The Cocktail Party Problem: WER we are, WER we are going

Presenter: Samuele Cornell

Email: *cornellsamuele@gmail.com*

See also "WER we are and WER we think we are" by Szymański et al.

Outline

- Transcribing multi-speaker conversational speech: tackling the Cocktail Party Problem
  - Why is an important problem
  - Why is it challenging ?
  - What is the current performance on the most challenging datasets

- How we try to solve this problem
  - Front-End methods
    - E.g., Speech Separation, target speaker extraction etc.
  - Back-End (ASR) methods
    - E.g., Serialized Output Training ASR, MIMO-Speech etc.

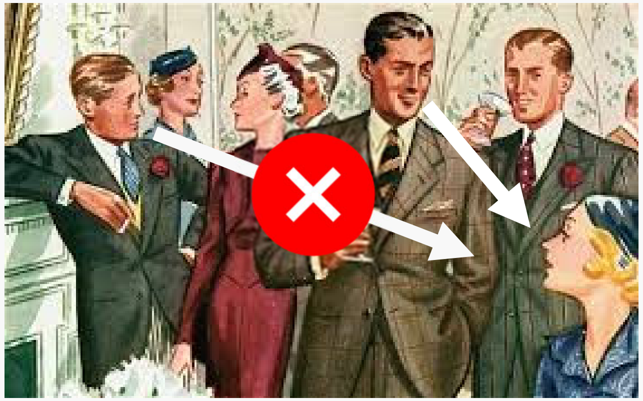
- Current Trends:
  - End-to-End Integration
    - Separate but together
  - Pretrained models
    - “There is no data like more data”
  - Iterative processing
    - Under-explored IMHO

Cocktail Party Problem/Effect

- The Cocktail Party Effect: auditory system has selective hearing.

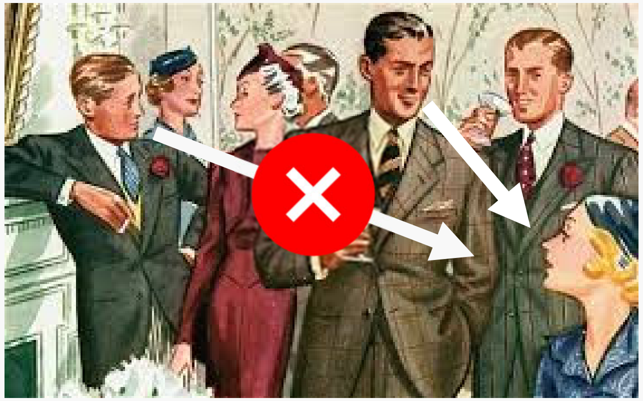
Cocktail Party Problem/Effect

- The Cocktail Party Effect: auditory system has selective hearing.
  - We can shift focus to different audio stimuli while ignoring others.



Cocktail Party Problem/Effect

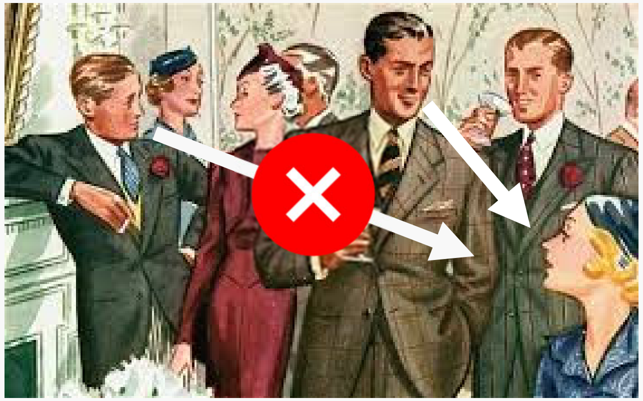
- The Cocktail Party Effect: auditory system has selective hearing.
  - We can shift focus to different audio stimuli while ignoring others.
    - See Prof. Mesgarani talk for details: "NIMA MESGARANI (COLUMBIA UNIVERSITY, USA): SPEECH PROCESSING IN THE HUMAN BRAIN MEETS DEEP LEARNING – YouTube" at JSALT2019.
    - Lower level auditory cortex separates audio in different streams then higher level (conscious level we can say) we decide on which to focus our attention.



Cocktail Party Problem/Effect

- The Cocktail Party Effect: auditory system has selective hearing.
  - We can shift focus to different audio stimuli while ignoring others.
    - See Prof. Mesgarani talk for details: "NIMA MESGARANI (COLUMBIA UNIVERSITY, USA): SPEECH PROCESSING IN THE HUMAN BRAIN MEETS DEEP LEARNING – YouTube" at JSALT2019.
    - Lower level auditory cortex separates audio in different streams then higher level (conscious level we can say) we decide on which to focus our attention.

- The Cocktail Party Problem:
  - We are nowhere near an automated system with such ability.
  - We are closer than 6 years ago for sure but significant challenges:
    - Reliability in real-world scenarios
    - "Continuous operation", low-latency and efficiency



Applications

- Most of audio applications since audio is “transparent” !

Applications

- Most of audio applications since audio is “transparent” !
    - Implicit or explicit separation are needed in the real world
        - e.g. ASR multi-condition training but other examples next
    - Fundamental problem to overcome for Conversational AI/Machine Listening

Applications

- Most of audio applications since audio is “transparent” !
  - Implicit or explicit separation are needed in the real world
    - e.g. ASR multi-condition training but other examples next
  - Fundamental problem to overcome for Conversational AI/Machine Listening

- Machine Listening
  - Sound Event Detection/Classification
  - Diarization (“who spoke when”)
  - Multi-Talker Automatic Speech Recognition
    - Meeting Transcription, live captioning etc.

Applications

- Most of audio applications since audio is “transparent” !
  - Implicit or explicit separation are needed in the real world
    - e.g. ASR multi-condition training but other examples next
  - Fundamental problem to overcome for Conversational AI/Machine Listening

- Machine Listening
  - Sound Event Detection/Classification
  - Diarization (“who spoke when”)
  - Multi-Talker Automatic Speech Recognition
    - Meeting Transcription, live captioning etc.

- “Human Listening”
  - Music applications
    - Music separation, genre classification etc.
  - Speech Enhancement/Separation
    - Hearing aids, hands-free communication etc.

Applications

- Most of audio applications since audio is “transparent” !
  - Implicit or explicit separation are needed in the real world
    - e.g. ASR multi-condition training but other examples next
  - Fundamental problem to overcome for Conversational AI/Machine Listening

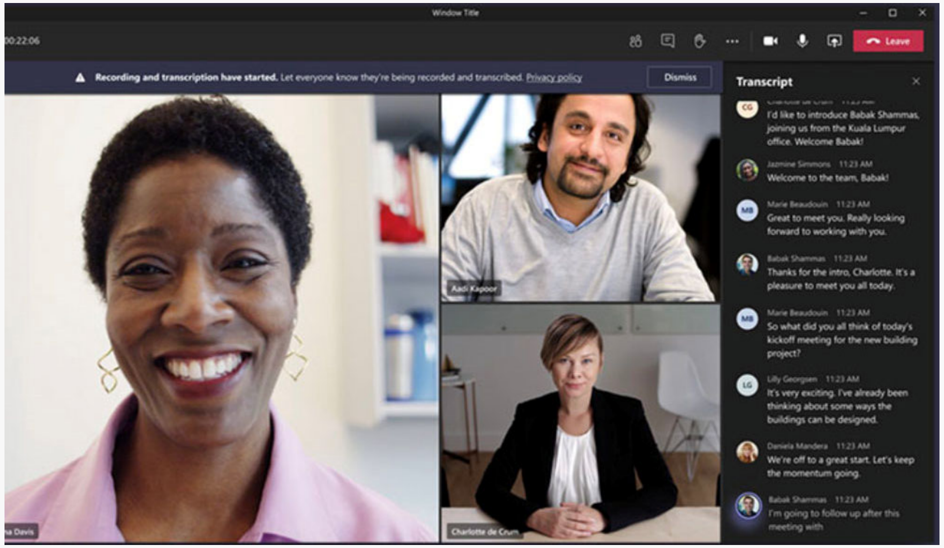
- Machine Listening
  - Sound Event Detection/Classification
  - Diarization (“who spoke when”)
  - **Multi-Talker Automatic Speech Recognition**
    - Meeting Transcription, live captioning etc.

- “Human Listening”
  - Music applications
    - Music separation, genre classification etc.
  - **Speech Enhancement/Separation**
    - Hearing aids, hands-free communication etc.

Transcribing The Cocktail Party

- Meeting Transcription
  - Diarization + multi-talker ASR with one or more devices

CHiME-5 Challenge Dataset



Microsoft Teams Live Transcriptions

Transcribing The Cocktail Party

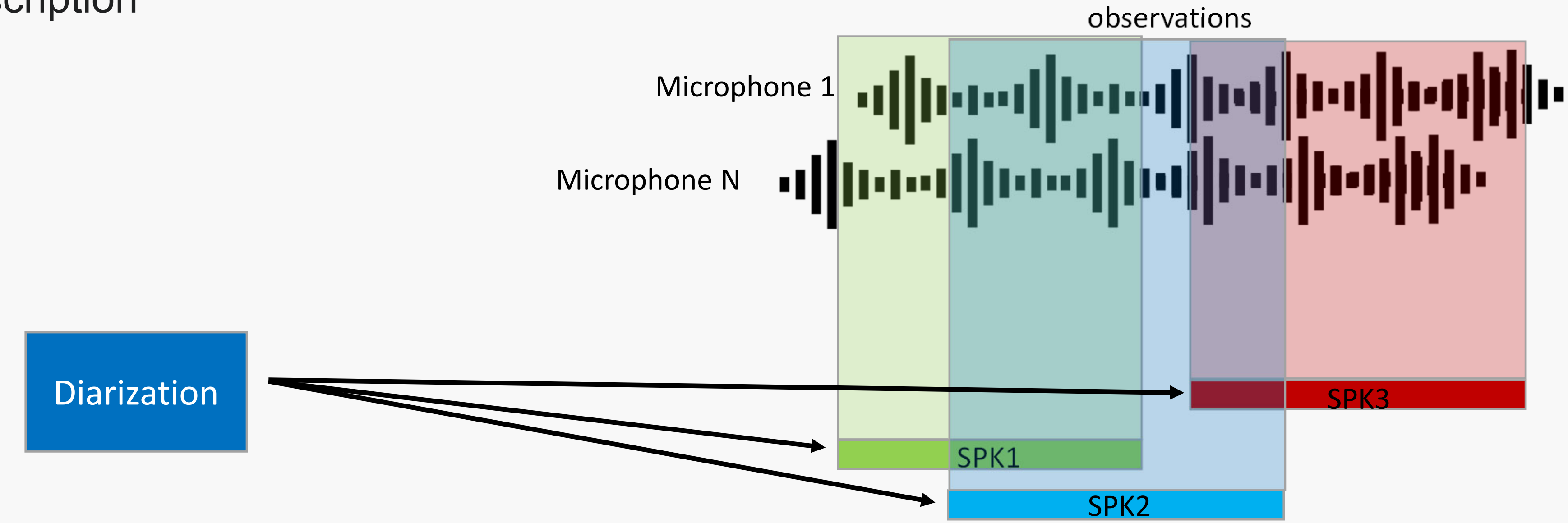
- Meeting Transcription
  - Diarization + multi-talker ASR with one or more devices

Microsoft Hybrid Meeting Transcription Demo

Meta EGO4D: smart glasses (live captions for translation)

