Data-independent Beamforming for End-to-end Multichannel Multi-speaker ASR

Can Cui, Paul Magron, Mostafa Sadeghi, Emmanuel Vincent

IEEE MMSP - Beijing, September 21st, 2025

Problem statement

Meeting transcription

- Use-case: a multi-person meeting - and an unhappy secretary.
- Main goal: who said what and when.

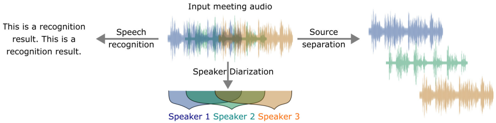
Problem statement

Meeting transcription

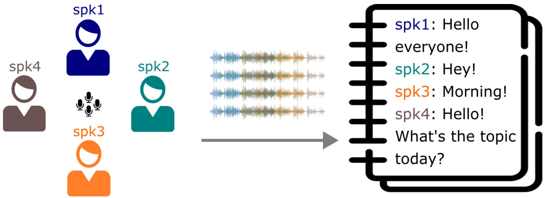
- Use-case: a multi-person meeting - and an unhappy secretary.
- Main goal: who said what and when.

Related tasks

- Source separation: extract one signal for each speaker.
- Diarization: knowing who speaks when.
- Automatic speech recognition (ASR): transcribe the content into text.



Multichannel ASR



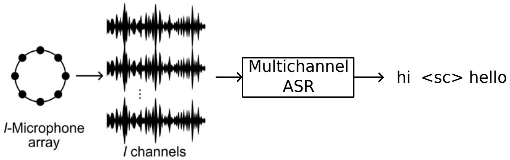
Key challenges

- Overlapping speakers, noise, reverberation.
- How to optimally exploit spatial information.
- Combination of modules and their input-output representations.

Existing strategies

Multichannel ASR using raw inputs (Yu et al. 2023).

- ▷ No need for an extra separation stage.
- ▷ Sensitive to noise, reverberation, and overlapping speakers.



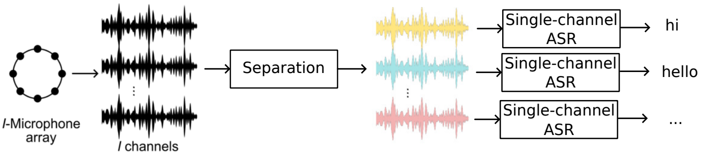
Existing strategies

Multichannel ASR using raw inputs (Yu et al. 2023).

- ▷ No need for an extra separation stage.
- ▷ Sensitive to noise, reverberation, and overlapping speakers.

Single-channel ASR using preprocessed signals (Raj et al. 2021; Masuyama et al. 2023).

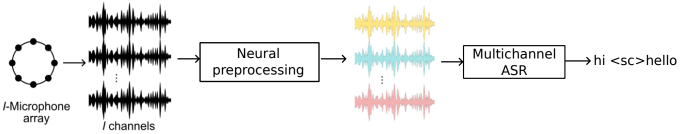
- ▷ Better performance / robustness.
- ▷ But the ASR system does not optimally use the spatial information.



Proposed approach

Alternative multichannel ASR systems use processed inputs (Kanda et al. 2023).

- A first network extracts a subset of signals.
- Better performance, but more computationally demanding.



Proposed approach

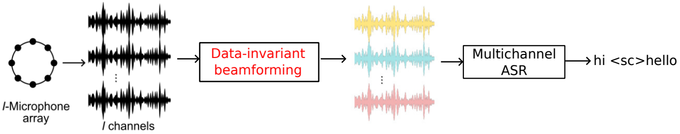
Alternative multichannel ASR systems use processed inputs (Kanda et al. 2023).

- ▷ A first network extracts a subset of signals.
- ▷ Better performance, but more computationally demanding.

Proposal

Processing multichannel inputs with **data-independent beamforming**.

- ▷ Extract signals from various positions that match the speakers' location.
- ▷ Enable the ASR system to leverage spatial / cross-channel information.
- ▷ Learning-free: no extra parameter / faster inference.



Proposed method

Setup

Multichannel signal: $\mathbf{x}(f) \in \mathbb{C}^I$ in the short-time Fourier transform domain.

