE TASLP 2010.*
D. H. Tran Vu and R. Haeb-Umbach, "Blind speech separation employing directional statistics in an Expectation Maximization framework," ICASSP2010.

cWMM-based mask + MVDR beamformer
Demo Video: Online meeting recognizer

NTT

[Araki+2017(NTT)]



Outside: exhibition noise

S. Araki, et al., "Online Meeting Recognition in Noisy Environments with Time-Frequency Mask Based MVDR Beamforming," HSCMA2017.
N. Ito+, "Data-driven and physical model-based designs of probabilistic spatial dictionary for online meeting diarization and adaptive beamforming," EUSIPCO2017.

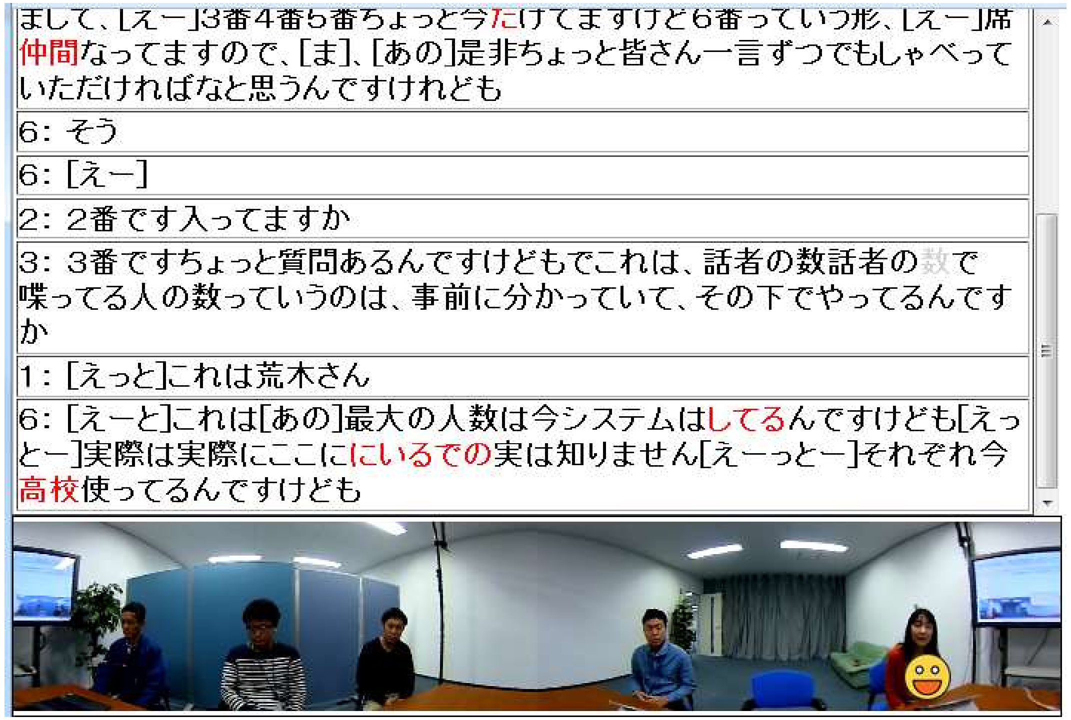
Online meeting recognition in noisy environments with mask-based beamforming

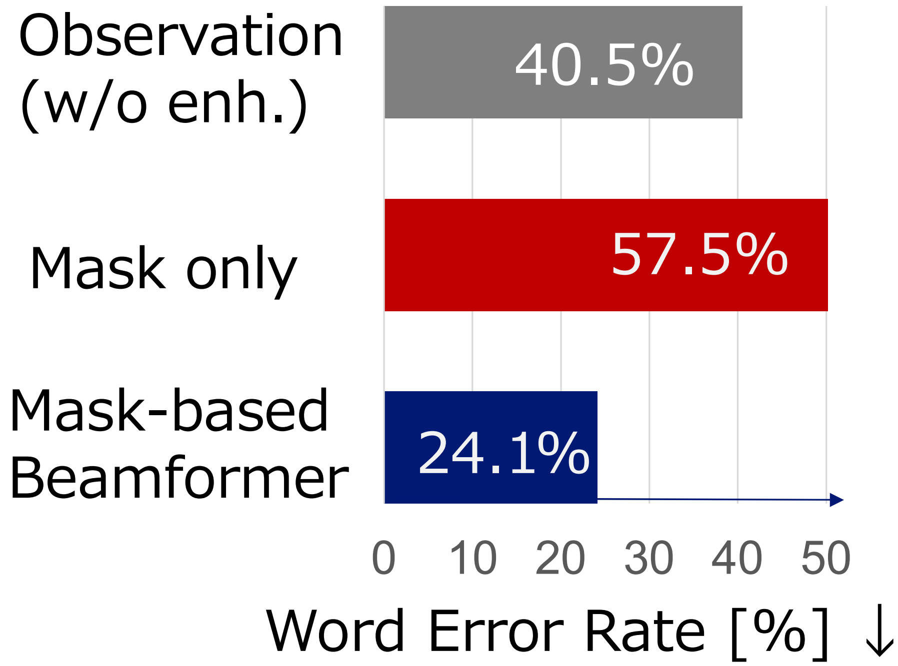
Presented at
NTT CS Labs. Open House 2016 & IEEE HSCMA2017

NTT Corporation

Demo video: Online prototype [Araki+2017 (NTT)] NTT

Worked in noisy and reverberant scenarios (e.g., research exhibition)

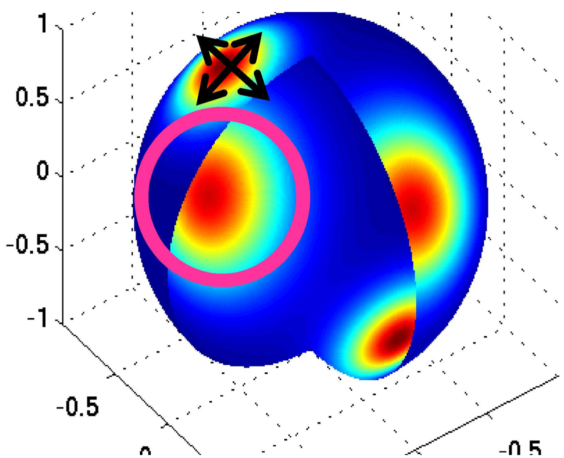




S. Araki, et al., "Online Meeting Recognition in Noisy Environments with Time-Frequency Mask Based MVDR Beamforming," HSCMA2017.