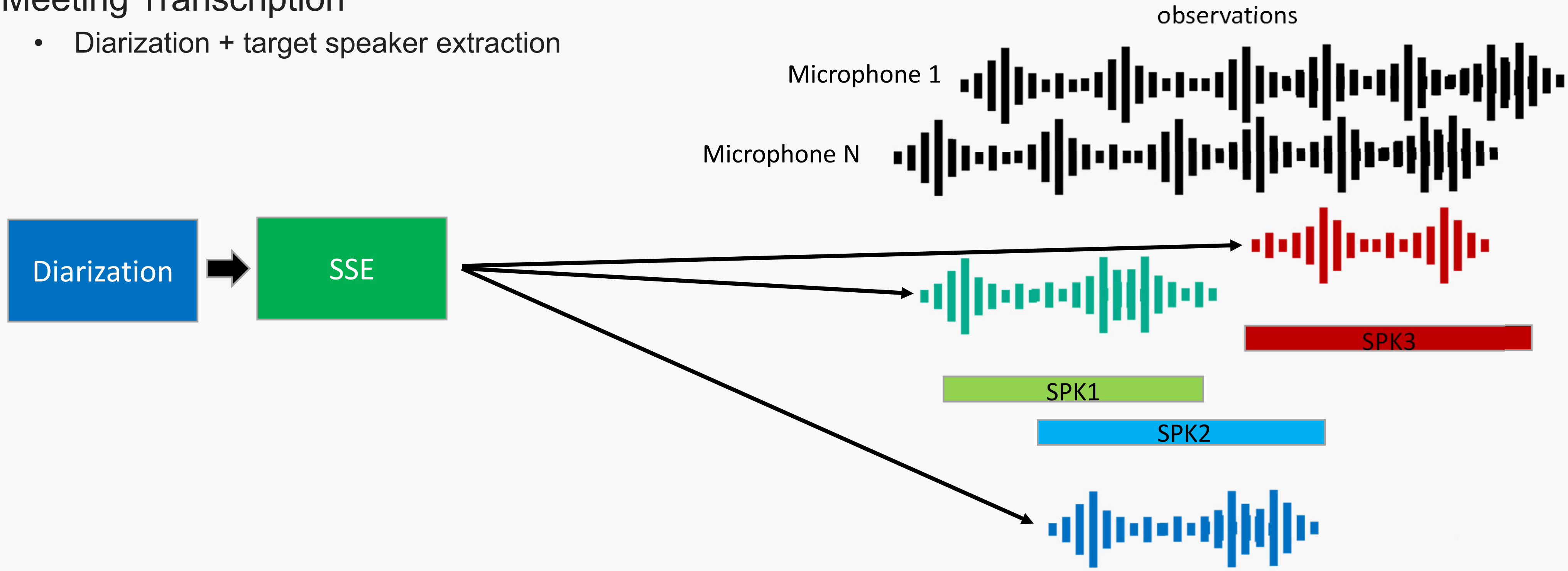
Transcribing The Cocktail Party

- Meeting Transcription
  - Diarization



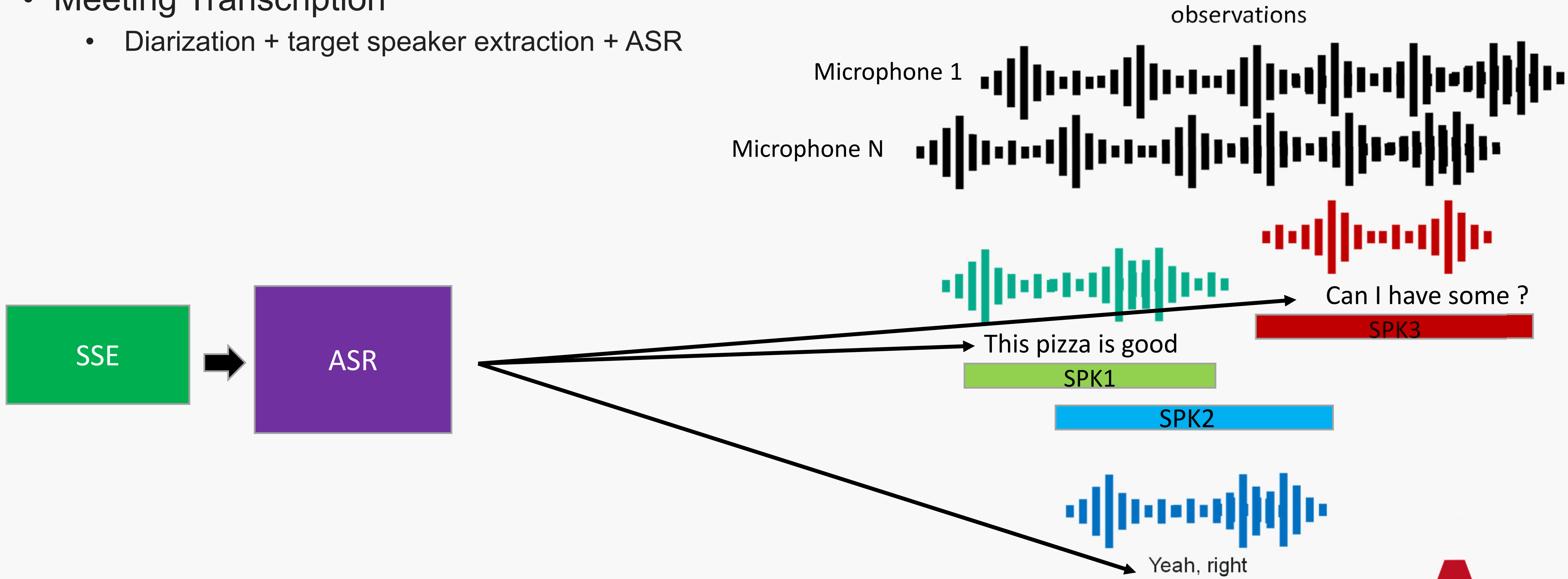
Transcribing The Cocktail Party

- Meeting Transcription
  - Diarization + target speaker extraction



Transcribing The Cocktail Party

- Meeting Transcription
  - Diarization + target speaker extraction + ASR



It’s a Challenging Problem

- Conversational Speech is hard !
  - Small speaker turns durations, laughs, many backchannel responses (“mmh”, “yeah”….) etc

It’s a Challenging Problem

- Conversational Speech is hard !
  - Small speaker turns durations, laughs, many backchannel responses (“mmh”, “yeah”….) etc
- Overlapped speech
  - Can exceed 15% of total speech e.g. CHiME-5/6 dinner party scenario

Table 1: Frame-level class frequency (%) for the speaker counting task on the AMI and CHiME-6 development and evaluation sets.

Class frequency		0-spk	1-spk	2-spk	3-spk	4-spk
AMI	dev	15.9	67.2	15.0	0.02	0.004
	eval	15.1	68.4	12.6	0.03	0.007
CHiME-6	dev	24.7	54.2	17.8	0.03	0.004
	eval	33.4	51.5	12.0	0.02	0.005

It's a Challenging Problem

- Conversational Speech is hard !
  - Small speaker turns durations, laughs, many backchannel responses ("mmh", "yeah"….) etc
- Overlapped speech
  - Can exceed 15% of total speech e.g. CHiME-5/6 dinner party scenario
- Colloquial language
  - More difficult to leverage text for LM

Table 1: Frame-level class frequency (%) for the speaker counting task on the AMI and CHiME-6 development and evaluation sets.

Class frequency		0-spk	1-spk	2-spk	3-spk	4-spk
AMI	dev	15.9	67.2	15.0	0.02	0.004
	eval	15.1	68.4	12.6	0.03	0.007
CHiME-6	dev	24.7	54.2	17.8	0.03	0.004
	eval	33.4	51.5	12.0	0.02	0.005

It's a Challenging Problem

- Conversational Speech is hard !
  - Small speaker turns durations, laughs, many backchannel responses ("mmh", "yeah"….) etc
- Overlapped speech
  - Can exceed 15% of total speech e.g. CHiME-5/6 dinner party scenario
- Colloquial language
  - More difficult to leverage text for LM
- Far-field Speech
  - Noisy/Reverberant Speech signal
  - Multiple devices help, but other problems:
    - Synchronization (clock drift)
    - Devices may be far, and processing multiple devices may be costly

Table 1: Frame-level class frequency (%) for the speaker counting task on the AMI and CHiME-6 development and evaluation sets.

Class frequency		0-spk	1-spk	2-spk	3-spk	4-spk
AMI	dev	15.9	67.2	15.0	0.02	0.004
	eval	15.1	68.4	12.6	0.03	0.007
CHiME-6	dev	24.7	54.2	17.8	0.03	0.004
	eval	33.4	51.5	12.0	0.02	0.005

WER we are now

- Many datasets are available in the literature for this kind of research:
  - AMI
  - LibriCSS (semi-simulated, only test and dev sets)
  - CHiME-5/6
  - AISHELL-4
  - Mixer 6 Speech
  - DipCo
  - AliMeeting (grand challenge at ICASSP 2022)
  - EGO4D
  - SPEAR (real and simulated)
  - Clarity Challenge 2 (simulated)

WER we are now

Current WER figures hint that we are far from reliable systems

	AMI eval WER
BigSSL [1]	17.7%
Pre-trained SOT [2]	21.2 %
VarArray + tSOT [3]	**15.5%**

[1] *Zhang, Yu, et al. "Bigssl: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition." IEEE Journal of Selected Topics in Signal Processing (2022).*
[2] *Kanda, Naoyuki, et al. "Large-scale pre-training of end-to-end multi-talker ASR for meeting transcription with single distant microphone." arXiv preprint arXiv:2103.16776 (2021).*
[3] *Kanda, Naoyuki, et al. "VarArray Meets t-SOT: Advancing the State of the Art of Streaming Distant Conversational Speech Recognition." arXiv preprint arXiv:2209.04974 (2022).*

WER we are now

Current WER figures hint that we are far from reliable systems

	CHiME-6 eval WER (oracle diarization)
BigSSL [1]	**31.0%**
USTC [2]	**31.0%**
Institute of Acoustics, CAS [3]	35.1%
STC-innovations Ltd, ITMO University [4]	35.8%

[1] *Zhang, Yu, et al. "Bigssl: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition." IEEE Journal of Selected Topics in Signal Processing (2022).*
[2] *Du, Jun, et al. "The USTC-NELSLIP systems for CHiME-6 challenge." CHiME-6 Workshop, Barcelona, Spain. 2020.*
[3] *Chen, Hangting, et al. "The IOA systems for CHiME-6 challenge." CHiME-6. 2020.*
[4] *Medennikov, Ivan, et al. "The STC system for the CHiME-6 challenge." CHiME 2020 Workshop on Speech Processing in Everyday Environments. 2020.*

WER we are now

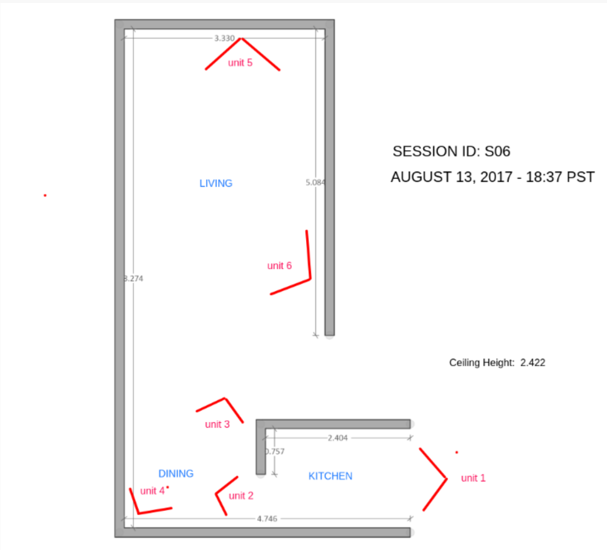
Current WER figures hint that we are far from reliable systems

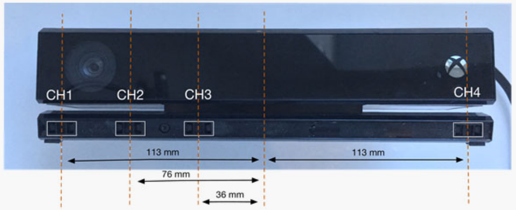
	CHiME-6 eval WER (non oracle diarization)
BigSSL [1]	n.a.
USTC [2]	68.5%
Institute of Acoustics, CAS [3]	n.a.
STC-innovations Ltd, ITMO University [4]	**44.5%**

[1] *Zhang, Yu, et al. "Bigssl: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition." IEEE Journal of Selected Topics in Signal Processing (2022).*
[2] *Du, Jun, et al. "The USTC-NELSLIP systems for CHiME-6 challenge." CHiME-6 Workshop, Barcelona, Spain. 2020.*
[3] *Chen, Hangting, et al. "The IOA systems for CHiME-6 challenge." CHiME-6. 2020.*
[4] *Medennikov, Ivan, et al. "The STC system for the CHiME-6 challenge." CHiME 2020 Workshop on Speech Processing in Everyday Environments. 2020.*

WER we are now

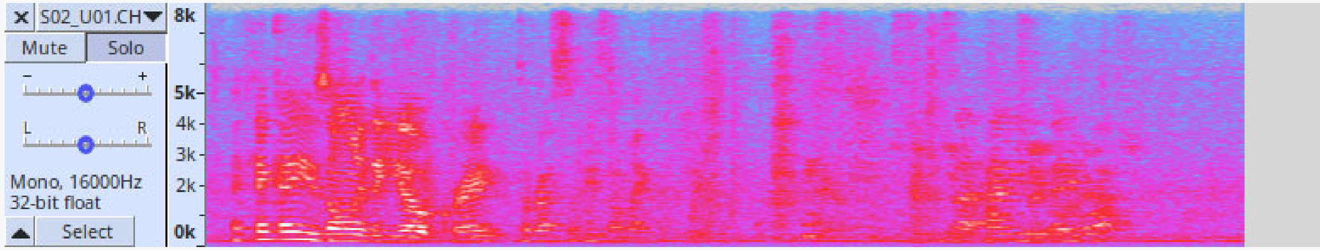
- An example from CHiME-6
- CHiME-6 Dataset:
    - 4 participants dinner party scenario
    - 6 Far-field Kinect array devices (4 microphones each)
        - + on-person close-talk binaural microphones for reference