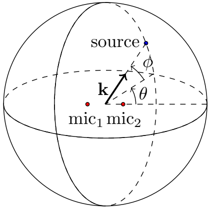
Setup

Multichannel signal: $\mathbf{x}(f) \in \mathbb{C}^I$ in the short-time Fourier transform domain.

Spherical coordinates with azimuth θ and elevation ϕ.

- Each point in space is described in terms of its unit vector:

$$\mathbf{k} = \begin{bmatrix} \cos\theta\cos\phi \\ \sin\theta\cos\phi \\ \sin\phi \end{bmatrix}$$



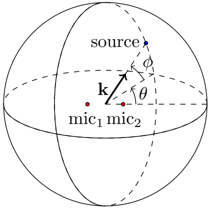
Setup

Multichannel signal: $\mathbf{x}(f) \in \mathbb{C}^I$ in the short-time Fourier transform domain.

Spherical coordinates with azimuth θ and elevation ϕ.

- Each point in space is described in terms of its unit vector:

$$\mathbf{k} = \begin{bmatrix} \cos\theta\cos\phi \\ \sin\theta\cos\phi \\ \sin\phi \end{bmatrix}$$



Sound propagation from a point $\mathbf{k}$ to a microphone $\mathbf{m}$ via a steering vector:

$$\mathbf{d}(\theta,\phi,f) = \begin{bmatrix} e^{-2\jmath\pi\mathbf{k}^T(\theta,\phi,f)\mathbf{m}_1/\lambda} \\ \vdots \\ e^{-2\jmath\pi\mathbf{k}^T(\theta,\phi,f)\mathbf{m}_I/\lambda} \end{bmatrix}$$

Angular sectors

Setting

- Each signal $\mathbf{x}(f)$ contains information from all directions.
- In practice (meetings), speakers are located around the microphone array.

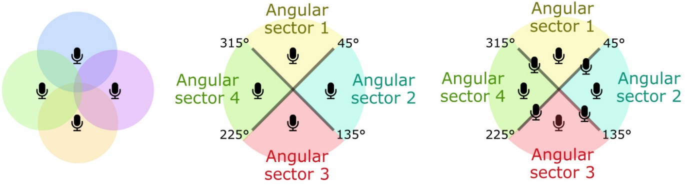
Angular sectors

Setting

- ▷ Each signal $\mathbf{x}(f)$ contains information from all directions.
- ▷ In practice (meetings), speakers are located around the microphone array.

Main idea

- ▷ Partition the space into S **angular sectors** Ψ_s, $s \in [1, S]$.
- ▷ Design a spatial filter $\mathbf{w}_s(f) \in \mathbb{C}^I$ for each angular sector.
- ▷ Filter the input data to get a set of sector-specific signals: $\mathbf{y}(f) = \mathbf{W}^H(f)\mathbf{x}(f) \in \mathbb{C}^S$



Data-independent beamforming

Data-independent beamforming: find a filter $\mathbf{w}(f)$ whose spatial response $\mathbf{w}(f)^H \mathbf{d}(\theta, \phi, f)$ is close to a predefined target response $b^{\text{tgt}}(\theta, \phi, f)$.

$$\arg\min_{\mathbf{w}(f)} \int_{\Omega} |\mathbf{w}(f)^H \mathbf{d}(\theta, \phi, f) - b^{\text{tgt}}(\theta, \phi, f)|^2 \cos\theta d\theta d\phi$$

Data-independent beamforming

Data-independent beamforming: find a filter $\mathbf{w}(f)$ whose spatial response $\mathbf{w}(f)^H\mathbf{d}(\theta, \phi, f)$ is close to a predefined target response $b^{\text{tgt}}(\theta, \phi, f)$.

$$\underset{\mathbf{w}(f)}{\arg\min} \int_{\Omega} |\mathbf{w}(f)^H\mathbf{d}(\theta, \phi, f) - b^{\text{tgt}}(\theta, \phi, f)|^2 \cos\theta d\theta d\phi$$

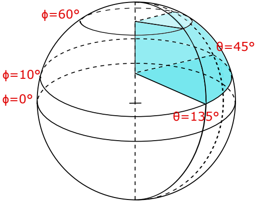
Extensive literature (Vincent et al. 2018) on possible solutions.