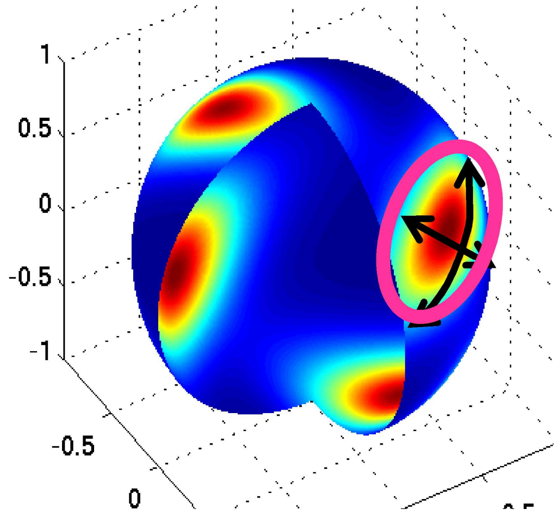
Directional statistics-based mask estimation

Complex Watson Mixture Model (cWMM)
[D. H. Tran Vu & Haeb-Umbach 2010]



Isotropic distribution
→Not always…
→Less accurate

Complex Angular Central Gaussian Mixture Model (cACGMM)
[Ito+2016 (NTT)]



Direction, shape

Elliptical distribution → More accurate

D. H. Tran Vu and R. Haeb-Umbach, "Blind speech separation employing directional statistics in an Expectation Maximization framework," ICASSP2010.
N. Ito, et al., "Complex angular central Gaussian mixture model for directional statistics in mask-based microphone array signal processing," *EUSIPCO2016*

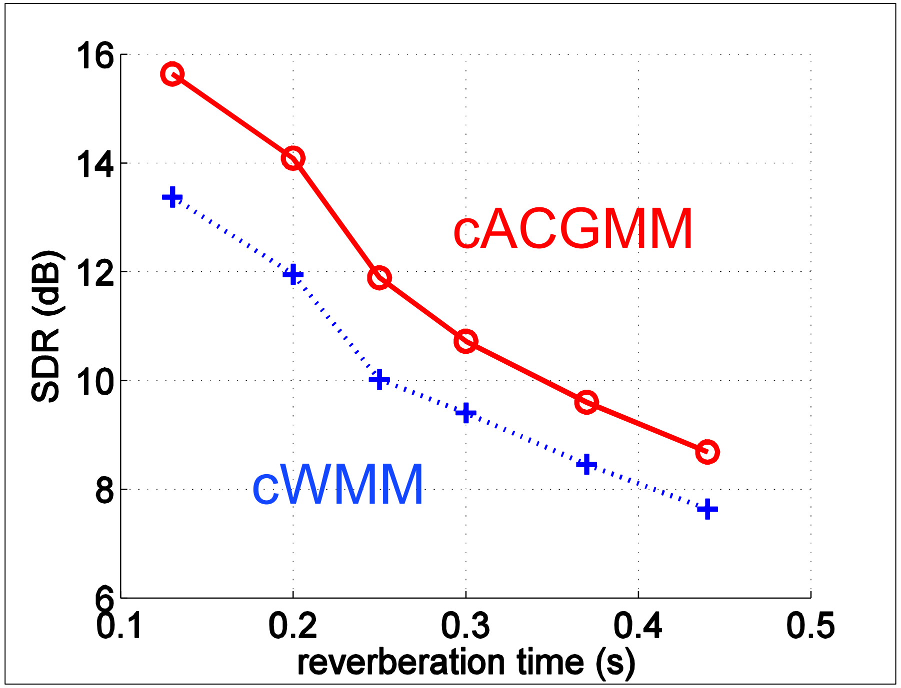
cWMM vs cACGMM [Ito+2016 (NTT)]

NTT

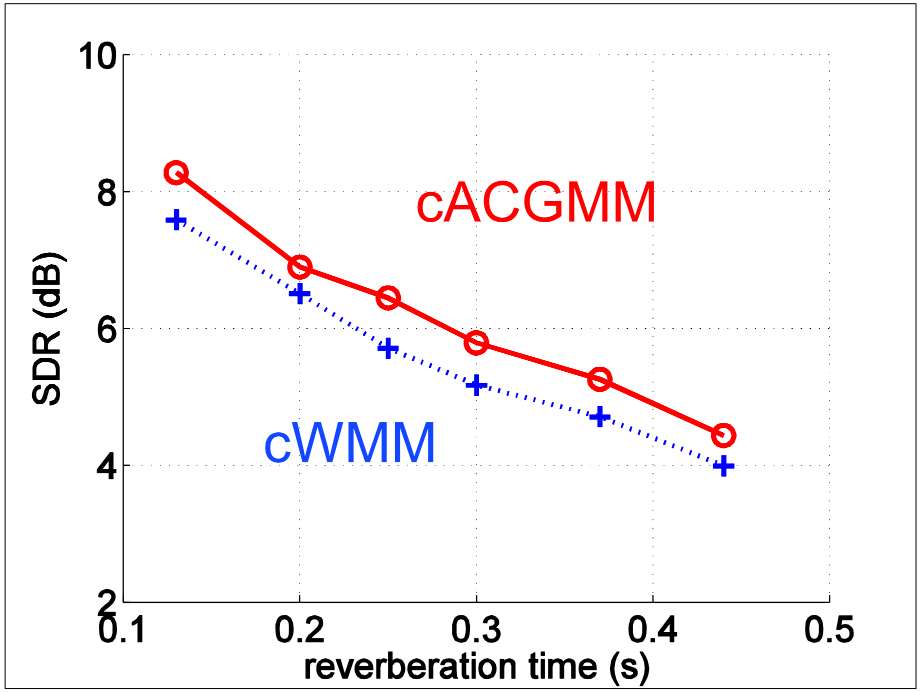
Good

Poor

2 speech separation with 2 microphones



3 speech separation with 2 microphones



- cACGMM outperforms cWMM
- cACGMM is employed by many SOTA systems

N. Ito, et al., "Complex angular central Gaussian mixture model for directional statistics in mask-based microphone array signal processing," EUSIPCO2016

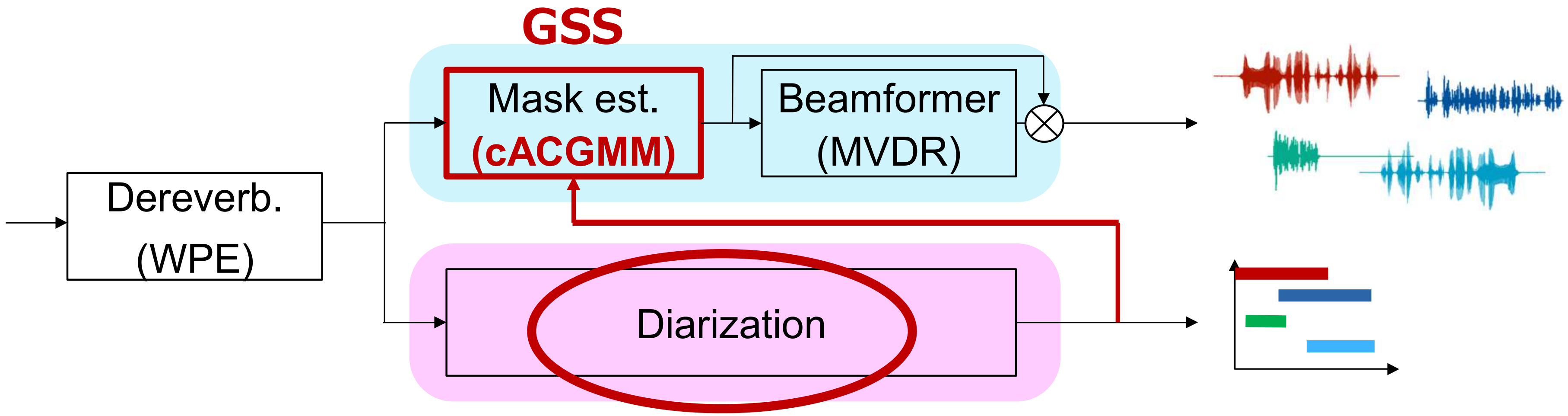
cACGMM-based mask estimation

GSS: Guided source separation

[Boeddecker+2018]

cACGMM-based mask estimation guided by **time annotation** with diarization

- Helps avoid frequency permutation problem in clustering
- Provides number of speakers (clusters)