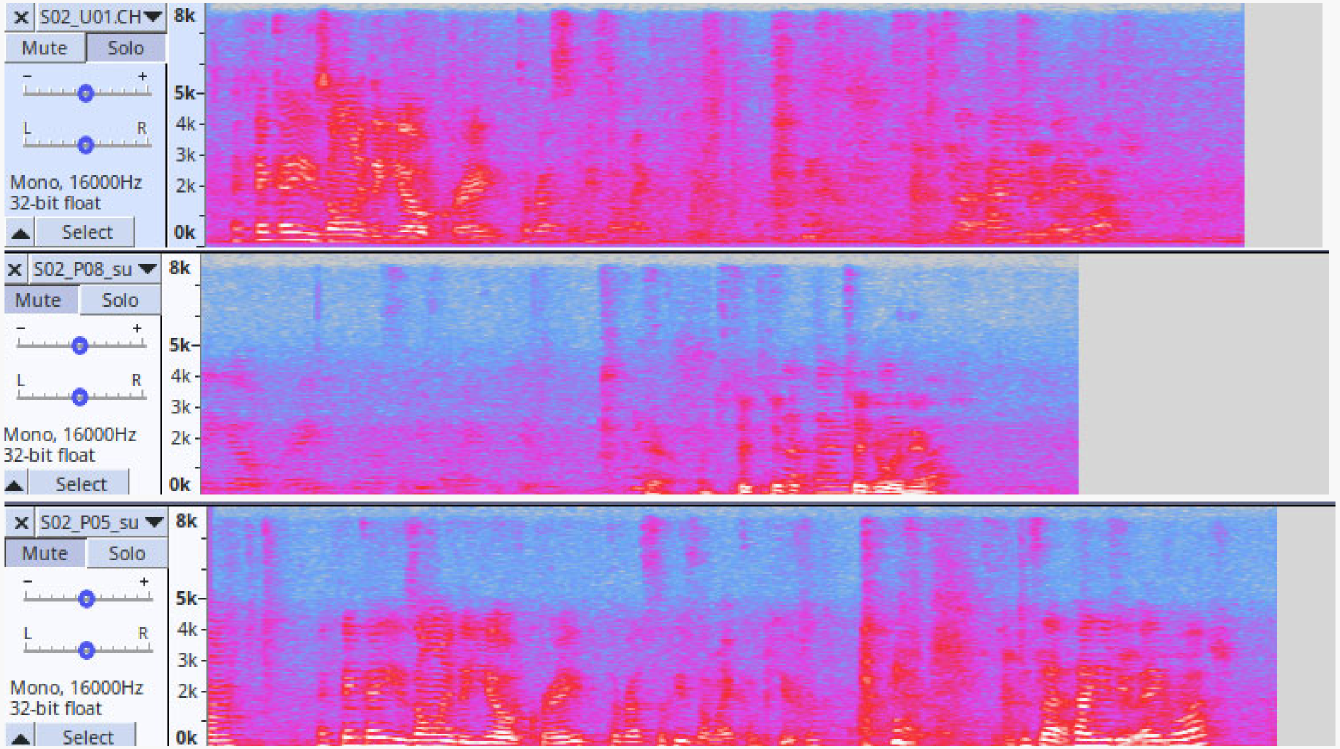
WER we are now

- An example from CHiME-6



WER we are now

- An example from CHiME-6

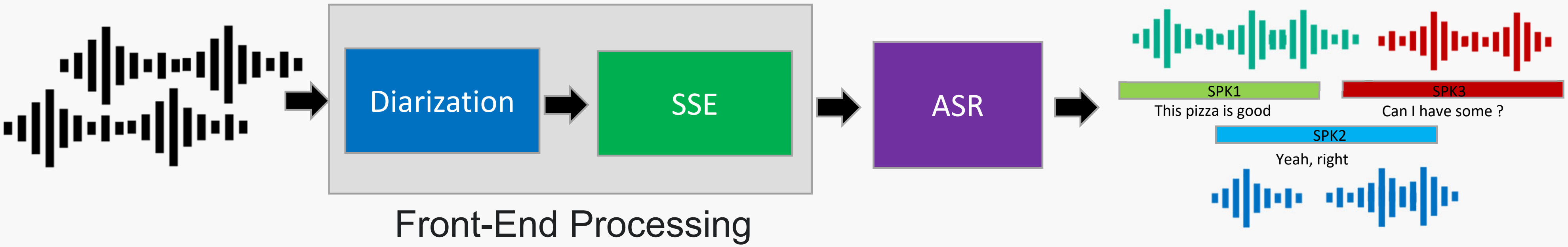


"Yeah, let's stick to the, take it with you."

"Okay. Um. I think I use only the yolk, right? The recipe is on my computer. [laughs] Is that how you do egg wash?"

How to address this problem

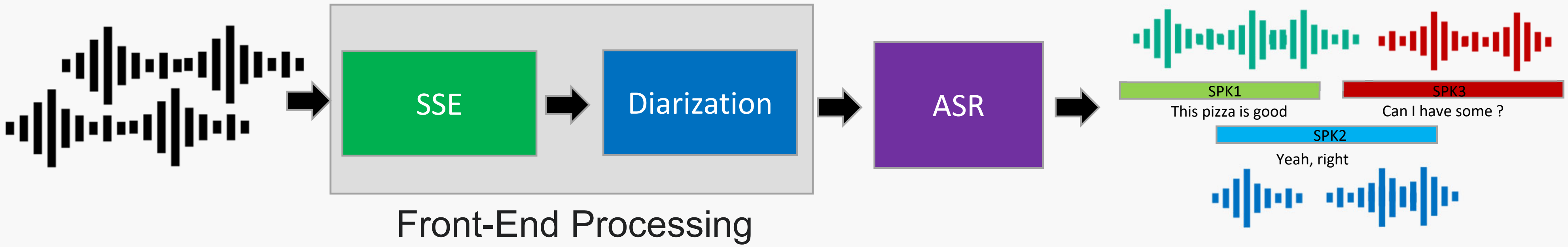
- “Multi-faceted problems require a multi-faceted solution“



- E.g. All best CHiME-5/6 systems use this pipeline:
  - Kanda, Naoyuki, et al. "Guided source separation meets a strong ASR backend: Hitachi/Paderborn University joint investigation for dinner party ASR." *arXiv preprint arXiv:1905.12230* (2019).
  - Medennikov, I., Korenevsky, M., Prisyach, T., Khokhlov, Y., Korenevskaya, M., Sorokin, I., ... & Romanenko, A. (2020). The STC system for the CHiME-6 challenge. In *CHiME 2020 Workshop on Speech Processing in Everyday Environments*.
  - Du, Jun, et al. "The USTC-NELSLIP systems for CHiME-6 challenge." *CHiME-6 Workshop, Barcelona, Spain*. 2020.

How to address this problem

- “Multi-faceted problems require a multi-faceted solution“

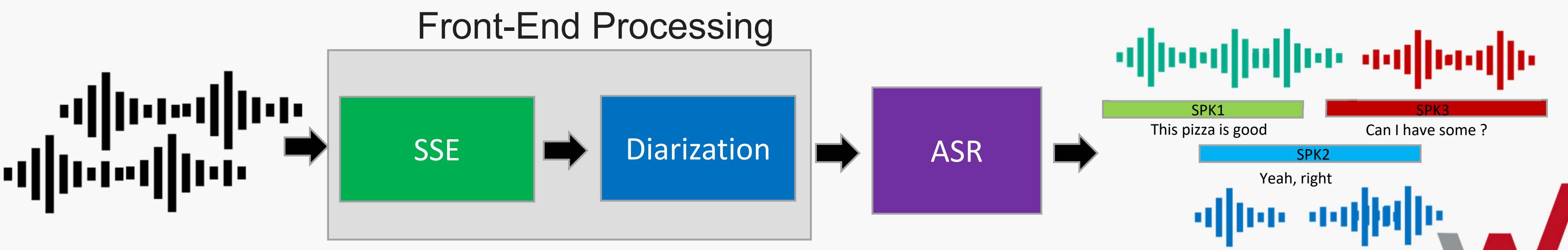


- E.g. SotA on AMI, VarArray + tSOT:
  - Kanda, Naoyuki, et al. "VarArray Meets t-SOT: Advancing the State of the Art of Streaming Distant Conversational Speech Recognition." *arXiv preprint arXiv:2209.04974* (2022).
    - But no diarization is performed and speakers attribution may be not consistent over whole recording. It may be added however.

How to address this problem

In general, we can divide the approaches to tackle multi-talker speech into:

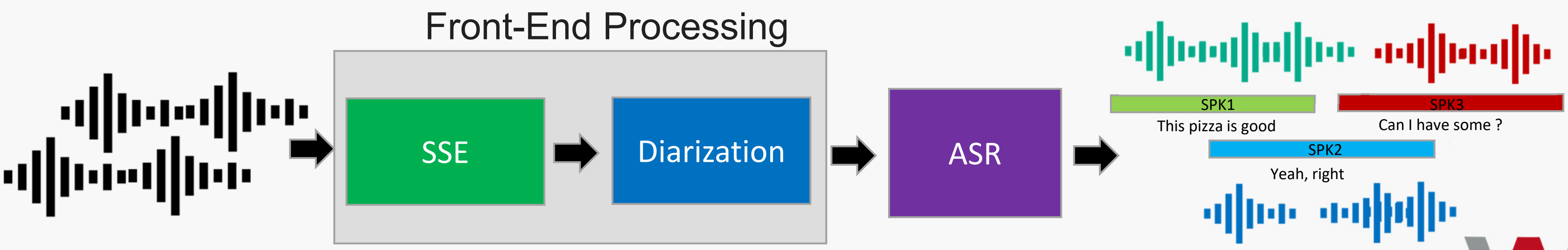
- Front-End methods



How to address this problem

In general, we can divide the approaches to tackle multi-talker speech into:

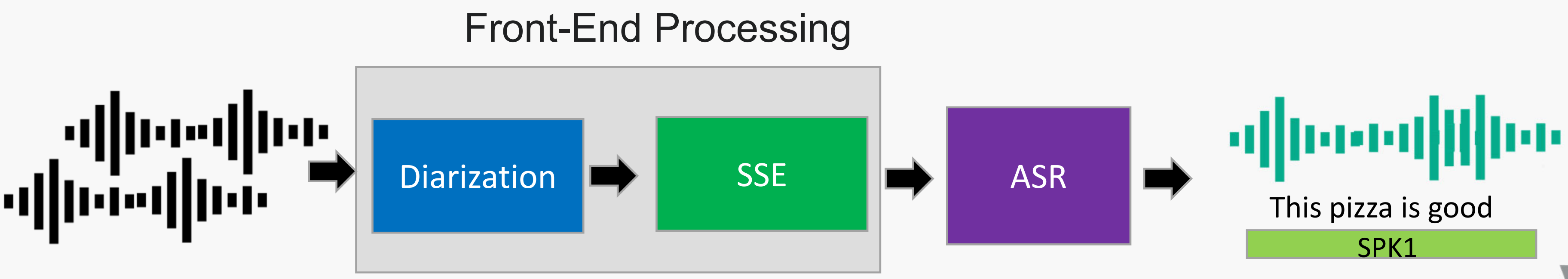
- Front-End methods
  - Speech Separation and Enhancement (SSE)
    - Continuous Speech Separation (CSS)



How to address this problem

In general, we can divide the approaches to tackle multi-talker speech into:

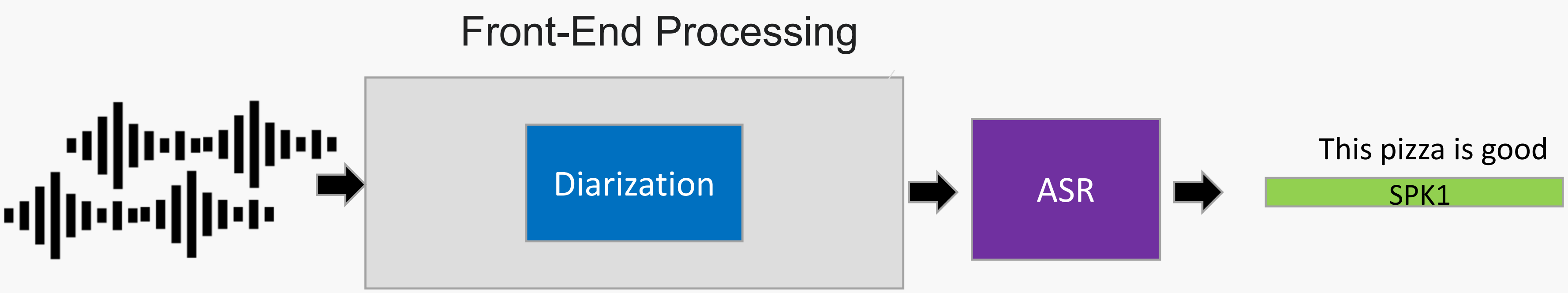
- Front-End methods
  - Speech Separation and Enhancement (SSE)
    - Continuous Speech Separation (CSS)
    - Target-speaker extraction



How to address this problem

In general, we can divide the approaches to tackle multi-talker speech into:

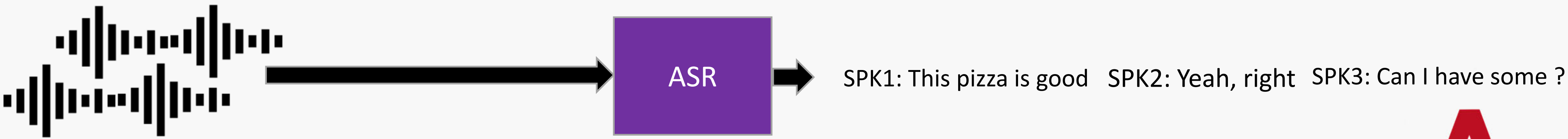
- Front-End methods
  - Speech Separation and Enhancement (SSE)
    - Continuous Speech Separation (CSS)
    - Target-speaker extraction
- Back-End (ASR) methods
  - Target-speaker ASR



How to address this problem

In general, we can divide the approaches to tackle multi-talker speech into:

- Front-End methods
  - Speech Separation and Enhancement (SSE)
    - Continuous Speech Separation (CSS)
    - Target-speaker extraction
- Back-End (ASR) methods
  - Target-speaker ASR
  - Serialized Output Training (SOT) and token-level SOT

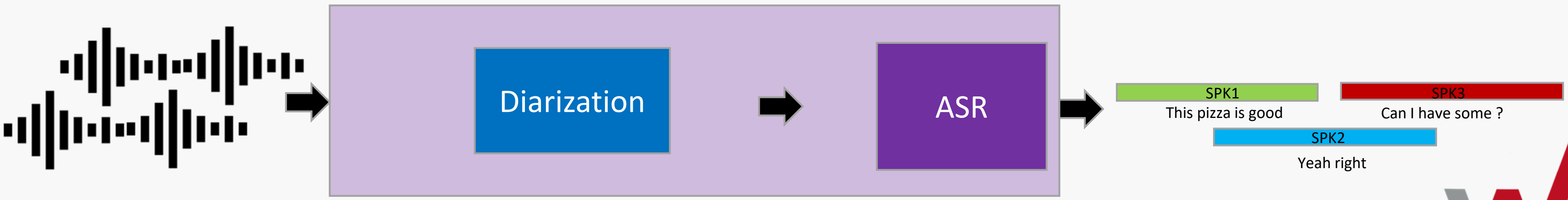


How to address this problem

In general, we can divide the approaches to tackle multi-talker speech into:

