Concurrent Speaker Detection: A Multi-microphone Transformer-Based Approach

Amit Eliav   Sharon Gannot

Faculty of Engineering, Bar-Ilan University, Israel

EUSIPCO 2024, Lyon

Concurrent Speaker Detection (CSD)

Goal

Identifying speakers' presence and overlapping activity in a given audio signal. Classifying into three classes:

'Noise', 'Single Speaker', and 'Concurrent Speakers'.

Concurrent Speaker Detection (CSD)

Goal

Identifying speakers' presence and overlapping activity in a given audio signal. Classifying into three classes:

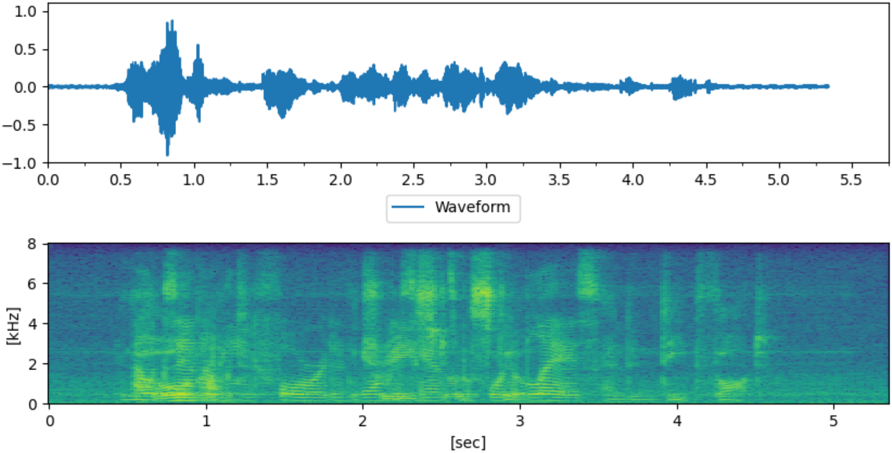
'Noise', 'Single Speaker', and 'Concurrent Speakers'.

Applications

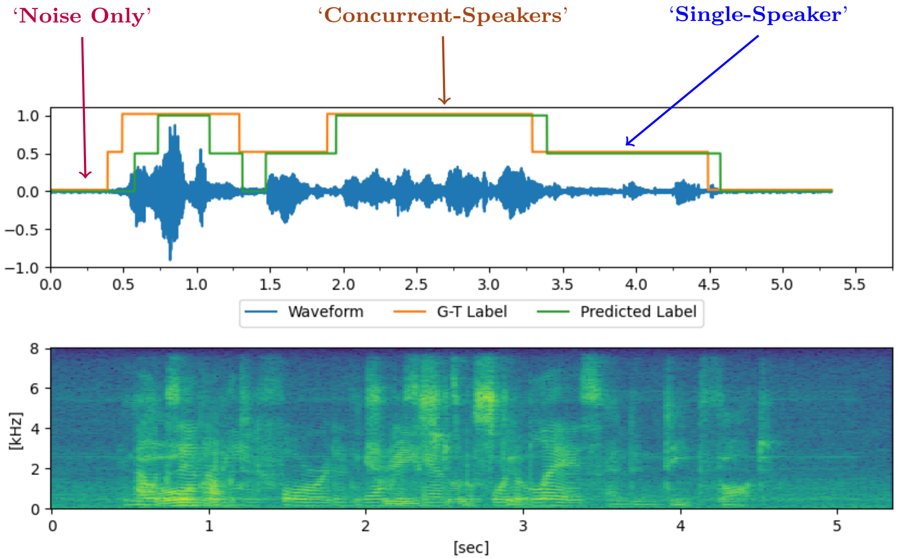
- Audio scene analysis and speech detection:
  - Speech detection
  - Speaker counting and speaker diarization
  - Speaker localization
  - Multi-microphone spatial processing in "cocktail party" scenarios
- Beamforming

CSD's output example

Can you spot all three classes in this signal?