- ▷ Choice for the predefined target.
- ▷ Particular microphone array geometries (e.g., linear, circular).
- ▷ Numerical approximation schemes / filter design.

Proposed solution

Proposal: a sector-specific target response.

$$\forall s \in [1, S], \quad b_s^{\text{tgt}}(\theta, \phi, f) = \begin{cases} 1 & \text{if} \quad (\theta, \phi) \in \Psi_s, \\ 0 & \text{otherwise.} \end{cases}$$

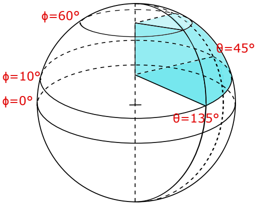


- The exact position (DOA) of each speaker is not needed.
- Expected to isolate speaker(s) located in sector s.

Proposed solution

Proposal: a sector-specific target response.

$$\forall s \in [1, S], \quad b_s^{\text{tgt}}(\theta, \phi, f) = \begin{cases} 1 & \text{if} \quad (\theta, \phi) \in \Psi_s, \\ 0 & \text{otherwise.} \end{cases}$$



- The exact position (DOA) of each speaker is not needed.
- Expected to isolate speaker(s) located in sector s.

Solution in closed form:

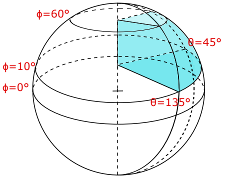
$$\mathbf{w}_s(f) = \left(\int_\Omega \cos\theta \times \mathbf{d}(\theta, \phi, f)\mathbf{d}(\theta, \phi, f)^H d\theta d\phi \right)^{-1} \times \int_{\Psi_s} \cos\theta \times \mathbf{d}(\theta, \phi, f) d\theta d\phi$$

- Only requires to approximate the integral (we use a step size of 1°).
- Applicable to every microphone array geometry.

Beamformer response

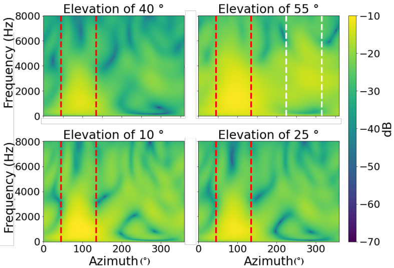
Proposed sectors

- $S = 4$, but it can be chosen freely.
- Elevation ϕ is kept in a realistic range $[10°, 60°]$.
- Azimuth range is split into four equal-size quadrants.



Filter response in the $[45°, 135°]$ angular sector, 4-microphones circular array.

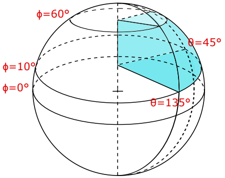
- High spatial response within the target sector.
- Degradation above 2 kHz.
- At high elevation, response from the opposite sector.



Beamformer response

Proposed sectors

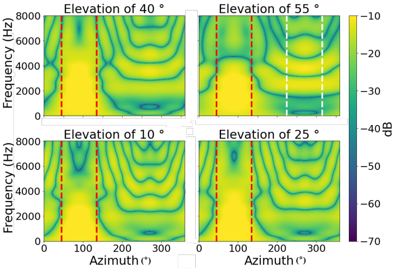
- ▷ $S = 4$, but it can be chosen freely.
- ▷ Elevation ϕ is kept in a realistic range $[10°, 60°]$.
- ▷ Azimuth range is split into four equal-size quadrants.



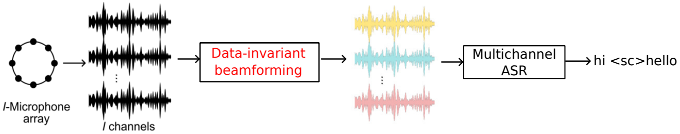
Filter response in the $[45°, 135°]$ angular sector, 4-microphones circular array.

- ▷ High spatial response within the target sector.
- ▷ Degradation above 2 kHz.
- ▷ At high elevation, response from the opposite sector.

Using 8 microphones: sharper response up to 4 kHz.