- Front-End methods
    - Speech Separation and Enhancement (SSE)
        - Continuous Speech Separation (CSS)
        - Target-speaker extraction
- Back-End (ASR) methods
    - Target-speaker ASR
    - Serialized Output Training (SOT) and token-level SOT

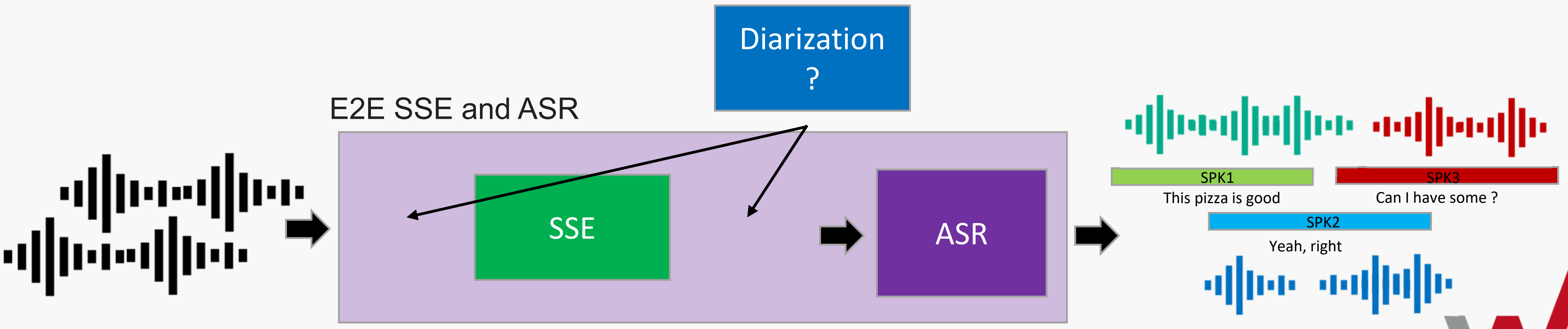
E2E Diarization and ASR



How to address this problem

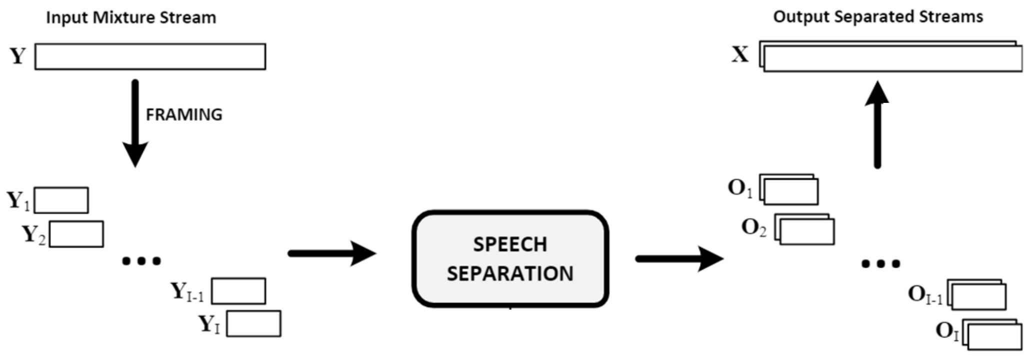
In general, we can divide the approaches to tackle multi-talker speech into:

- Front-End methods
  - Speech Separation and Enhancement (SSE)
    - Continuous Speech Separation (CSS)
    - Target-speaker extraction
- Back-End (ASR) methods
  - Target-speaker ASR
  - Serialized Output Training (SOT) and tSOT
  - Multi-channel ASR: e.g. MIMO-Speech, Directional-ASR



Continuous Speech Separation

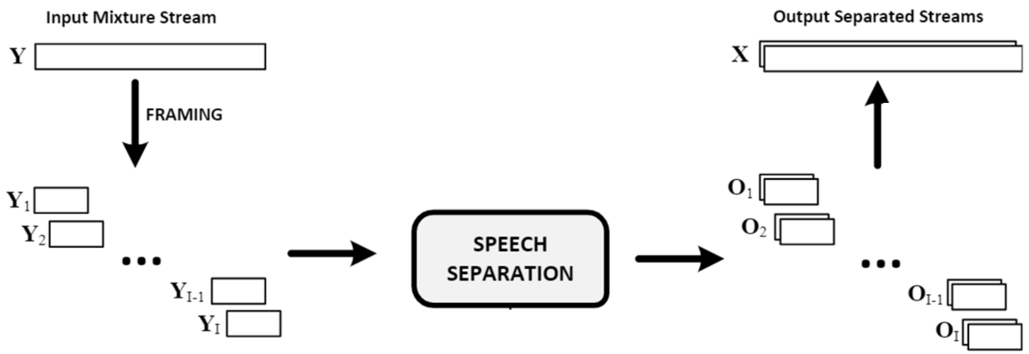
Current SotA speech separation models (e.g. DPRNN, ConvTasNet, SepFormer, TF-GridNet) are trained with a permutation invariant objective (PIT)



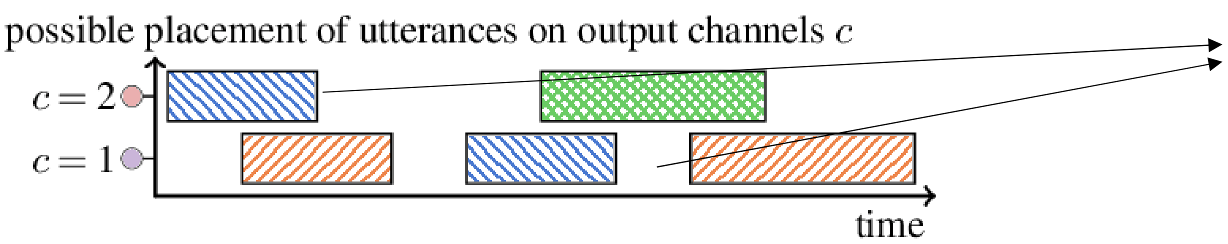
Separation model applied on rolling windows because we don't have infinite memory

Continuous Speech Separation

Current SotA speech separation models (e.g. DPRNN, ConvTasNet, SepFormer, TF-GridNet) are trained with a permutation invariant objective (PIT)



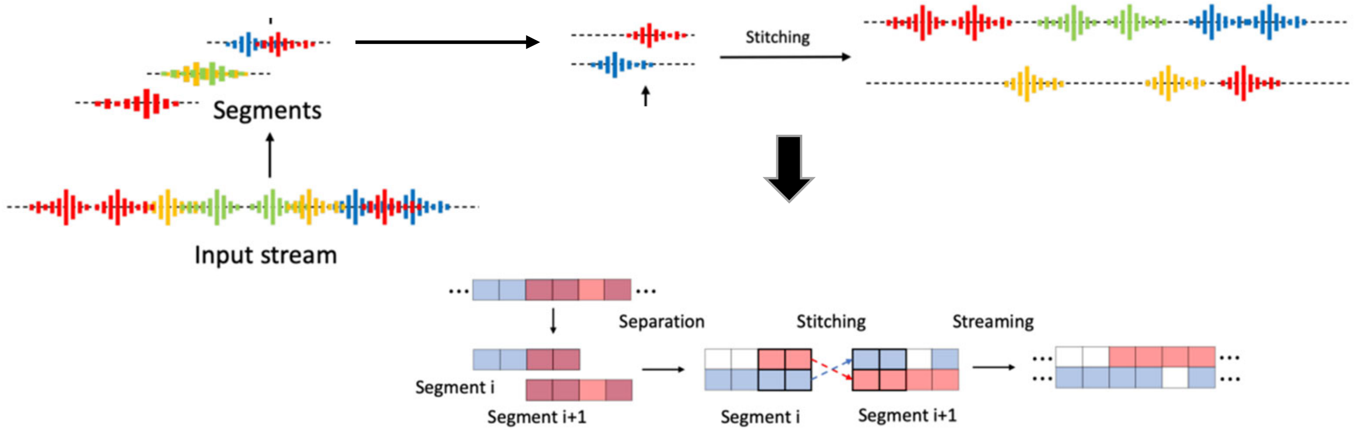
Separation model applied on rolling windows because we don't have infinite memory



Output placement might be not consistent. E.g. here for Speaker 1 (blue) on two consecutive windows.

Continuous Speech Separation

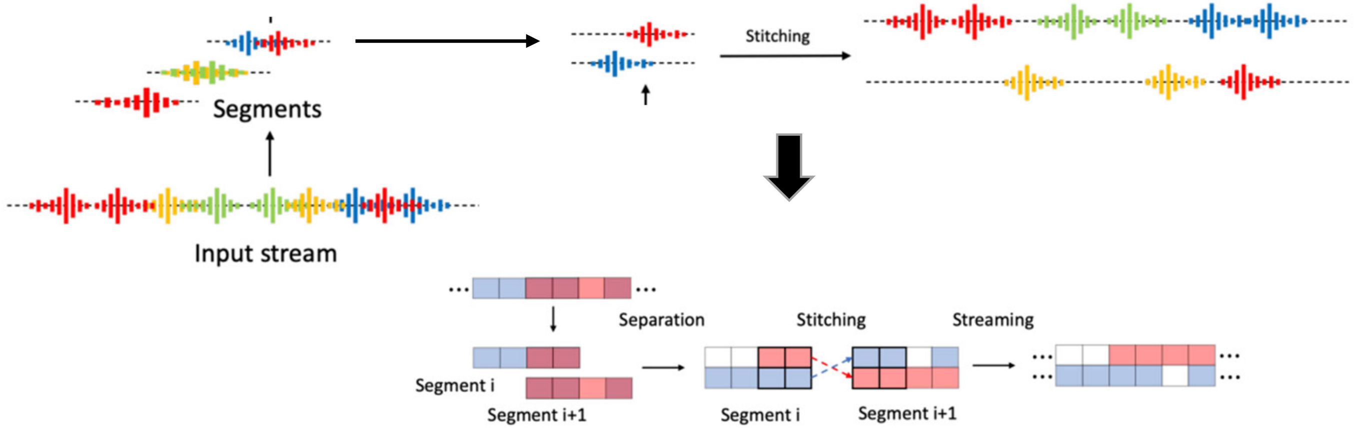
Solution: use overlapped windows and reorder based on a similarity measure the windows



Images from "Han, Cong, et al. "Continuous speech separation using speaker inventory for long multi-talker recording." *arXiv preprint arXiv:2012.09727* (2020)."

Continuous Speech Separation

Solution: use overlapped windows and reorder based on a similarity measure the windows

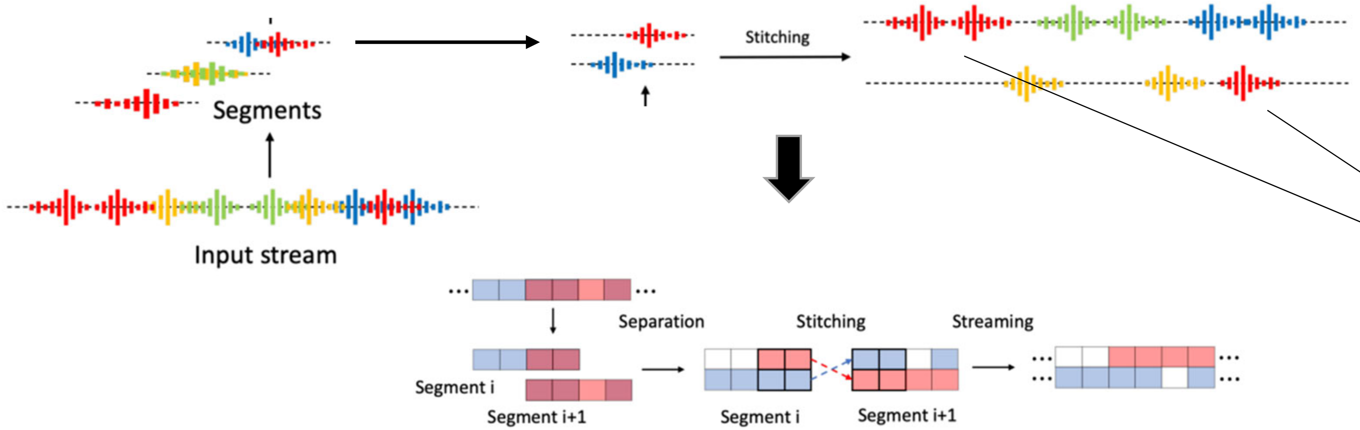


Here we need to swap the segments in output

Images from "Han, Cong, et al. "Continuous speech separation using speaker inventory for long multi-talker recording." *arXiv preprint arXiv:2012.09727* (2020)."

Continuous Speech Separation

Solution: use overlapped windows and reorder based on a similarity measure the windows



Here we need to swap the segments in output

NOTE:

If one window is smaller than the pause between one speaker utterances it looses ordering !

Images from "Han, Cong, et al. "Continuous speech separation using speaker inventory for long multi-talker recording." *arXiv preprint arXiv:2012.09727* (2020)."

Continuous Speech Separation

- CSS:
  - Original work: Chen, Zhuo, et al. "Continuous speech separation: Dataset and analysis." ICASSP, 2020.