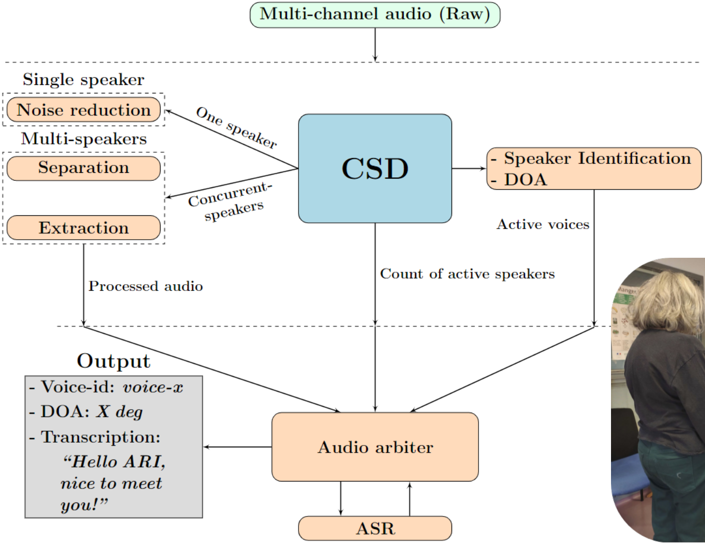


CSD's output example



Concurrent Speaker Detection

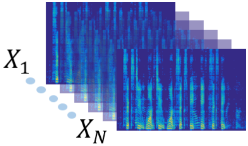
Applications: *'Audio Pipeline'* for robot Audition [Alameda-Pineda et al., 2024]:



Problem Formulation

Input Data

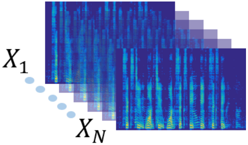
Let $X_i(\ell, k)$, $i = 1, \ldots, N$ represent the Short-Time Fourier transform (STFT) of the microphone signals, where N is the number of microphones, ℓ and k represent the frame index and the frequency index, respectively.



Problem Formulation

Input Data

Let $X_i(\ell, k)$, $i = 1, \ldots, N$ represent the Short-Time Fourier transform (STFT) of the microphone signals, where N is the number of microphones, ℓ and k represent the frame index and the frequency index, respectively.



CSD Task - A Multi-Class Classification

$$\mathrm{CSD}(\ell) = \begin{cases} \text{Class \#0} & \text{Noise only} \\ \text{Class \#1} & \text{Single-speaker activity} \\ \text{Class \#2} & \text{Concurrent-speaker activity} \end{cases} \tag{1}$$

Challenges in CSD

- Variability in speech characteristics:
  - Accents and dialects
  - Speaking rate and rhythm

Challenges in CSD

- Variability in speech characteristics:
  - Accents and dialects
  - Speaking rate and rhythm
- Environmental factors:
  - Different noise statistics
  - Presence of ‘babble noise’
  - Varying levels of reverberation

Challenges in CSD

- Variability in speech characteristics:
  - Accents and dialects
  - Speaking rate and rhythm
- Environmental factors:
  - Different noise statistics
  - Presence of 'babble noise'
  - Varying levels of reverberation
- Intricate activity patterns:
  - Varying number of concurrent speakers (2, 3, 4, or more)
  - Different overlap patterns and durations
  - Uneven energy distribution among overlapping speakers

Proposed Model Overview

Input Features and Labels

- The log-magnitude of the STFT of the audio signals
- Input of 0.5s audio signals ($[257 \times 32]$ time-frequency bins)
- Labels correspond to the middle 0.1s segment (6 time-frames)

Proposed Model Overview

Input Features and Labels

- The log-magnitude of the STFT of the audio signals
- Input of 0.5s audio signals ([257 × 32] time-frequency bins)
- Labels correspond to the middle 0.1s segment (6 time-frames)

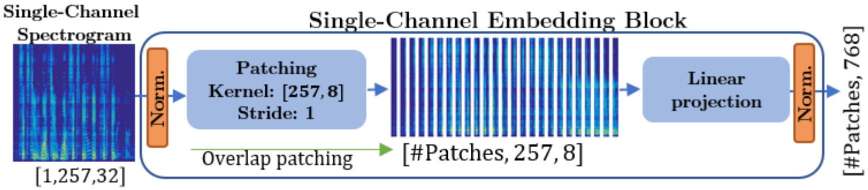
High-level Architecture

- Based on the Vision Transformer (ViT) model [Dosovitskiy et al., 2021]
- Modified the original ViT model to better suit **audio** requirements
- Handle both single-channel and multichannel audio
- Three main blocks: Embedding, Transformer, and Classification