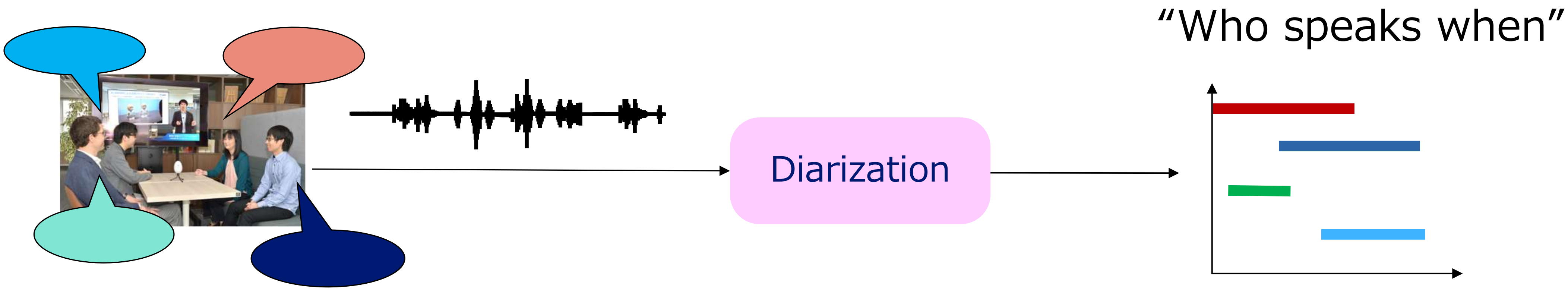
→ Employed by most of current SOTA systems (e.g., All systems in CHiME-7 (2023))

C. Boeddecker, et al., "Front-end processing for the CHiME-5 dinner party scenario," CHiME-2018

Contents

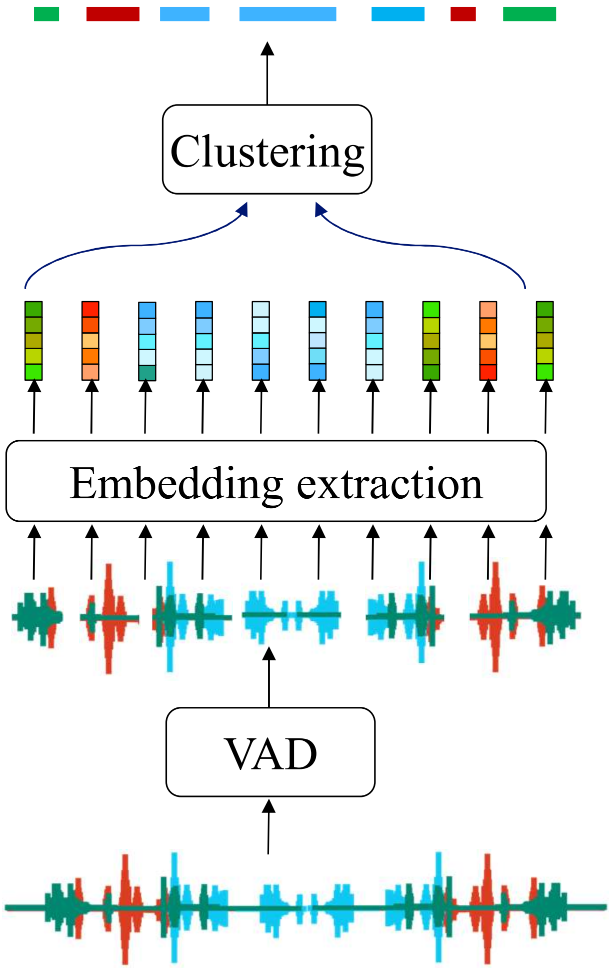
1. Frontend for conversational speech processing
    - Mask-based Beamformer
2. Key technologies for handling various recording conditions
    - Blind processing: Spatial feature clustering
    - Arbitrary number of speakers:
        - Speaker Diarization
        - Target speech extraction
    - Dynamic conditions: Beamformer for moving speakers
3. Remaining challenges & Closing remarks

Diarization



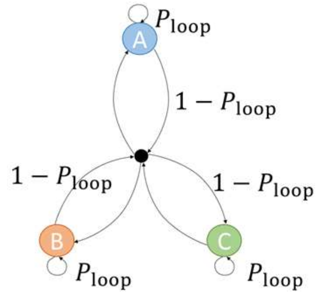
- Fundamental technology **essential for conversational speech processing**
  - **E.g., speaker-attributed ASR**
  - **Useful for speech enhancement (e.g., GSS)**
- Difficulties:
  - Some utterances are **overlap with other speaker's voice**
  - **The number of speakers are unknown**

Embedding vector clustering



1-stream output

- AHC (Agglomerative Hierarchical Clustering)
- VBx (Variational Bayesian clustering of x-vectors) [Landini+2022]



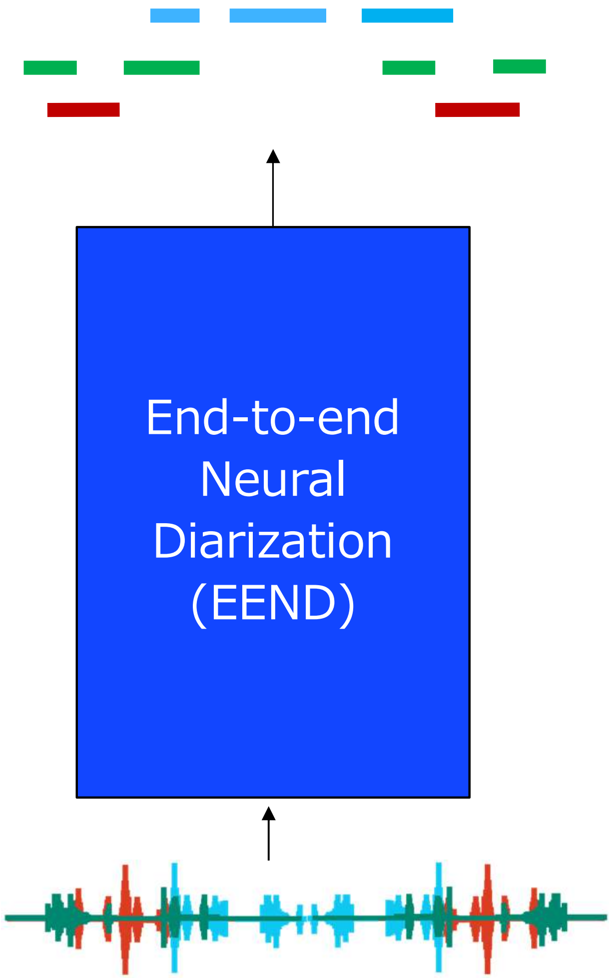
- i-vector
- x-vector (TDNN, ECAPA-TDNN, Resnet…)

(assuming 1 speaker @each segment)

	VC
Overlap	☹
Arbitrary num. speaker	☺

F. Landini, et al., "Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: Theory, implementation and analysis on standard tasks," Computer Speech & Language, 2022.