System overview



ASR system (Yu et al. 2023):

- Conformer-based encoder and decoder.
- Core part: a multi-frame cross-channel attention (MFCCA) mechanism to handle multichannel signals.
- Output: a single stream of transcript for all speakers, with a speaker change token <sc> between utterances.

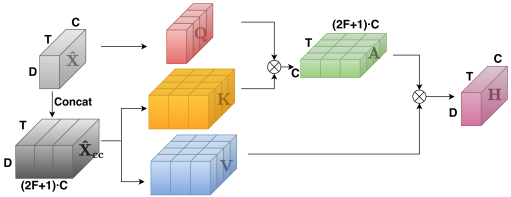


Image from (Yu et al. 2023)

Experiments

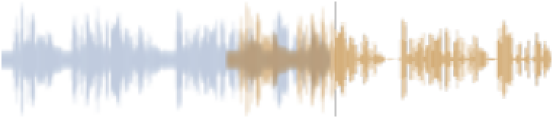
Protocol

1. Pretraining on a large set of *simulated* mixtures: Librispeech.
2. Fine-tuning on a smaller set of *real* meeting recordings: Real AMI.

Protocol

1. Pretraining on a large set of *simulated* mixtures: Librispeech.
2. Fine-tuning on a smaller set of *real* meeting recordings: Real AMI.

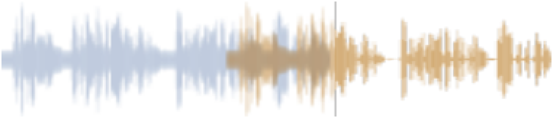
- ▷ Circular array, 4 or 8 microphones.
- ▷ Ensure no overlapping speakers at the beginning / end of each utterance.



Protocol

1. Pretraining on a large set of *simulated* mixtures: Librispeech.
2. Fine-tuning on a smaller set of *real* meeting recordings: Real AMI.

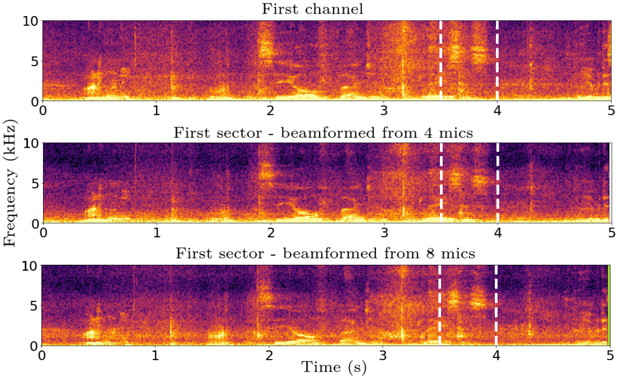
- ▷ Circular array, 4 or 8 microphones.
- ▷ Ensure no overlapping speakers at the beginning / end of each utterance.



Serialized output training

- ▷ Reference transcripts are build by inserting <sc> between references utterance labels.
- ▷ Sort speakers utterances by starting time (first-in first-out).
- ▷ Minimize the cross entropy between true / estimated serialized transcript.

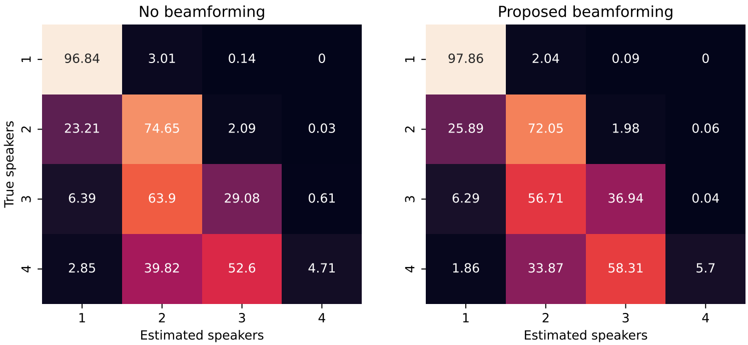
Results: signal quality



- ▷ Noise / reverberation reduction in the beamformed signals.
- ▷ Enhances a specific speaker (the one located in the sector) and reduces interferring speakers.
- ▷ Better formant preservation / enhancement when using 8 microphones over 4.