Model Architecture: **Embedding Block**

Single-Channel Case

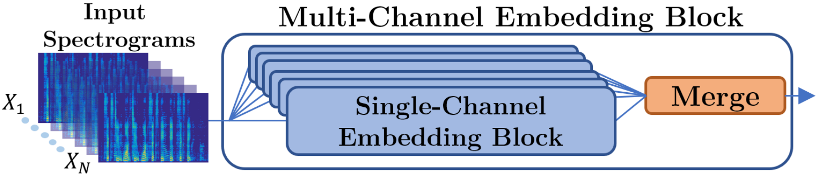
- Patches of $[257 \times 8]$: $[1, 257, 32] \rightarrow [\#Patches, 257, 8]$
- Overlapping patches
- Linear projection into a dimension of $D = 768$: $[\#Patches, 257, 8] \rightarrow [\#Patches, 768]$
- Dual normalization layers [Kumar et al., 2023]



Model Architecture: **Embedding Block**

Multi-Channel Case

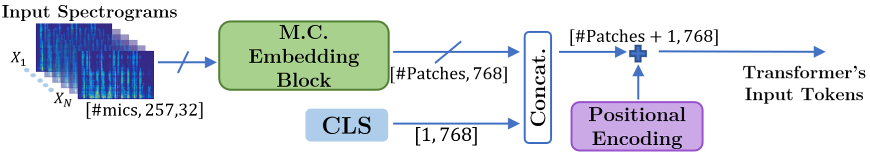
- Three merging strategies:
  - Summation
  - Siamese layers and averaging
  - Concatenation
- Chosen strategy:
  - N single-channel embedding blocks (no shared weights)
  - Concatenation - along the microphone dimension, which allows cross-channel attention.



Model Architecture: **Embedding Block**

Tokens for the Transformer block

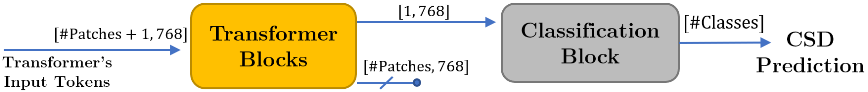
- Single/Multi embedding scheme
- Class-Token ('CLS')
- Positional encoding



Model Architecture: **Transformer & Classification Blocks**

Transformer Block

- Token dimension of $D = 768$
- 12 consecutive Multi-Head Attention (MHA) layers, with 12 heads each.



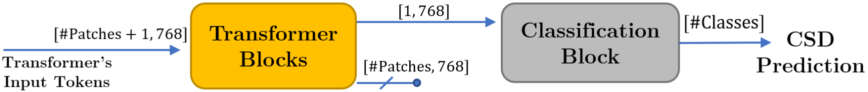
Model Architecture: **Transformer & Classification Blocks**

Transformer Block

- Token dimension of $D = 768$
- 12 consecutive Multi-Head Attention (MHA) layers, with 12 heads each.

Classification Block

- Consisting of fully connected layers (single hidden layer)
- Takes only the token corresponding to the 'CLS' token as input
  - Reduces computational cost
  - Unbiased classification towards any particular token
- Outputs the final classification logits



Comparison of ViT, AST, and Our CSD Model

Key comparison of ViT, AST [Gong et al., 2021], and our CSD model

Vision Transformer (ViT)

- Image classification
- 16×16 patch size
- RGB color images

Audio Spectrogram Transformer (AST)

- Based on ViT
- **Single**-microphone
- Similar to "grayscale" images
- 16×16 patch size
- Acoustic scene analysis

Our CSD Model

- Based on ViT
- **Multi**-microphone
- Similar to "color" images
- Optimized patch size: 257×8
- CSD task, with 3 classes

Objective Functions

Classification loss function

- Cross-Entropy (CE) loss
  - Class weights
  - Label-Smoothing (LS) regularization

Objective Functions

Classification loss function

- Cross-Entropy (CE) loss
  - Class weights
  - Label-Smoothing (LS) regularization

Are all mistakes equal?

Objective Functions

Classification loss function

- Cross-Entropy (CE) loss
  - Class weights
  - Label-Smoothing (LS) regularization

Are all mistakes equal? **No! → Cost-Sensitive (CS) loss**

We add CS loss as a regularization: [Galdran et al., 2020]

- Define 3×3 matrix by:
  - Giving more weight to frequent classification errors
  - The CSD task requirements
- Iterative training procedure:
  - Train the model with no CS loss
  - Modify the CS loss weights and retrain the model
  - Repeat until satisfactory convergence

CS weights

$$\begin{array}{cc} \textbf{True Label} & \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \end{bmatrix} \\ & \textbf{Predicted Label} \end{array}$$