End-to-end neural diarization (EEND)



N-stream output
(N: given)

[Fujita+, 2019]

Multi-label classification
Ex.) WavLM+Transformer

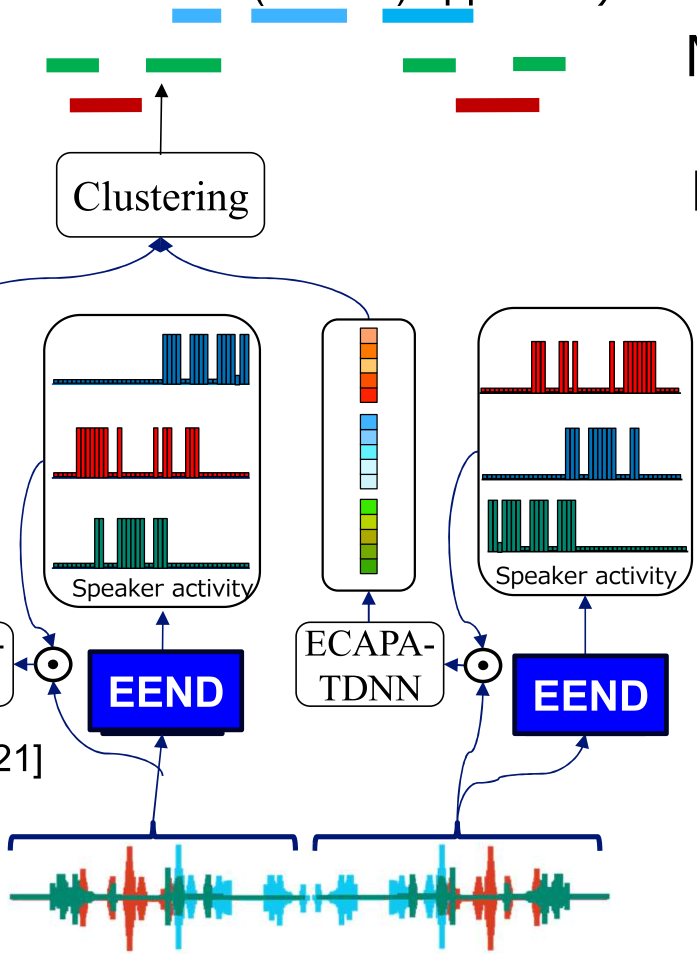
	VC	EEND
Overlap	☹	☺
Arbitrary num. speaker	☺	☹

Complementary
→Integrate them to get the most out of both

Y. Fujita, et al., "End-to-end neural speaker diarization with self attention," ASRU2019.

EEND-VC* *End-to-end neural diarization and vector clustering [Kinoshita+2021 (NTT)]

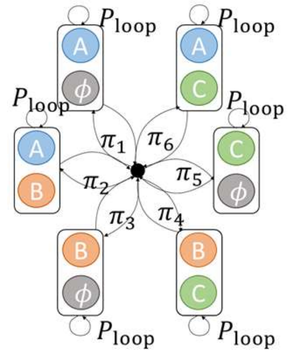
(Best of both worlds (BOBW) approach)



[Bredin+2021]

Multi-stream output

Multi-stream VBx [Delcroix+2023 (NTT)]



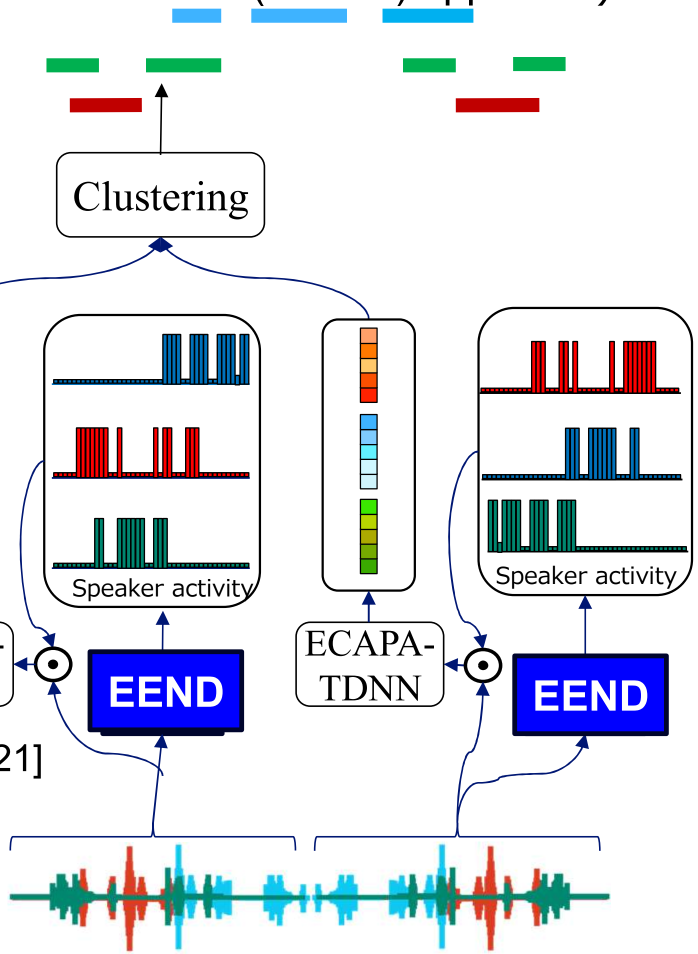
Estimate diarization results and speaker embeddings.

DER (%)	CALLHOME	DIHARD-III
VC	13.6	20.5
EEND	11.8	19.5
EEND-VC	11.1	19.3
EEND-VC +MS-VBx	10.4	18.2

K. Kinoshita, et al., “Integrating end-to-end neural and clustering-based diarization: Getting the best of both worlds,”ICASSP2021.
M. Delcroix, et al., "Multi-Stream Extension of Variational Bayesian HMM Clustering (MS-VBx) for Combined End-to-End and Vector Clustering-based Diarization," Interspeech2023.

EEND-VC* *End-to-end neural diarization and vector clustering [Kinoshita+2021 (NTT)]

(Best of both worlds (BOBW) approach)



[Bredin+2021]

	EEND-VC +MS-VBx
Overlap	☺
Arbitrary num. speaker	☺

- Adopted in pyannote
- Worked quite well even for multiple recording conditions (e.g., CHIME-7/8) [Tawara+2024 (NTT)] [Kamo+2024 (NTT)]

N. Tawara et al., “NTT speaker diarization system for CHiME-7: multi-domain, multi-microphone End-to-end and vector clustering diarization,” ICASSP2024
N. Kamo, et al, ,” NTT Multi-Speaker ASR System for the DASR Task of CHiME-8 Challenge, “ CHiME2024 workshop.