- Can be used to perform diarization:
  - Speech Separation Guided Diarization:
    - Fang, Xin, et al. "A deep analysis of speech separation guided diarization under realistic conditions." *APSIPA ASC*, 2021.
  - Near SotA results on CALLHOME for two speakers
    - *Morrone, Giovanni, et al. "Leveraging Speech Separation for Conversational Telephone Speaker Diarization." arXiv (2022).*
  - NOTE: instead of CSS one can also use a streaming separation model as causal ConvTasNet, DPRNN or SkiM [1]

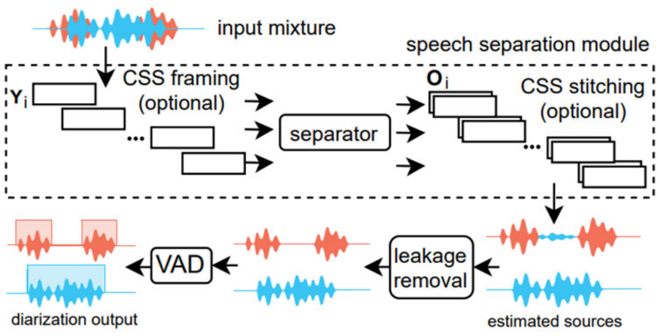
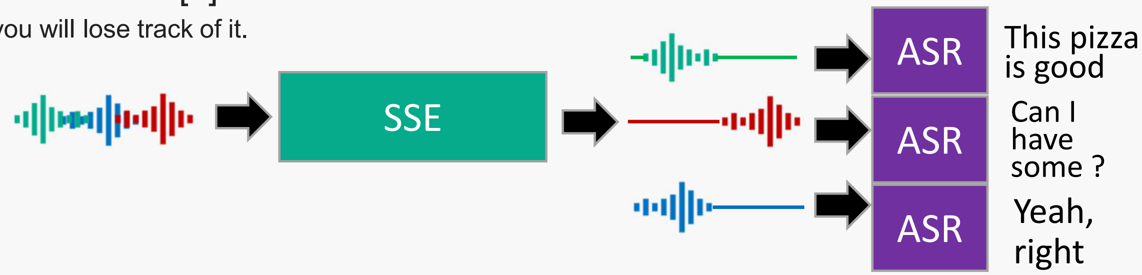


Fig. 1. General diagram for the SSGD method.

[1] *Li, Chenda, et al. "SkiM: Skipping Memory LSTM for Low-Latency Real-Time Continuous Speech Separation." ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) . IEEE, 2022.*

Continuous Speech Separation

- Cons
  - Difficult to optimize end-to-end with the ASR model:
    - Need to use so-called “utterance groups“ (groups of utterances that overlap)
      - You cannot truncate the transcripts (unless you use forced-alignment first !).
      - In CHiME-6 these groups can be several minutes long.
  - Not easy to handle arbitrary number of speakers
    - Output size is usually fixed: e.g. max number of local speakers 2 or 3.
      - One channel needs to be zero when there is only one speaker but in practice you may have leakage.
      - See for the leakage problem [1].
  - For long meetings diarization may still be necessary as outlined in [2].
    - As explained previously, if one speaker does not talk for long time, you will lose track of it.
      - [3] provides a partial solution



[1] Morrone, Giovanni, et al. "Leveraging Speech Separation for Conversational Telephone Speaker Diarization." arXiv, 2022.

[2] *Raj, Desh, et al. "Integration of speech separation, diarization, and recognition for multi-speaker meetings: System description, comparison, and analysis." SLT, 2021*

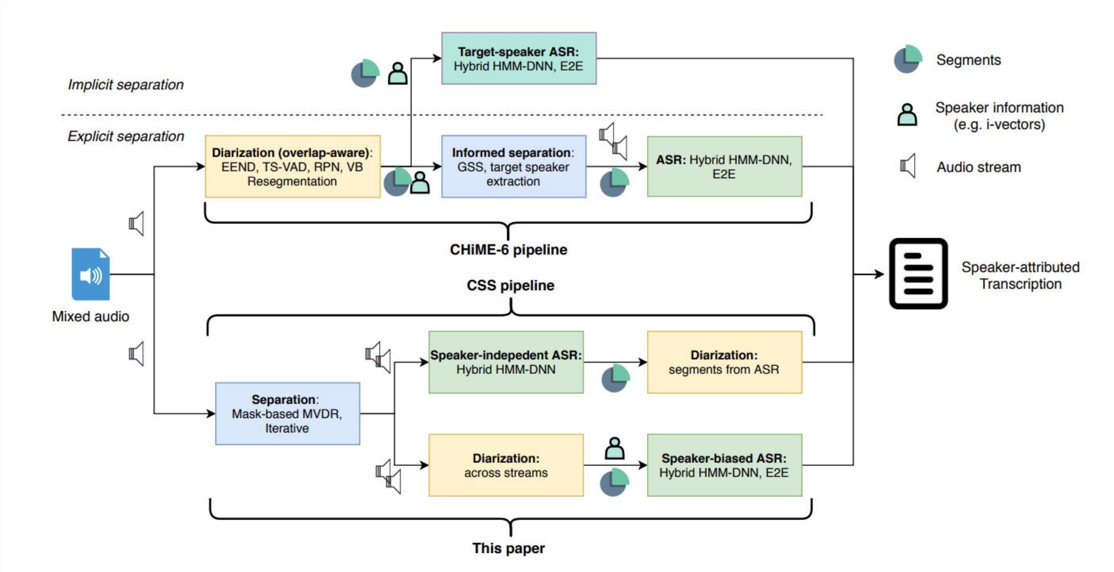
[3] *Han, Cong, et al. Continuous speech separation using speaker inventory for long multi-talker recording." arXiv, 2020.*

Continuous Speech Separation

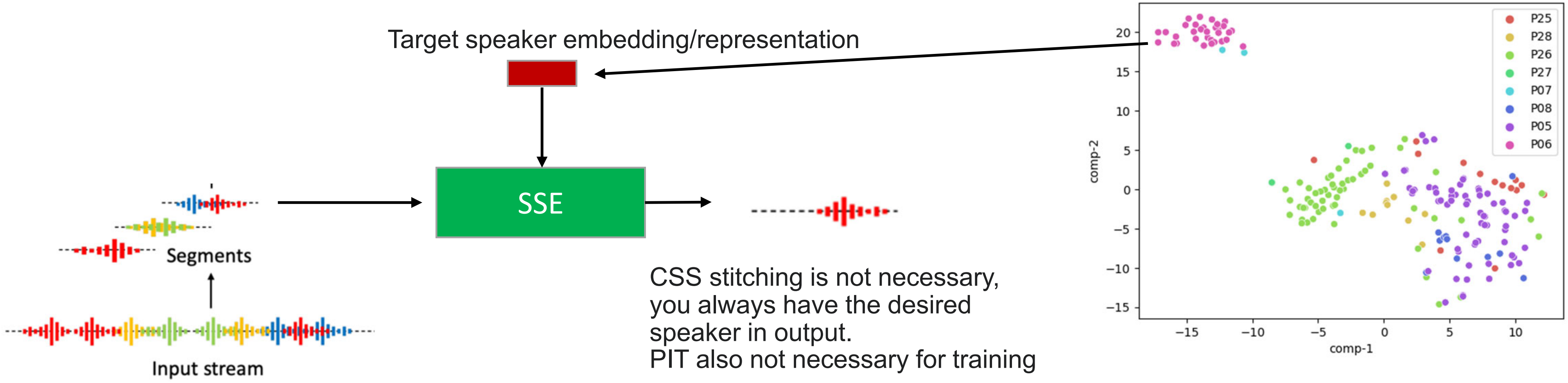
- Recent works confirmed CSS can be effective in real-world multi-speaker meetings:
  - Yoshioka, Takuya, et al. "VarArray: Array-geometry-agnostic continuous speech separation." *ICASSP*, 2022.

Continuous Speech Separation

- Recent works confirmed CSS can be effective in real-world multi-speaker meetings:
  - Yoshioka, Takuya, et al. "VarArray: Array-geometry-agnostic continuous speech separation.“ *ICASSP*, 2022.
  - *Raj, Desh, et al. "Integration of speech separation, diarization, and recognition for multi-speaker meetings: System description, comparison, and analysis." SLT, 2021.*

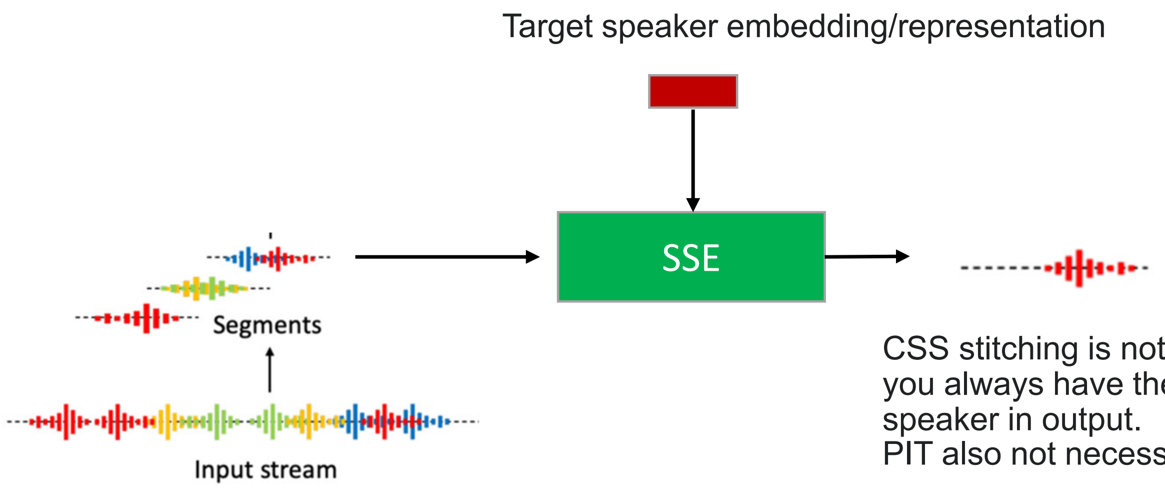


Target Speaker Extraction



CSS stitching is not necessary,
you always have the desired
speaker in output.
PIT also not necessary for training

Target Speaker Extraction



CSS stitching is not necessary,
you always have the desired
speaker in output.
PIT also not necessary for training

- SpeakerBeam: *Delcroix, Marc, et al. "End-to-End SpeakerBeam for Single Channel Target Speech Recognition." Interspeech. 2019*.
- Guided Source Separation (GSS): *Boeddeker, Christoph, et al. "Front-end processing for the CHiME-5 dinner party scenario." CHiME5 Workshop, 2018.*

Target Speaker Extraction

Pros:

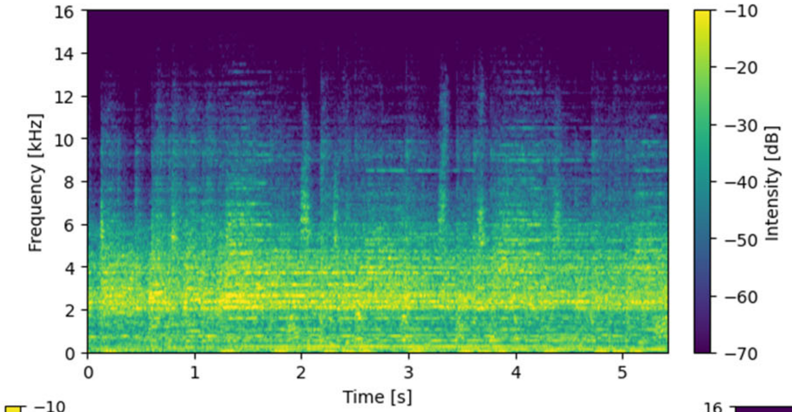
- No stitching required, conceptually easier inference.
- Easier also to optimize end-to-end with the ASR back-end.
  - Less memory requirements as you can truncate utterances from competing speakers without caring of utterance groups.

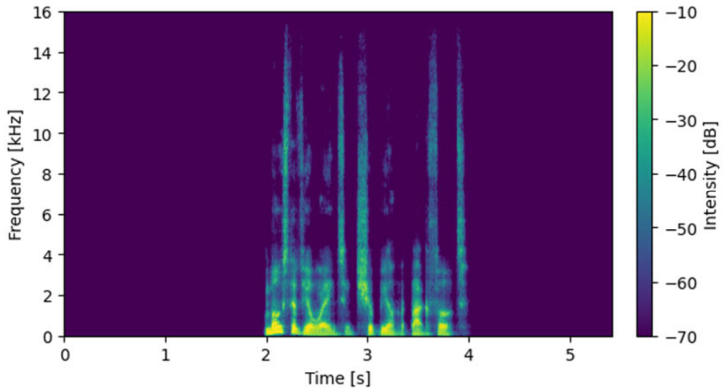
Cons:

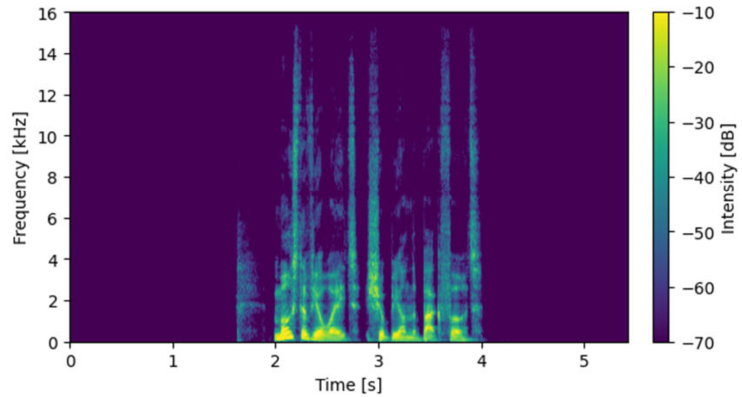
- Performance depends largely on diarization e.g. see CHiME-6 results:
  - Best systems are the ones with better diarization as everyone uses GSS [1].
    - “Chicken and egg problem” for diarization and target speaker extraction: Speaker extraction needs diarization but also diarization can in principle benefit from target speaker extraction.