Datasets

3 Real-World Datasets

The following real-world datasets were used:

- AMI [Carletta et al., 2006]
- AliMeeting [Yu et al., 2022]
- CHiME 5 [Barker et al., 2018]

Datasets' properties:

- Multi-microphone arrays
- English speakers (AMI, CHiME) and Mandarin speakers (AliMeeting) engaging in different tasks
- Different rooms and acoustic setups
- Fully transcribed with phrase-level resolution
- Highly unbalanced class distribution

Datasets

Class frequency [%]:

Dataset/Class	#0 'Noise only'	#1 'Single-speaker activity'	#2 'Concurrent-speaker activity'
Ali-Meeting	6.9%	67.2%	25.9%
AMI	16.8%	71.8%	11.4%
CHiME 5	20.5%	50.9%	28.6%

Datasets

Class frequency [%]:

Dataset/Class	#0 'Noise only'	#1 'Single-speaker activity'	#2 'Concurrent-speaker activity'
Ali-Meeting	6.9%	67.2%	25.9%
AMI	16.8%	71.8%	11.4%
CHiME 5	20.5%	50.9%	28.6%

Highly unbalanced datasets!

Addressed by:

- Tuning the loss function (Class weights, CS loss)
- Balancing the train set

Results

The confusion matrix results [%] normalized to the ground-truth labels:

Single-Microphone:

AMI

Ground-truth label \ Predicted label	0	1	2
0	78%	20%	2%
1	9%	75%	16%
2	1%	37%	62%

AliMeeting

Ground-truth label \ Predicted label	0	1	2
0	88%	11%	1%
1	10%	77%	13%
2	2%	30%	68%

CHiME

Ground-truth label \ Predicted label	0	1	2
0	55%	32%	13%
1	10%	51%	39%
2	3%	34%	63%

Multi-Microphone:

AMI

Ground-truth label \ Predicted label	0	1	2
0	80%	18%	2%
1	10%	74%	16%
2	2%	35%	63%

AliMeeting

Ground-truth label \ Predicted label	0	1	2
0	85%	13%	2%
1	7%	82%	11%
2	2%	28%	70%

CHiME

Ground-truth label \ Predicted label	0	1	2
0	73%	21%	6%
1	19%	51%	30%
2	8%	33%	59%

Comparative Results

To compare our results we first define an important related task:

Overlapped Speech Detection (OSD) Task - Binary Classification

$$\mathrm{OSD}(\ell) = \begin{cases} \text{Class } \#0 & \text{Non-overlapped speech} \\ \text{Class } \#1 & \text{Overlapped speech} \end{cases} \quad (2)$$

'Non-overlapped speech' holds both 'Noise only' and 'Single-speaker'

CSD
'Noise only'
'Single-speaker activity'
'Concurrent-speaker activity'

OSD
'Non-overlapped speech'
'Overlapped speech'

Comparative Results

OSD Comparison

A comparison between the proposed single- and multi-microphone variants and various competing methods in evaluating the performance on the OSD task.
Including Precision, Recall, and mean Average Precision (mAP) [%] measures on the AMI dataset.

Variant	Method	Precision	Recall	mAP
Single-channel	[Cornell et al., 2022]	N/A	N/A	59.1
	[Kyoung et al., 2023]	N/A	N/A	62.7
	[Bullock et al., 2020]	86.8	65.8	N/A
	pyannote 2.0 [Bredin and Laurent, 2021]	80.7	70.5	N/A
	Ours	**91.4**	**88.9**	**69.3**
Multichannel	[Zheng et al., 2021]	87.8	87	N/A
	[Cornell et al., 2022]	87.8	87	60.3
	Ours	**92.4**	**89**	**73.1**

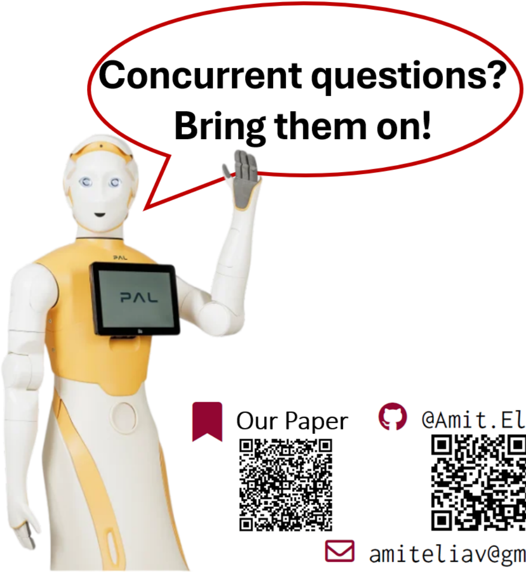
Classes #0 ('Noise only') and #1 ('Single-speaker activity') from the CSD model were aggregated to obtain OSD prediction.