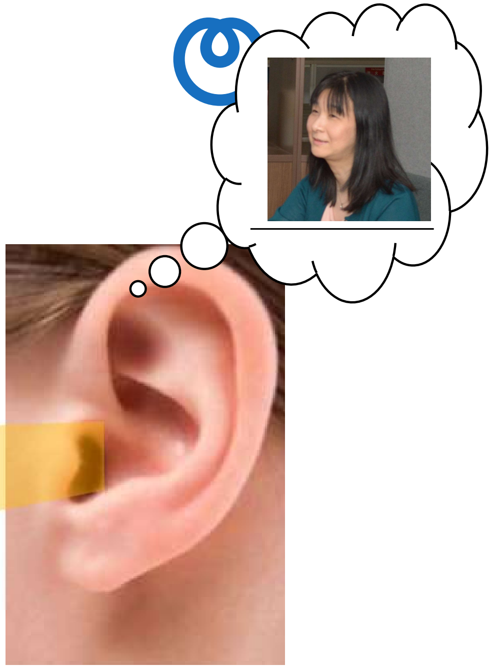
Contents

1. Frontend for conversational speech processing
   - Mask-based Beamformer
2. Key technologies for handling various recording conditions
   - Blind mask estimation: Spatial feature clustering
   - Arbitrary number of speakers:
     - Speaker Diarization
     - Target speech extraction
   - Dynamic conditions: Beamformer for moving speakers
3. Remaining challenges & Closing remarks

Turn-taking: speakers change

Number of simultaneous speakers is changing (& unknown).
→ Require speech enhancement that does not depend on num. targets

Separate all → Target speech extraction



Listening only to the “Target” voice, not everyone

TSE enables speech enhancement regardless of the number of speakers

SpeakerBeam:
Deep learning based target speech extraction

First successful attempt to extract the voice of a target speaker based on the characteristics of his/her voice

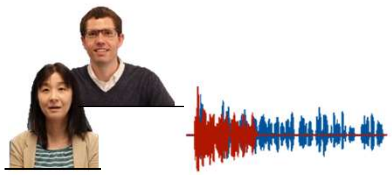
Demo@Youtube

Use recording of the voice of the target speaker (10 sec) as auxiliary information

Speaker characteristic Neural net

Compute the characteristics of the voice of the target speaker

Speech mixture



Target speech extraction Neural net

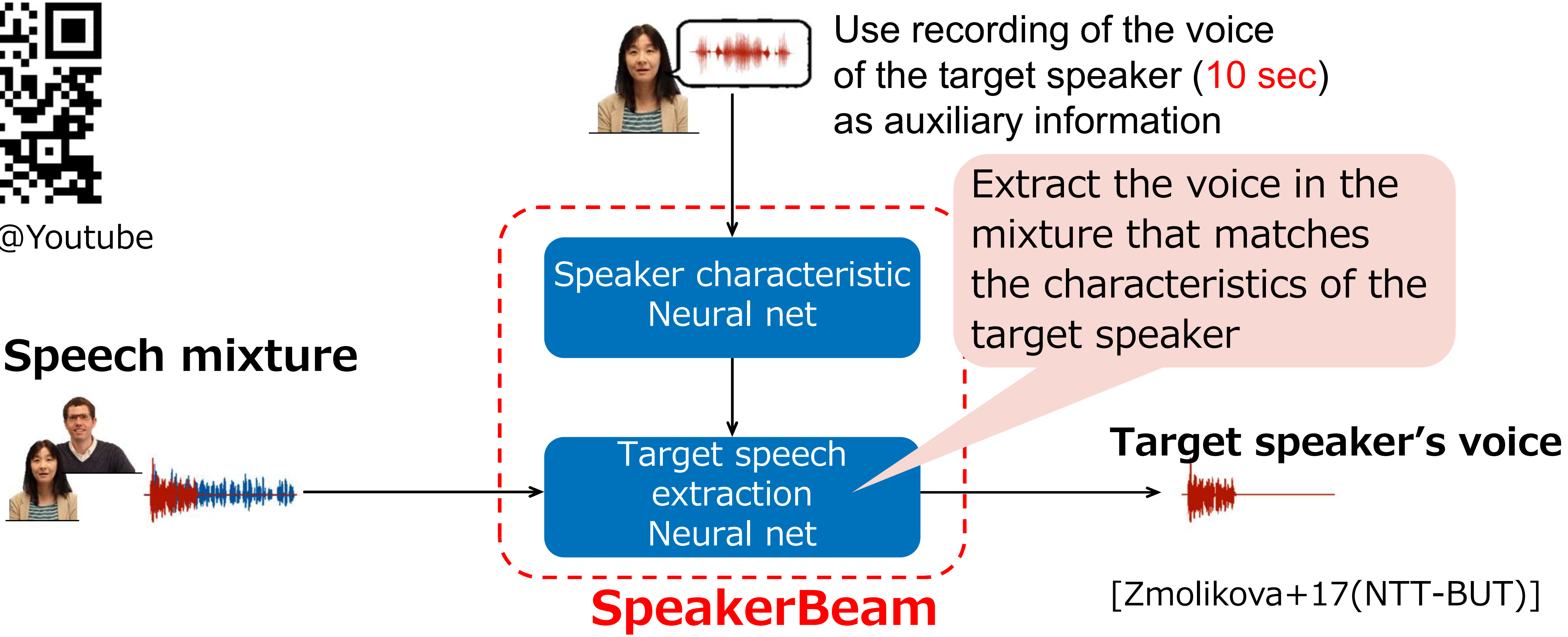
[Zmolikova+17(NTT-BUT)]

K. Zmolikova, et al., "Speaker-aware neural network based beamformer for speaker extraction in speech mixtures," Interspeech2023.

SpeakerBeam:
Deep learning based target speech extraction

First successful attempt to extract the voice of a target speaker based on the characteristics of his/her voice

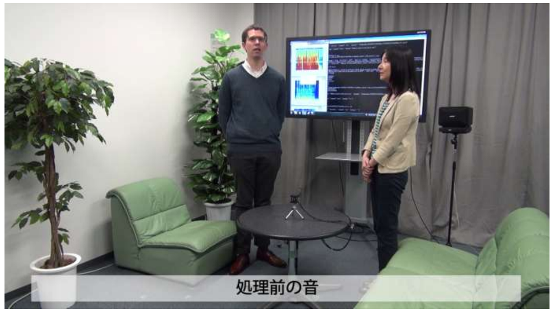
Demo@Youtube



[Zmolikova+17(NTT-BUT)]

K. Zmolikova, et al., “Speaker-aware neural network based beamformer for speaker extraction in speech mixtures,” Interspeech2023.