[1] Guided Source Separation: Boeddeker, Christoph, et al. "Front-end processing for the CHiME-5 dinner party scenario." *CHiME5 Workshop, Hyderabad, India*. Vol. 1. 2018.

Front-End Methods: SSE is getting stronger !







Front-End Methods: SSE is getting stronger !

Surpassing human-level performance on simulated datasets:

- WSJ0-2mix anechoic separation
  - Wang, Zhong-Qiu, et al. "TF-GridNet: Making Time-Frequency Domain Models Great Again for Monaural Speaker Separation." *arXiv* 2022.

Table 1: Comparison with other systems on WSJ0-2mix.

Systems	Domain	Year	#params (M)	SI-SDRi (dB)	SDRi (dB)
DPCL++ [3]	T-F	2016	13.6	10.8	-
uPIT-BLSTM-ST [2]	T-F	2017	92.7	-	10.0
ADANet [27]	T-F	2018	9.1	10.4	10.8
WA-MISI-5 [5]	T-F	2018	32.9	12.6	13.1
Sign Prediction Net [6]	T-F	2019	56.6	15.3	15.6
Conv-TasNet [10]	Time	2019	5.1	15.3	15.6
Deep CASA [7]	T-F	2019	12.8	17.7	18.0
Conv-TasNet-MBT [11]	Time	2020	8.8	15.6	-
FurcaNeXt [12]	Time	2020	51.4	-	18.4
SUDO RM -RF [13]	Time	2020	2.6	18.9	-
DPRNN [14]	Time	2020	2.6	18.8	19.0
Gated DPRNN [15]	Time	2020	7.5	20.1	20.4
DPTNet [16]	Time	2020	2.7	20.2	20.6
DPTCN-ATPP [17]	Time	2021	4.7	19.6	19.9
SepFormer [18]	Time	2021	26.0	20.4	20.5
Sandglasset [19]	Time	2021	2.3	20.8	21.0
Wavesplit [20]	Time	2021	29.0	21.0	21.2
TFPSNet [24]	T-F	2022	2.7	21.1	21.3
MTDS (DPTNet) [21]	Time	2022	4.0	21.5	21.7
SFSRNet [22]	Time	2022	59.0	22.0	22.1
QDPN [23]	Time	2022	200.0	22.1	-
TF-GridNet	T-F	2022	14.4	**23.4**	**23.5**

Front-End Methods: SSE is getting stronger !

Surpassing human-level performance on simulated datasets:

- WSJ0-2mix anechoic separation
  - Wang, Zhong-Qiu, et al. "TF-GridNet: Making Time-Frequency Domain Models Great Again for Monaural Speaker Separation." *arXiv,* 2022.
- Multi-channel Noisy/Reverberant speech enhancement e.g. L3DAS22
  - Lu, Yen-Ju, et al. "Towards Low-Distortion Multi-Channel Speech Enhancement: The ESPNET-Se Submission to the L3DAS22 Challenge." ICASSP, 2022.

Table 1: Results of one-DNN systems on dev. set. Approaches marked with * use additional STOI loss and ASR-based Deep Feature loss.

Approaches	WER (%)	STOI	Task1 Metric
Challenge Baseline [9]	25.0	0.870	0.810
FasNet* [8]	18.2	0.874	0.846
Conv-TasNet [36] MVDR*	5.56	0.821	0.883
DCCRN* [33]	18.8	0.907	0.860
Demucs v2* [34]	26.3	0.851	0.794
Demucs v3* [38]	15.3	0.874	0.860
DNN_1	**3.90**	**0.964**	**0.963**

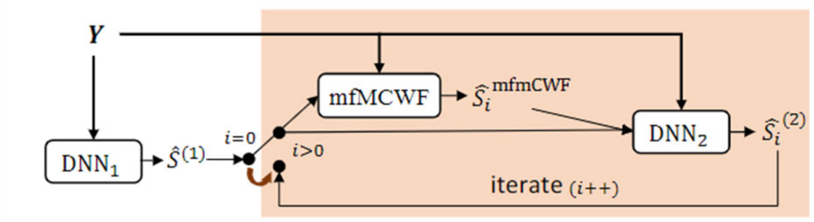


Fig. 1: Overview of proposed iterative neural/beamforming enhancement (iNeuBe) framework. A multi-frame multi-channel Wiener filter (mfMCWF) beamformer is applied between the two DNN MISO networks.

Front-End Methods: SSE is getting stronger !

Surpassing human-level performance on simulated datasets:

- WSJ0-2mix anechoic separation
  - Wang, Zhong-Qiu, et al. "TF-GridNet: Making Time-Frequency Domain Models Great Again for Monaural Speaker Separation." *arXiv, 2022*.
- Multi-channel Noisy/Reverberant speech enhancement e.g. L3DAS22
  - Lu, Yen-Ju, et al. "Towards Low-Distortion Multi-Channel Speech Enhancement: The ESPNET-Se Submission to the L3DAS22 Challenge." *ICASSP*, 2022.

Table 3: Results of two-DNN systems on dev. set.

Approaches	l	r	WER (%)	STOI	Task1 Metric
Challenge Baseline [9]	-	-	25.0	0.870	0.810
DNN_1	-	-	3.90	0.964	0.963
DNN_1+MVDR+DNN_2	-	-	3.62	0.970	0.968
DNN_1+mfMCWF+DNN_2	0	0	3.36	0.971	0.969
DNN_1+mfMCWF+DNN_2	7	0	2.63	0.978	0.976
DNN_1+mfMCWF+DNN_2	6	1	2.36	0.982	0.979
DNN_1+mfMCWF+DNN_2	5	2	2.53	0.982	0.978
DNN_1+mfMCWF+DNN_2	4	3	2.35	0.983	0.980
DNN_1+(mfMCWF+DNN_2)×2	4	3	**2.14**	**0.986**	**0.982**

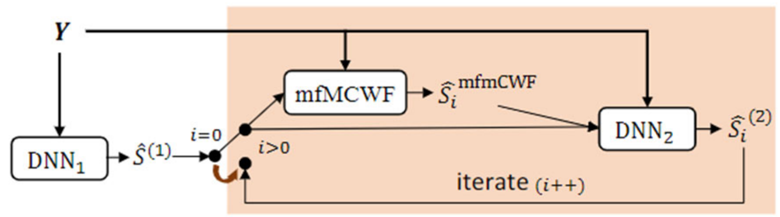
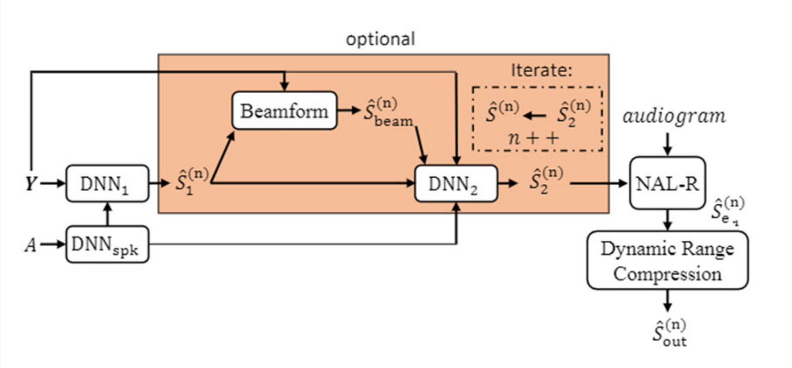


Fig. 1: Overview of proposed iterative neural/beamforming enhancement (iNeuBe) framework. A multi-frame multi-channel Wiener filter (mfMCWF) beamformer is applied between the two DNN MISO networks.

Front-End Methods: SSE is getting stronger !

Surpassing human-level performance on simulated datasets:

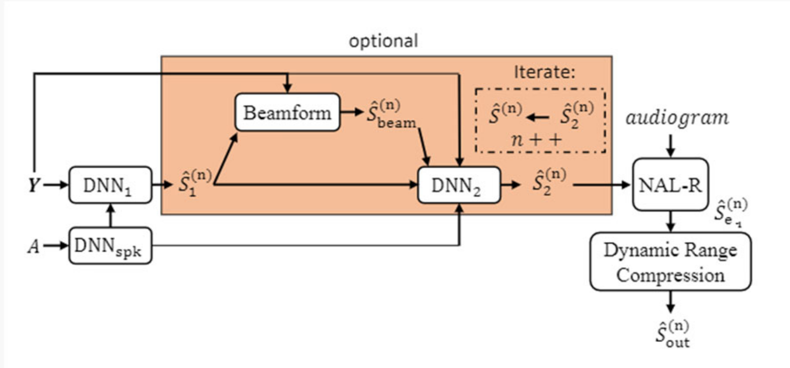
- WSJ0-2mix anechoic separation
  - Wang, Zhong-Qiu, et al. "TF-GridNet: Making Time-Frequency Domain Models Great Again for Monaural Speaker Separation." *arXiv* 2022.
- Multi-channel Noisy/Reverberant speech enhancement e.g. L3DAS22
  - Lu, Yen-Ju, et al. "Towards Low-Distortion Multi-Channel Speech Enhancement: The ESPNET-Se Submission to the L3DAS22 Challenge." *ICASSP 2022*.
- Multi-channel Noisy/Reverberant speech separation e.g. Clarity Challenge 2
  - iNeuBe+TF-GridNet+Target Speaker Extraction
  - Submitted to the Clarity Challenge 2 (Fingers crossed !)

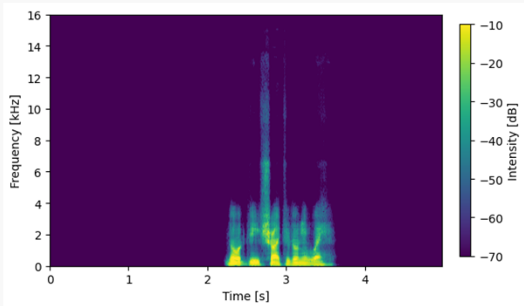


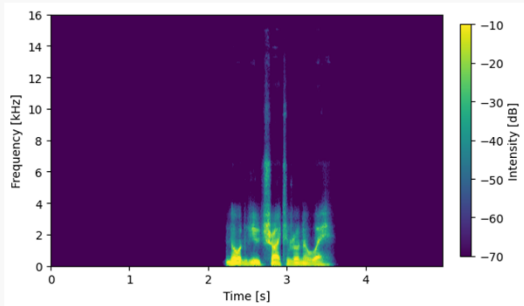
Front-End Methods: SSE is getting stronger !

What seems to work best:

- Strong DNN models
  - E.g. TCNDenseUNet, TF-GridNet
- Complex Spectral Mapping
  - Especially for noisy/reverberant scenarios
- Iterative Processing
  - Two iterations DNN1+DNN2 suffice



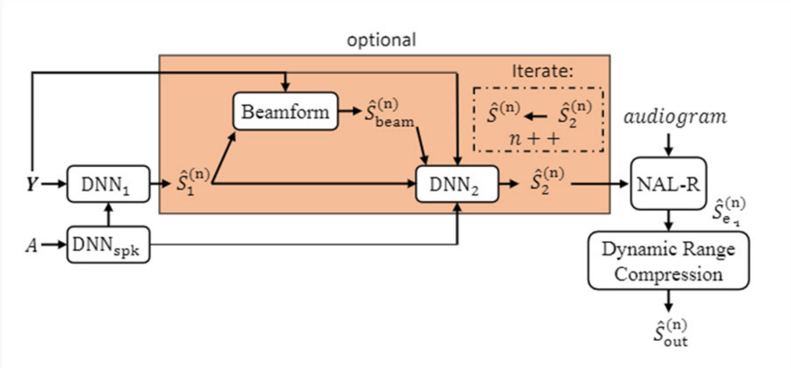




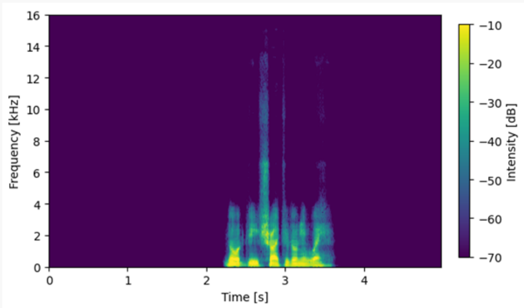
Front-End Methods: SSE is getting stronger !

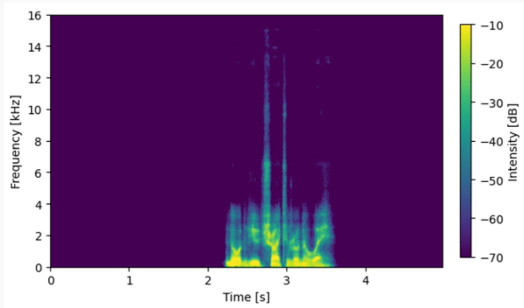
What seems to work best:

- Strong DNN models
  - E.g. TCNDenseUNet, TF-GridNet
- Complex Spectral Mapping
  - Especially for noisy/reverberant scenarios
- Iterative Processing
  - Two iterations DNN1+DNN2 suffice



Do these amazing techniques also work on CHiME-6 like scenarios ?





Front-end fine-tuning with the ASR back-end

No, usually they do not work on on-the-wild data without fine-tuning with the ASR back-end !

Main plague of SSE:

- Mismatch between training (synthetic dataset) and testing conditions (real-data).
- With no fine-tuning ASR performance can degrade.

Front-end fine-tuning with the ASR back-end

No, usually they do not work on on-the-wild data without fine-tuning with the ASR back-end !