Conclusions

Summary

- Multi-microphone transformer-based CSD model
- Extend the use of ViT and adapt it to the multi-microphone case
- Training scheme:
  - Weights based on class importance
  - Tuning the loss function (CS loss)
  - Addressing highly unbalanced datasets
- Evaluate the performance with three real-world datasets
- Advantages compared to existing methods

Thank you for listening!



Our Paper

@Amit.Eliav

Amit Eliav

amiteliav@gmail.com

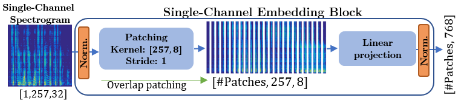
Image source: `pal-robotics.com/robots/ari`

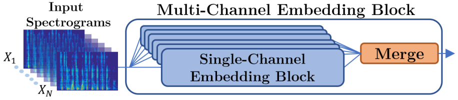
Model Architecture Breakdown

Quick Recap

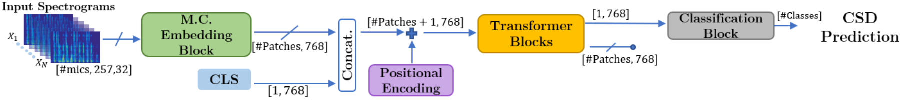
CSD Architecture Breakdown - Quick Recap

High-level model overview





Detailed model architecture:



References and Further Reading I

Alameda-Pineda, X., Addlesee, A., García, D. H., Reinke, C., Arias, S., Arrigoni, F., Auternaud, A., Blavette, L., Beyan, C., Camara, L. G., et al. (2024).
Socially pertinent robots in gerontological healthcare.
arXiv preprint arXiv:2404.07560.

Barker, J., Watanabe, S., Vincent, E., and Trmal, J. (2018).
The fifth 'chime' speech separation and recognition challenge: Dataset, task and baselines.
In *Proceedings Interspeech*, Hyderabad, India.

Bredin, H. and Laurent, A. (2021).
End-to-end speaker segmentation for overlap-aware resegmentation.
arXiv preprint arXiv:2104.04045.

Bullock, L., Bredin, H., and Garcia-Perera, L. P. (2020).
Overlap-aware diarization: Resegmentation using neural end-to-end overlapped speech detection.
In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7114–7118.

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D., and Wellner, P. (2006).
Machine Learning for Multimodal Interaction, chapter The AMI Meeting Corpus: A Pre-announcement, pages 28–39.
Springer Berlin Heidelberg.

Cornell, S., Omologo, M., Squartini, S., and Vincent, E. (2022).
Overlapped speech detection and speaker counting using distant microphone arrays.
Computer Speech & Language, 72:101306.

References and Further Reading II

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2021).
An image is worth 16x16 words: Transformers for image recognition at scale.
In *International Conference on Learning Representations (ICLR)*.

Galdran, A., Dolz, J., Chakor, H., Lombaert, H., and Ben Ayed, I. (2020).
Cost-sensitive regularization for diabetic retinopathy grading from eye fundus images.
In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 665–674.

Gong, Y., Chung, Y.-A., and Glass, J. (2021).
AST: Audio Spectrogram Transformer.
In *Proc. Interspeech*, pages 571–575.

Kumar, M., Dehghani, M., and Houlsby, N. (2023).
Dual PatchNorm.
Transactions on Machine Learning Research.

Kyoung, M., Jeon, H., and Park, K. (2023).
Audio-visual overlapped speech detection for spontaneous distant speech.
IEEE Access, 11:27426–27432.

Yu, F., Zhang, S., Fu, Y., Xie, L., Zheng, S., Du, Z., Huang, W., Guo, P., Yan, Z., Ma, B., Xu, X., and Bu, H. (2022).
M2MeT: The ICASSP 2022 multi-channel multi-party meeting transcription challenge.
In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

References and Further Reading III

Zheng, S., Zhang, S., Huang, W., Chen, Q., Suo, H., Lei, M., Feng, J., and Yan, Z. (2021).
Beamtransformer: Microphone array-based overlapping speech detection.
arXiv preprint arXiv:2109.04049.