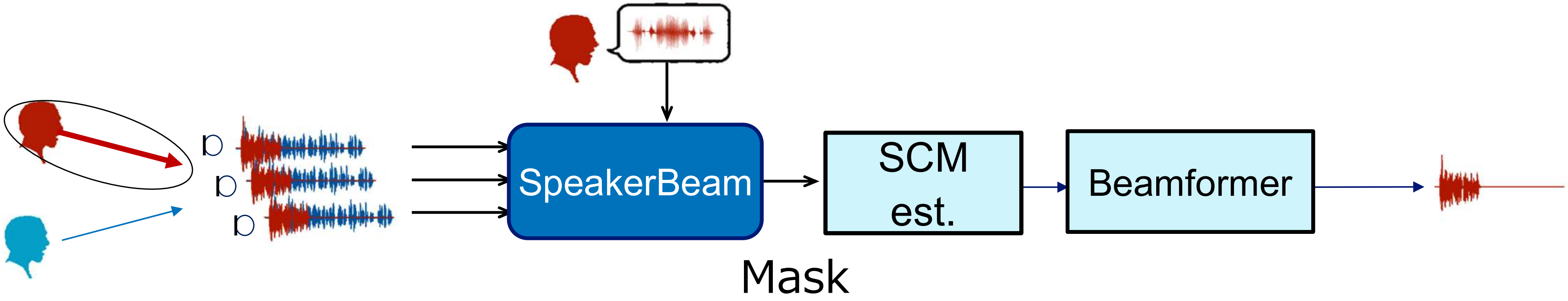
SpeakerBeam

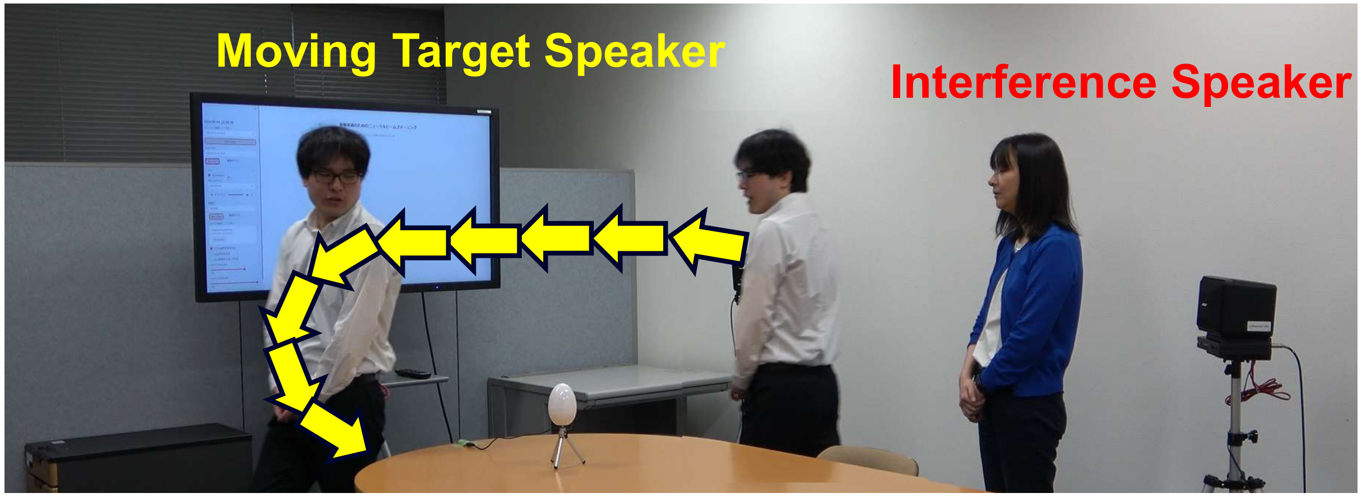


- Demo video in the next section & Youtube→→
- TSE concept has been employed for conversational speech processing
  - Speech enhancement independent of number of speakers [Ye+2023]
  - SOTA diarization approach (Target speaker VAD (TS-VAD)) [Medennikov+2020]
- Online and real-time implementation is also available
  - Related paper on Thursday (in Session A8-P5) [Sato+2024]

H. Sato et al., "SpeakerBeam-SS: Real-time target speaker extraction with lightweight Conv-TasNet and state space modeling," Interspeech 2024. (Thursday, Session A8-P5)

Mask-based beamformer for moving speakers





Microphone Array

Demonstration of
mask-based neural beamforming for moving speakers with self-attention-based tracking

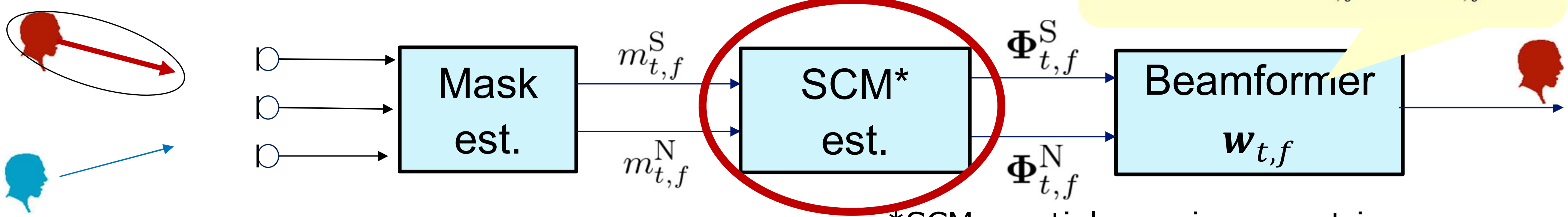
NTT Corporation

Mask-based beamformer for moving speakers

NTT

Moving → Time varying SCM

$$\mathbf{w}_{t,f} = \frac{(\boldsymbol{\Phi}^{\mathrm{N}}_{t,f})^{-1}\boldsymbol{\Phi}^{\mathrm{S}}_{t,f}}{\mathrm{Tr}((\boldsymbol{\Phi}^{\mathrm{N}}_{t,f})^{-1}\boldsymbol{\Phi}^{\mathrm{S}}_{t,f})}\mathbf{u}$$

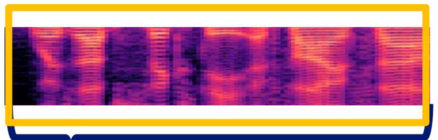


*SCM: spatial covariance matrix

Conventional mask-based SCM estimation

$\nu \in \{\mathrm{S}, \mathrm{N}\}$

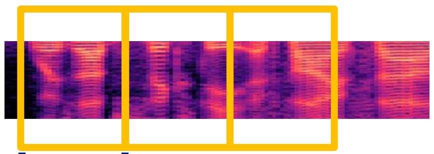
Time-invariant



Instantaneous SCM (ISCM)

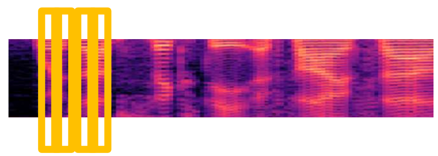
$$\boldsymbol{\Phi}^{\nu}_{f} = \sum_{\tau=1}^{T} \frac{1}{\sum_{\tau'=1}^{T} m^{\nu}_{\tau',f}} \underbrace{m^{\nu}_{\tau,f}\mathbf{x}_{\tau,f}\mathbf{x}^{\mathsf{H}}_{\tau,f}}_{\triangleq \boldsymbol{\Psi}^{\nu}_{\tau,f}}$$

Blockwise



$$\boldsymbol{\Phi}^{\nu}_{t,f} = \sum_{\tau=t-L}^{t+L} \frac{1}{\sum_{\tau'=t-L}^{t+L} m^{\nu}_{\tau',f}} \boldsymbol{\Psi}^{\nu}_{\tau,f}$$

Online



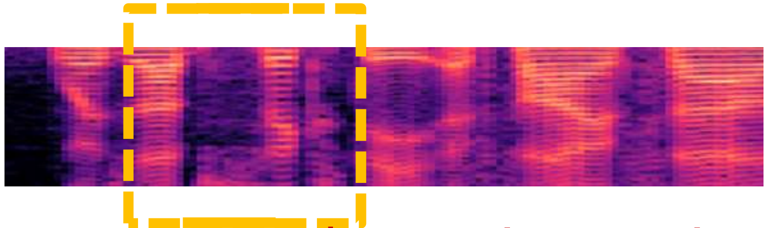
$$\boldsymbol{\Phi}^{\nu}_{t,f} = \alpha\boldsymbol{\Phi}^{\nu}_{t-1,f} + \boldsymbol{\Psi}^{\nu}_{t,f} = \sum_{\tau=1}^{t} \alpha^{t-\tau}\boldsymbol{\Psi}^{\nu}_{\tau,f}$$

Attention weight for SCM computation

Conventional: ☹ Preset fixed range → non-optimal for moving sources

$$\mathbf{\Phi}^{\nu}_{t,f} = \sum_{t'=1}^{T} c^{\nu}_{t,t'} \underbrace{\mathbf{\Psi}^{\nu}_{t',f}}_{\text{ISCM}}$$

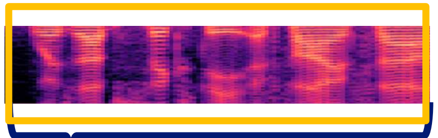
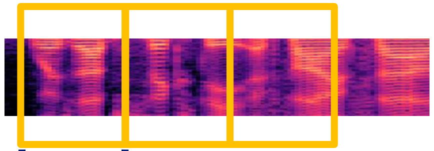
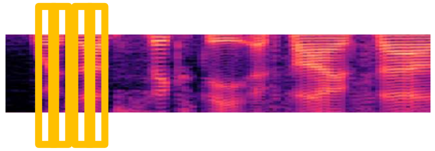
Attention weight



How can we determine optimal range for moving sources?