Main plague of SSE:

- Mismatch between training (synthetic dataset) and testing conditions (real-data).
- With no fine-tuning ASR performance can degrade.

Case study 1: IRIS

- Chang, Xuankai, et al. "End-to-End Integration of Speech Recognition, Speech Enhancement, and Self-Supervised Learning Representation." *arXiv preprint arXiv:2204.00540* (2022).

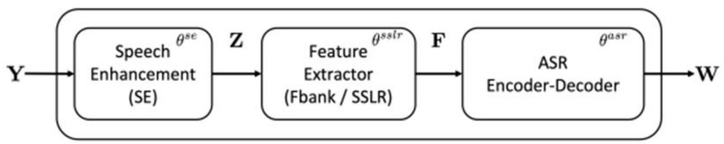


Figure 1: *Overview of the proposed end-to-end model.*

Table 1: *Single-channel CHiME-4 ASR performance (%WER) of the E2E-ASR model and previous studies on monaural dev and test sets. In system 6 and 7, HuBERT and WavLM models are pre-trained with large amount of unlabelled data.*

ID	System	Model	Dev. Set		Test Set	
			Simu.	Real	Simu.	Real
1	Kaldi Baseline [33]	Hybrid	6.81	5.58	12.15	11.42
2	Du *et al.* [34]	Hybrid	6.61	4.55	11.81	9.15
3	Yang *et al.* [7]	Hybrid	**4.99**	**3.35**	8.61	6.25
4	Wav2Vec-Switch [25]	E2E	-	3.5	-	6.6
5	E2E Transformer - Fbank	E2E	11.32	9.43	19.67	17.99
6	E2E Transformer - HuBERT	E2E	11.56	9.13	18.02	20.41
7	E2E Transformer - WavLM	E2E	5.93	4.03	**8.25**	**4.47**

Table 2: *Monaural CHiME-4 ASR performance (%WER) of the IRIS model. Different combinations of fine-tuning SE (FT. SE) and fine-tuning ASR (FT. ASR) are evaluted.*

Enhancement	Feature	FT. SE	FT. ASR	Dev. Set		Test Set	
				Simu.	Real	Simu.	Real
Conv-TasNet	Fbank	✗	✗	17.22	16.76	30.28	32.50
	Fbank	✗	✓	11.42	9.92	21.16	21.82
	Fbank	✓	✗	9.20	8.33	17.01	16.56
	Fbank	✓	✓	9.52	7.94	17.42	15.24
	WavLM	✗	✗	5.96	4.37	13.52	12.11
	WavLM	✗	✓	5.45	4.04	12.68	11.57
	WavLM	✓	✗	3.54	2.27	6.73	4.90
	WavLM	✓	✓	**3.16**	**2.03**	**6.12**	**3.92**

Front-end fine-tuning with the ASR back-end

No, usually they do not work on on-the-wild data without fine-tuning with the ASR back-end !

Main plague of SSE:

- Mismatch between training (synthetic dataset) and testing conditions (real-data)
- With no fine-tuning ASR performance can degrade

Case study 1: IRIS

- Chang, Xuankai, et al. "End-to-End Integration of Speech Recognition, Speech Enhancement, and Self-Supervised Learning Representation." *arXiv preprint arXiv:2204.00540* (2022).

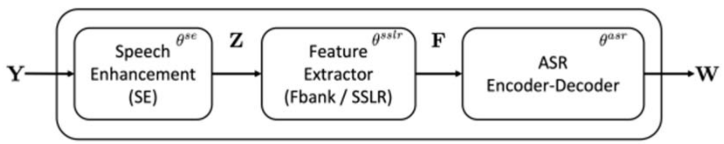


Figure 1: *Overview of the proposed end-to-end model.*

Table 1: *Single-channel CHiME-4 ASR performance (%WER) of the E2E-ASR model and previous studies on monaural dev and test sets. In system 6 and 7, HuBERT and WavLM models are pre-trained with large amount of unlabelled data.*

ID	System	Model	Dev. Set		Test Set	
			Simu.	Real	Simu.	Real
1	Kaldi Baseline [33]	Hybrid	6.81	5.58	12.15	11.42
2	Du *et al.* [34]	Hybrid	6.61	4.55	11.81	9.15
3	Yang *et al.* [7]	Hybrid	**4.99**	**3.35**	8.61	6.25
4	Wav2Vec-Switch [25]	E2E	-	3.5	-	6.6
5	E2E Transformer - Fbank	E2E	11.32	9.43	19.67	17.99
6	E2E Transformer - HuBERT	E2E	11.56	9.13	18.02	20.41
7	E2E Transformer - WavLM	E2E	5.93	4.03	**8.25**	**4.47**

Table 2: *Monaural CHiME-4 ASR performance (%WER) of the IRIS model. Different combinations of fine-tuning SE (FT. SE) and fine-tuning ASR (FT. ASR) are evaluted.*

Enhancement	Feature	FT. SE	FT. ASR	Dev. Set		Test Set	
				Simu.	Real	Simu.	Real
Conv-TasNet	Fbank	✗	✗	17.22	16.76	30.28	32.50
	Fbank	✗	✓	11.42	9.92	21.16	21.82
	Fbank	✓	✗	9.20	8.33	17.01	16.56
	Fbank	✓	✓	9.52	7.94	17.42	15.24
	WavLM	✗	✗	5.96	4.37	13.52	12.11
	WavLM	✗	✓	5.45	4.04	12.68	11.57
	WavLM	✓	✗	3.54	2.27	6.73	4.90
	WavLM	✓	✓	**3.16**	**2.03**	**6.12**	**3.92**

Front-end fine-tuning with the ASR back-end

Case study 2: multi-IRIS

- Masuyama, Yoshiki, et al. "End-to-End Integration of Speech Recognition, Dereverberation, Beamforming, and Self-Supervised Learning Representation." *arXiv preprint arXiv:2210.10742* (2022).
- Demo page:

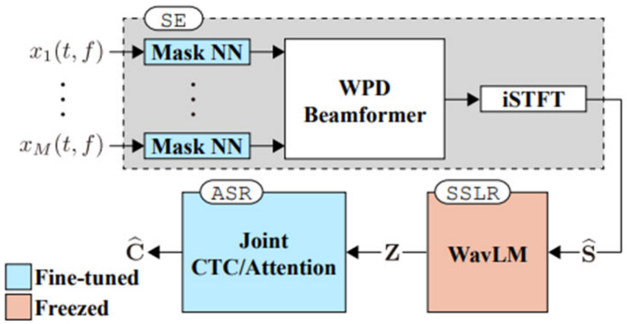


Table 3: WER with different beamformers on CHiME-4 dataset. WavLM was used for feature extraction in all systems.

		Dev. Set		Test Set		Ave.
		Simu.	Real	Simu.	Real	
2ch.	BeamformIt	4.17	5.33	5.58	4.57	4.89
	MPDR	2.53	2.03	2.26	2.98	2.43
	+ Joint training	2.45	1.93	2.19	2.89	2.35
	MVDR	2.38	2.13	2.11	3.14	2.41
	+ Joint training	2.30	1.98	**2.04**	2.86	2.28
	WPD	2.28	2.06	2.30	3.63	2.52
	+ Joint training	**2.04**	**1.66**	**2.04**	**2.65**	**2.07**
6ch.	BeamformIt	2.78	4.28	3.80	3.57	3.60
	MPDR	1.36	1.44	1.39	1.84	1.49
	+ Joint training	1.36	1.42	1.36	1.79	1.47
	MVDR	1.21	1.38	1.23	1.91	1.41
	+ Joint training	1.25	**1.31**	**1.21**	1.85	1.39
	WPD	**1.19**	1.32	1.29	1.85	1.39
	+ Joint training	1.22	1.33	1.24	**1.77**	**1.38**

Front-end fine-tuning with the ASR back-end

Takeaways:

- Thou shall fine-tune ! (especially if monaural)
  - NOTE that retraining/fine-tuning the ASR may not be possible in many applications however ! Preferable to only tune the front-end even if sub-optimal.
- Fine-tuning/Retraining may be not necessary when using distortion-less beamforming (e.g. WPD, MVDR)
  - Why ? Because beamformed signals are "natural" (linear combination of input signals) see [1].
  - But the scenario considered is arguably very simple.
  - Nonetheless e.g. VarArray [2] results are very encouraging on AMI and show fine-tuning works quite good.
    - Again, VarArray uses MVDR.

[1] *Iwamoto, Kazuma, et al. "How bad are artifacts?: Analyzing the impact of speech enhancement errors on asr." arXiv preprint arXiv:2201.06685 (2022).*

[2] *Yoshioka, Takuya, et al. "VarArray: Array-geometry-agnostic continuous speech separation." ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022.*

Front-end fine-tuning with the ASR back-end

Under-explored direction to tackle mismatch:

- Using unsupervised techniques such as MixIT [1] to adapt the SSE model to real-world mixtures.
  - Preliminary work: *Sivaraman, Aswin, et al. "Adapting speech separation to real-world meetings using mixture invariant training." ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022.*
    - No back-end ASR evaluation (or diarization) however, only signal-based metrics and subjective listening tests.

[1] Wisdom, Scott, et al. "Unsupervised sound separation using mixture invariant training." *Advances in Neural Information Processing Systems* 33 (2020): 3846-3857.

Back-End Methods: PIT-based

Permutation Invariant ASR methods:

- MIMO-Speech
  - Chang, Xuankai, et al. "MIMO-Speech: End-to-end multi-channel multi-speaker speech recognition." *2019 IEEE Automatic ASRU*, 2019.
- DASR: Directional ASR
  - Subramanian, Aswin Shanmugam, et al. "Directional ASR: A new paradigm for E2E multi-speaker speech recognition with source localization." ICASS, 2021.

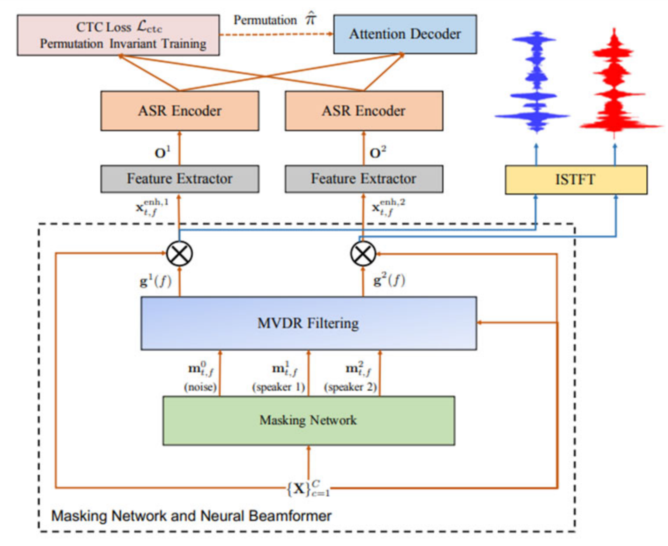
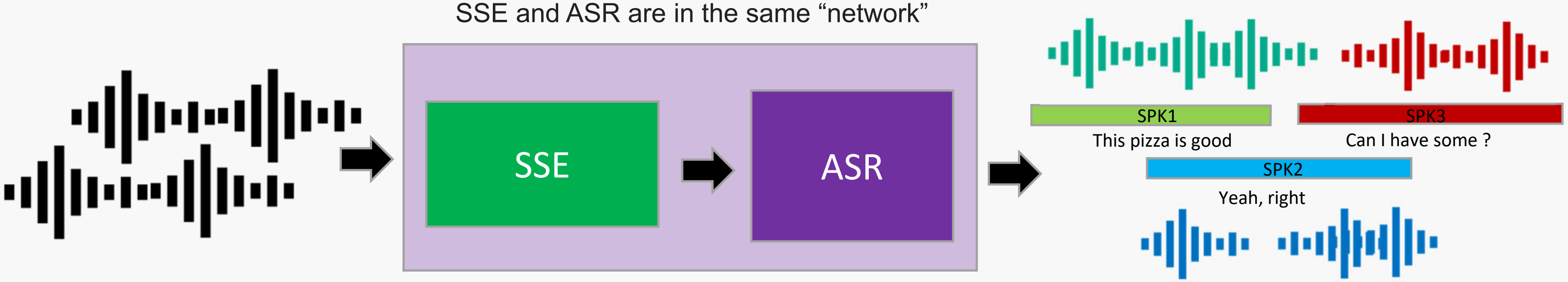


Fig. 1. End-to-End Multi-channel Multi-speaker Model

Image from MIMO Speech paper

SSE and ASR are in the same “network”



Back-End Methods: PIT-based

Permutation Invariant ASR methods:

- MIMO-Speech
  - Chang, Xuankai, et al. "MIMO-Speech: End-to-end multi-channel multi-speaker speech recognition." *2019 IEEE Automatic ASRU*, 2019.
- DASR: Directional ASR
  - Subramanian, Aswin Shanmugam, et al. "Directional ASR: A new paradigm for E2E multi-speaker speech recognition with source localization." ICASS, 2021.
- Might also be non-interpretable