Speaker counting

Confusion score:



- ▷ Similar performance when there are few speakers (1 or 2).
- ▷ The proposed beamforming improves speaker counting for large numbers of speakers.

ASR performance

Word error rate (lower is better):

Beamforming	**# Mics**	**1-speaker**	**2-speaker**	**3-speaker**	**Average**
None	4	25.89	41.70	54.68	45.25
Proposed	4	24.52	**39.61**	52.19	43.14
	8	**22.96**	40.01	**49.59**	**41.64**

- ▷ Improvements of the beamforming over the unprocessed signals.
- ▷ Beamforming with 8 mics is better than with 4 mics.
  - ▷ Sector-wise enhancement is better, which in turns improves ASR performance.

Comparison to MVDR

MVDR = Minimum Variance Distortionless Response beamformer.

- ▷ Returns 1 reference signal, then single-channel ASR.

Comparison to MVDR

MVDR = Minimum Variance Distortionless Response beamformer.

- ▷ Returns 1 reference signal, then single-channel ASR.

Beamforming	# Mics	1-speaker	2-speaker	3-speaker	Average
None	4	25.89	41.70	54.68	45.25
MVDR	4	25.11	40.07	54.17	44.43
	8	26.35	42.13	55.89	46.08

- ▷ MVDR improves performance over no beamforming.

Comparison to MVDR

MVDR = Minimum Variance Distortionless Response beamformer.

- ▷ Returns 1 reference signal, then single-channel ASR.

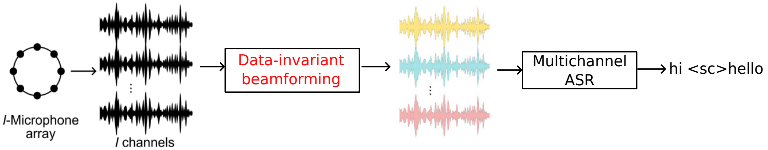
Beamforming	# Mics	1-speaker	2-speaker	3-speaker	Average
None	4	25.89	41.70	54.68	45.25
MVDR	4	25.11	40.07	54.17	44.43
	8	26.35	42.13	55.89	46.08
Proposed	4	24.52	**39.61**	52.19	43.14
	8	**22.96**	40.01	**49.59**	**41.64**

- ▷ MVDR improves performance over no beamforming.
- ▷ The proposed beamformer outperforms MVDR.

Conclusion

Conclusion

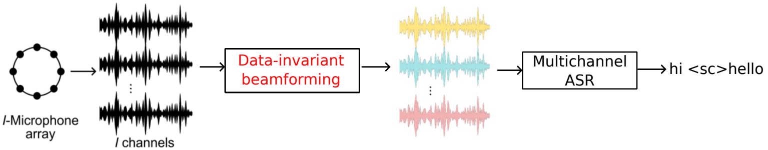
A data-independent beamformer for multichannel ASR



✓ Improves performance, training-free, flexible, fast.

Conclusion

A data-independent beamformer for multichannel ASR



✓ Improves performance, training-free, flexible, fast.

Perspectives

- Adapt and evaluate in more diverse situations: moving sources, alternative arrays.
- Automatic determination of the optimal number and/or geometry of the sectors.

References i

Kanda, N. et al. (2023). **"Vararray Meets T-Sot: Advancing the State of the Art of Streaming Distant Conversational Speech Recognition".** In: *Proc. of IEEE ICASSP*.

Masuyama, Y. et al. (2023). **"End-to-end integration of speech recognition, dereverberation, beamforming, and self-supervised learning representation".** In: *Proc. of IEEE SLT*.

Raj, D. et al. (2021). **"Integration of speech separation, diarization, and recognition for multi-speaker meetings: System description, comparison, and analysis".** In: *Proc. of IEEE SLT*.

Vincent, E. et al. (2018). **Audio Source Separation and Speech Enhancement.** John Wiley & Sons.

Yu, F. et al. (2023). **"MFCCA: Multi-frame cross-channel attention for multi-speaker ASR in multi-party meeting scenario".** In: *Proc. of IEEE SLT*.