Conventional mask-based SCM estimation

$\nu \in \{\mathrm{S}, \mathrm{N}\}$

Time-invariant	Blockwise	Online
 Instantaneous SCM (ISCM) $\mathbf{\Phi}^{\nu}_{f} = \sum_{\tau=1}^{T} \frac{1}{\sum_{\tau'=1}^{T} m^{\nu}_{\tau',f}} \underbrace{m^{\nu}_{\tau,f}\mathbf{x}_{\tau,f}\mathbf{x}^{\mathsf{H}}_{\tau,f}}_{\triangleq \mathbf{\Psi}^{\nu}_{\tau,f}}$	 $\mathbf{\Phi}^{\nu}_{t,f} = \sum_{\tau=t-L}^{t+L} \frac{1}{\sum_{\tau'=t-L}^{t+L} m^{\nu}_{\tau',f}} \mathbf{\Psi}^{\nu}_{\tau,f}$	 $\mathbf{\Phi}^{\nu}_{t,f} = \alpha \mathbf{\Phi}^{\nu}_{t-1,f} + \mathbf{\Psi}^{\nu}_{t,f} = \sum_{\tau=1}^{t} \alpha^{t-\tau} \mathbf{\Psi}^{\nu}_{\tau,f}$

Attention-based SCM aggregate

[Ochiai+2023]

NTT

$$\mathbf{\Phi}^{\nu}_{t,f} = \sum_{t'=1}^{T} c^{\nu}_{t,t'} \underline{\mathbf{\Psi}^{\nu}_{t',f}}$$

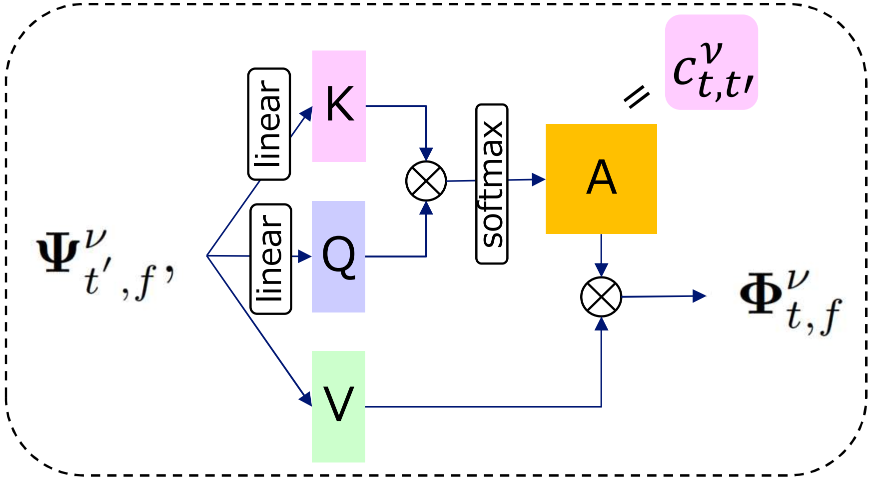
Instantaneous spatial covariance

Attention weight

This equation is similar to self-attention NN

Adopting self-attention-based NN

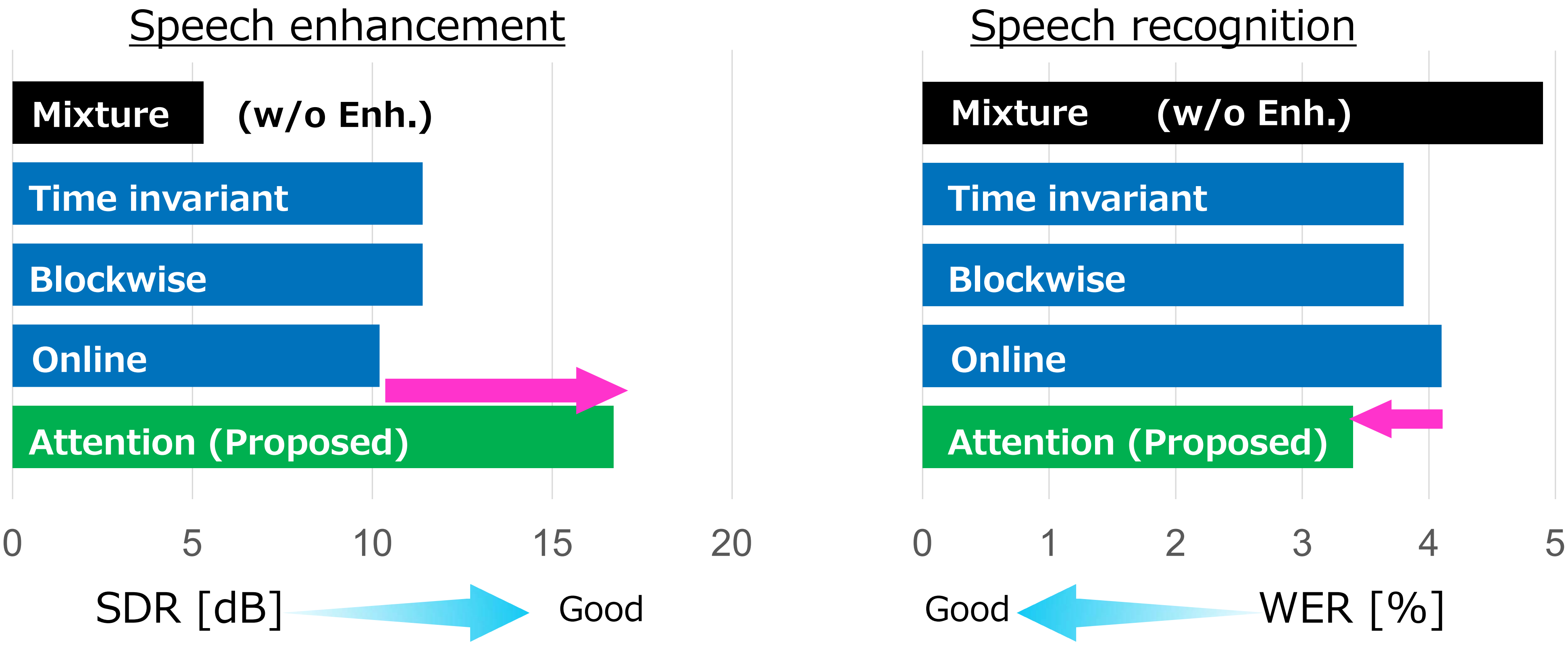
- Related paper on Wednesday [Tammen+2024] (Session A6-O4)



T. Ochiai, et al., "Mask-Based Neural Beamforming for Moving Speakers With Self-Attention-Based Tracking," IEEE TASLP 2023.
M. Tammen, et al., "Array Geometry-Robust Attention-Based Neural Beamformer for Moving Speakers," Interspeech 2024. (Wednesday, Session A6-O4)

Evaluation result

- 1 moving source (in a straight line) + noise (SNR = 2～8 dB)
- 5 microphones



Proposed attention-based Neural BF can handle moving sources.