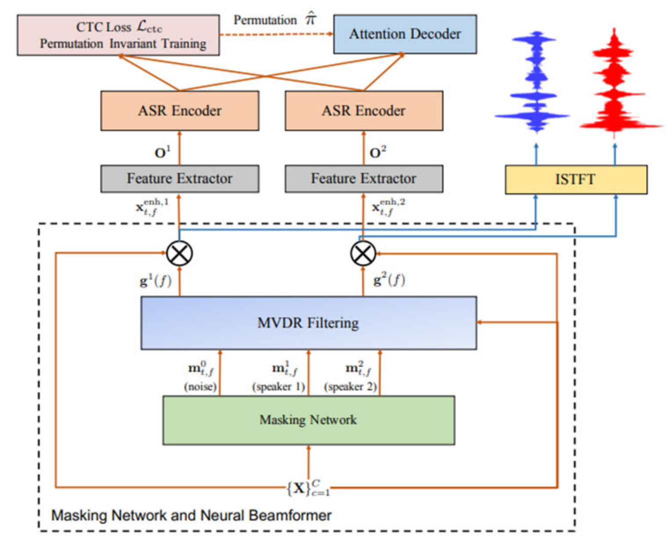
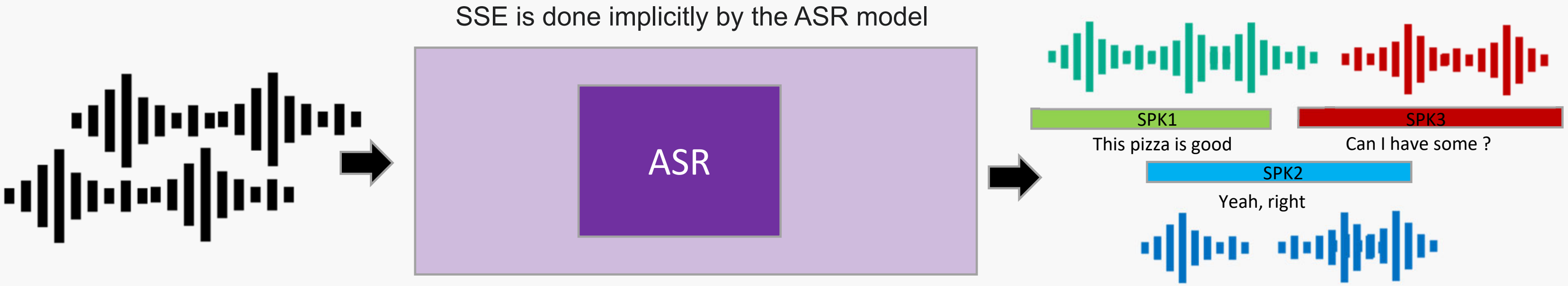


Fig. 1. End-to-End Multi-channel Multi-speaker Model

Image from MIMO Speech paper



Back-End Methods: PIT-based

Practically, MIMO-Speech and DASR are equivalent to SSE+ASR pipeline with fine-tuning, e.g. Multi-IRIS.

- But are engineered to be trained from scratch with the ASR objective
  - No synthetic to real-world domain mismatch problem.
  - Convergence may be an issue however on challenging datasets.
    - E.g. MIMO-Speech uses curriculum learning

Back-End Methods: PIT-based

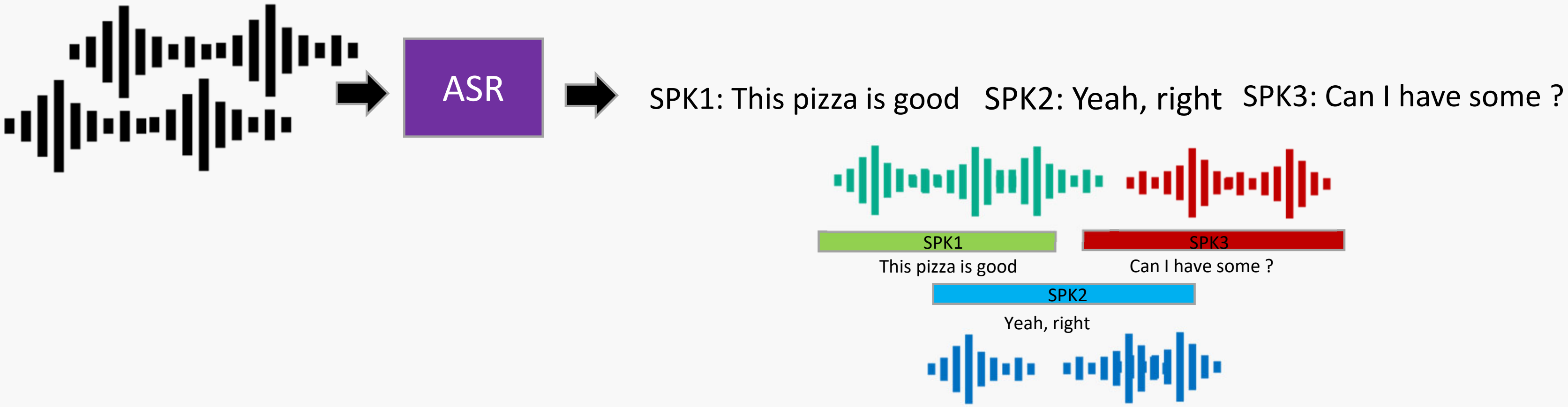
Practically, MIMO-Speech and DASR are equivalent to SSE+ASR pipeline with fine-tuning, e.g. Multi-IRIS.

- But are engineered to be trained from scratch with the ASR objective
  - No synthetic to real-world domain mismatch problem.
  - Convergence may be an issue however on challenging datasets.
    - E.g. MIMO-Speech uses curriculum learning

- Same Pros/Cons of CSS:
  - They may require to perform CSS on long inputs or diarization otherwise we lose speaker tracking !
  - May be difficult to train on datasets such as CHiME-6 (large memory requirements for whole utterance groups)
  - Not easy to generalize to arbitrary large number of speakers

Back-End Methods: Serialized Output Training

Similar to PIT-based methods but trained to output the speakers transcripts in a FIFO way. Examples [1], [2], [3]



[1] N. Kanda, Y. Gaur et al., “Serialized output training for endto-end overlapped speech recognition,” in Proc. Interspeech, 2020

[2] N. Kanda, J. Wu et al., “Streaming multi-talker ASR with token-level serialized output training,” in Proc. Interspeech, 2022

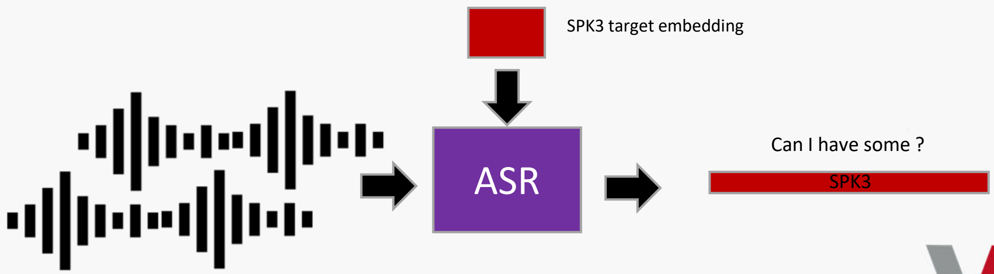
[3] Kanda, Naoyuki, et al. "Transcribe-to-Diarize: Neural Speaker Diarization for Unlimited Number of Speakers using End-to-End Speaker-Attributed ASR." ICASSP, 2022.

Back-End Methods: target-speaker-based

Practically, also these back-end techniques are equivalent to target speaker extraction:

- Implicit extraction: the ASR model ignores competing speakers and transcribes only the target.
  - Can be interpretable, “target DASR”:
    - *Subramanian, Aswin Shanmugam, et al. "Far-field location guided target speech extraction using end-to-end speech recognition objectives." ICASSP 2020.*
  - Also non interpretable:
    - *Huang, Zili, et al. "Adapting self-supervised models to multi-talker speech recognition using speaker embeddings." arXiv preprint arXiv:2211.00482 (2022).*

- Same pros/cons of target speaker extraction minus mismatch problem.
  - Performance largely depends on accurate diarization



WER we are going: Current Trends

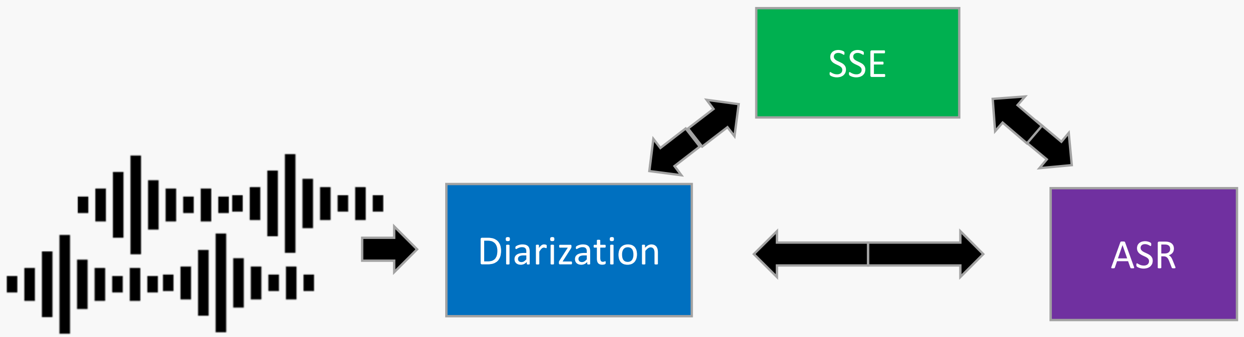
End-to-End/Tight integration of front-end and back-end

- SSE+ASR:
  - Back-end methods: MIMO-Speech, DASR, “target-DASR”
  - Front-end methods: IRIS, multi-IRIS, VarArray etc.
  - SSE helps ASR but also vice-versa is true
    - E.g. Erdogan, Hakan, et al. "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks." ICASSP. IEEE, 2015.
- SSE+Diarization:
  - Speech Separation Guided Diarization (SSGD)
  - EEND-SS
    - Ueda, Yushi, et al. "EEND-SS: Joint End-to-End Neural Speaker Diarization and Speech Separation for Flexible Number of Speakers." *arxiv* 2022.
  - More work will come here since both speech separation and EEND use PIT
- Diarization+ASR:
  - Kanda, Naoyuki, et al. "Transcribe-to-Diarize: Neural Speaker Diarization for Unlimited Number of Speakers using End-to-End Speaker-Attributed ASR." ICASSP, 2022.
  - Khare, Aparna, et al. "ASR-aware end-to-end neural diarization." ICASSP, 2022.

Separate but Together !

Diarization, ASR and separation are intimately related

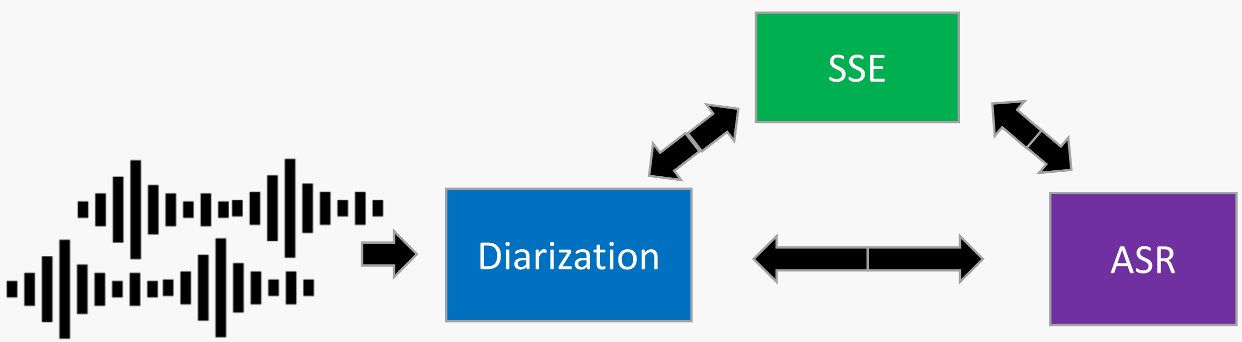
- Can we devise a way on how to integrate all of these ?
  - Ravanelli, Mirco, et al. "A network of deep neural networks for distant speech recognition." *ICASSP*, 2017.



Separate but Together !

Diarization, ASR and separation are intimately related

- Can we devise a way on how to integrate all of these ?
  - Ravanelli, Mirco, et al. "A network of deep neural networks for distant speech recognition." *ICASSP*, 2017.



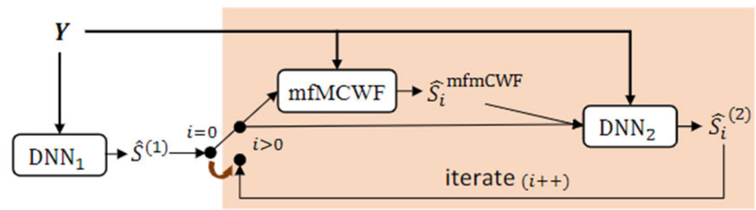


Fig. 1: Overview of proposed iterative neural/beamforming enhancement (iNeuBe) framework. A multi-frame multi-channel Wiener filter (mfMCWF) beamformer is applied between the two DNN MISO networks.

Iterative processing, like in iNeuBe or Target-Speaker VAD [1].

[1] Medennikov, I., et al. "Target-speaker voice activity detection: A novel approach for multi-speaker diarization in a dinner party scenario." *INTERSPEECH*. 2020.

WER we are going: Current Trends

Pretrained models, leveraging massive datasets:

- Self-supervised learning representation:
  - IRIS, multi-IRIS
  - Multi-talker adaptation of pre-trained SSL models:
    - *Huang, Zili, et al. "Adapting self-supervised models to multi-talker speech recognition using speaker embeddings." arXiv preprint arXiv:2211.00482 (2022).*
  - Large supervised ASR models such as Whisper

- How to adapt these to multi-channel scenarios ?
  - Open question, simple selection already can show how much powerful these models are

	CHiME-6 eval WER (oracle diarization)
BigSSL [1] (GSS)	31.0%
Whisper (reference array)	56.63% (49.09% with fine tuning)
Whisper (oracle selection)	**27.91%** (**19.80%** with fine tuning)
Whisper (MicRank selection)	33.87% (26.40% with fine tuning)

Thank you for your time

Any questions ?