Contents

Summary & Challenges

Key technologies of frontend for conversation speech processing

- **Mask-based beamformer** is widely adopted
- For handling **various recording conditions**
  - **Blind** mask estimation: Spatial feature clustering
  - **Arbitrary number of speakers**: Speaker Diarization, Target speech extraction
  - **Dynamic** conditions: Beamformer for moving speakers

Remaining challenges

- Light weight, low latency, online
- Artifact-free 1-ch speech enhancement (2 more slides!)
- Simulate/Measure RIRs of moving speakers for training data augmentation

1-ch speech enhancement: Artifact matters

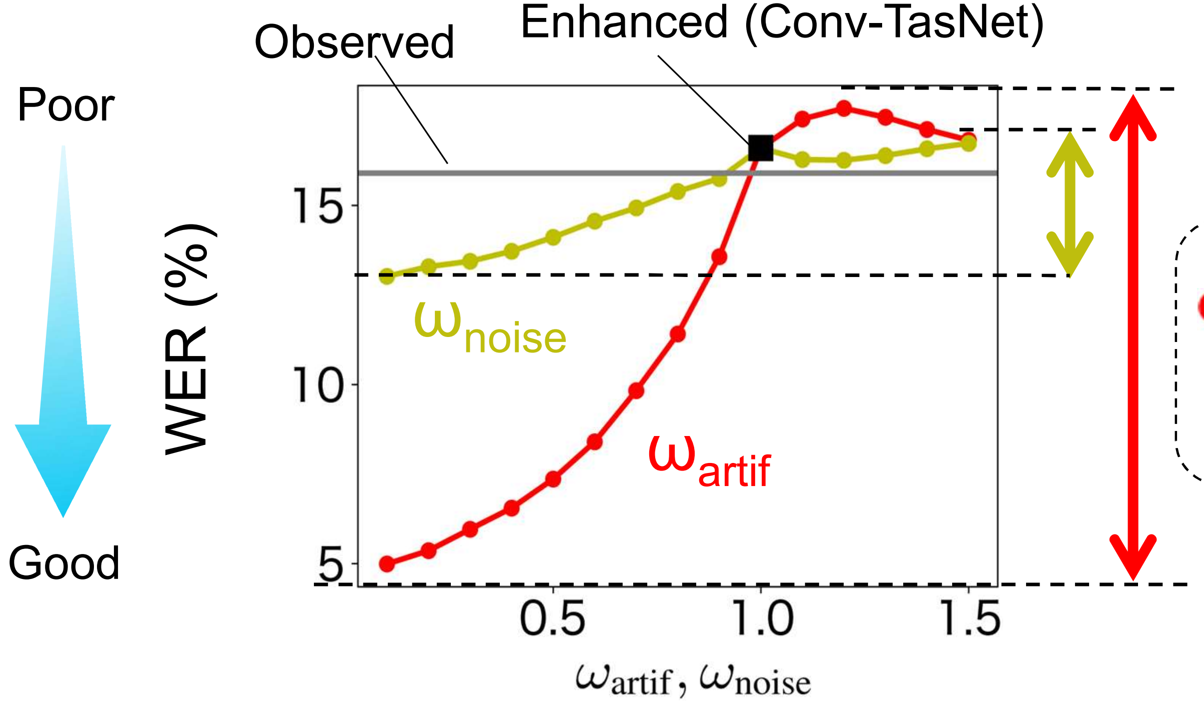
[Iwamoto+2022 (NTT+Doshisha-U)]

- Usually degrades ASR performance

→ Quantitative investigation of enhanced speech by 1-ch DNN:

$$\underbrace{\widehat{\mathbf{s}}_{\omega} = \mathbf{s}_{\text{target}}}_{\text{Speech}} + \underbrace{\omega_{\text{noise}}\ \mathbf{e}_{\text{noise}}}_{\text{Residual noise}} + \underbrace{\omega_{\text{artif}}\ \mathbf{e}_{\text{artif}}}_{\text{Artifact (nonlinear distortion)}}$$

cf.) BSS_EVAL [Vincent+2006]



$\mathbf{e}_{\text{artif}}$ has more impact than $\mathbf{e}_{\text{noise}}$ [Iwamoto+2022 (NTT)]

Same tendency in human intelligibility [Araki+2023 (NTT)]

K. Iwamoto, T. Ochiai, et al., “How bad are artifacts?: Analyzing the impact of speech enhancement errors on ASR,” Interspeech2022.

S. Araki et al.,”Impact of Residual Noise and Artifacts in Speech Enhancement Errors on Intelligibility of Human and Machine, “ Interspeech2023.

How to reduce $\mathbf{e}_{\text{artif}}$ for 1-ch SE

- Artifact boosted training loss [Ochai+2024 (NTT)]

$$\mathcal{L}_{\text{AB-SDR}} = -10 \log_{10} \frac{\|\mathbf{s}_{\text{target}}\|^2}{\|\mathbf{e}_{\text{interf}} + \mathbf{e}_{\text{noise}} + \alpha \mathbf{e}_{\text{artif}}\|^2}$$

- Observation adding

  [Iwamoto+2022 (NTT)]

$$\hat{\mathbf{s}} \leftarrow \hat{\mathbf{s}} + \omega_{obs} \mathbf{x}$$

- Joint train of SE and ASR

  [Iwamoto+2024 (NTT)]

	SAR [dB]↑	WER [%]↓
Obs. (No SE)	∞	15.9
SDR-loss (Conv.)	14.8	14.8
Artifact-loss (Prop.)	16.7	13.0
+ Obs-add (Prop.)	17.1	12.8

Improve

T. Ochiai, et al., “Rethinking Processing Distortions: Disentangling the Impact of Speech Enhancement Errors on Speech Recognition Performance,“ IEEE TASLP, (to appear)
K. Iwamoto, T. Ochiai, et al., “How bad are artifacts?: Analyzing the impact of speech enhancement errors on ASR,” Interspeech2022.
K. Iwamoto, T. Ochiai, et al., “How Does End-To-End Speech Recognition Training Impact Speech Enhancement Artifacts?,“ ICASSP2024.

Summary & Challenges

NTT

Key technologies of frontend for conversation speech processing

- **Mask-based beamformer** is widely adopted
- For handling **various recording conditions**
  - **Blind** mask estimation: Spatial feature clustering
  - **Arbitrary number of speakers**: Speaker Diarization, Target speech extraction
  - **Dynamic** conditions: Beamformer for moving speakers

Remaining challenges

- Light weight, low latency, online
- Artifact-free 1-ch speech enhancement
- Simulate/Measure RIRs of moving speakers for training data augmentation

Special thanks to

Signal processing research group members,

alumni (especially Prof. N. Ito and video performers!),

and collaborators

Prof. J. Cernocky, Dr. K. Zmolikova*, Dr. M. Diez, Dr. F. Landini, Dr. A. Silnova, Dr. L. Burget
(Brno University of Technology) (*Currently with Meta)

Mr. K. Iwamoto and S. Katagiri (Doshisha Univ.)