Enabling Speech Technologies "In the Wild"

Samuele Cornell,
Language Technologies Institute,
Carnegie Mellon University

Watanabe's
Audio and Voice Lab

Language
Technologies
Institute

Table of Contents

What's "Speech-in-the-Wild"?

What is “Speech-in-the-Wild” ?

Addressing the social aspect: «everyday speech» / «speech-in-the-wild»

What is “Speech-in-the-Wild” ?

Addressing the social aspect: «everyday speech» / «speech-in-the-wild»

- **2 or more people involved**: «speech-in-the-wild» / conversational speech

What is “Speech-in-the-Wild” ?

Addressing the social aspect: «everyday speech» / «speech-in-the-wild»

- **2 or more people involved**: «speech-in-the-wild» / conversational speech
  - Colloquial terms, dialect, overlapped speech

What is “Speech-in-the-Wild” ?

Addressing the social aspect: «everyday speech» / «speech-in-the-wild»

- **2 or more people involved**: «speech-in-the-wild» / conversational speech
  - Colloquial terms, dialect, overlapped speech
- Spontaneous interaction -> **long-form audio**
  - Not pre-segmented (e.g. via keyword spotting «Hey Siri !»)

What is “Speech-in-the-Wild” ?

Addressing the social aspect: «everyday speech» / «speech-in-the-wild»

- **2 or more people involved**: «speech-in-the-wild» / conversational speech
  - Colloquial terms, dialect, overlapped speech
- Spontaneous interaction -> **long-form audio**
  - Not pre-segmented (e.g. via keyword spotting «Hey Siri !»)

What is “Speech-in-the-Wild” ?

Addressing the social aspect: «everyday speech» / «speech-in-the-wild»

- **2 or more people involved**: «speech-in-the-wild» / conversational speech
  - Colloquial terms, dialect, overlapped speech
- Spontaneous interaction -> **long-form audio**
  - Not pre-segmented (e.g. via keyword spotting «Hey Siri !»)
- Usually **captured by far-field microphone devices** (laptop, smart-speaker, robots, smart-glasses)
  - Noise, reverberation, overlapped speech (again !)

What is “Speech-in-the-Wild” ?

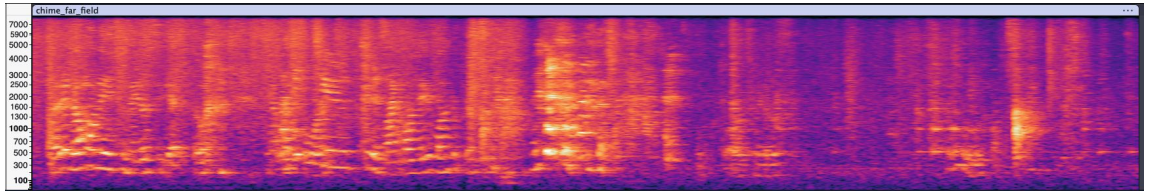
E.g. CHiME-6 dataset [1] («dinner-party» scenario between 4 friends)

[1] Watanabe, Shinji, et al. "CHiME-6 challenge: Tackling multi-speaker speech recognition for unsegmented recordings.” CHiME Workshop. 2020

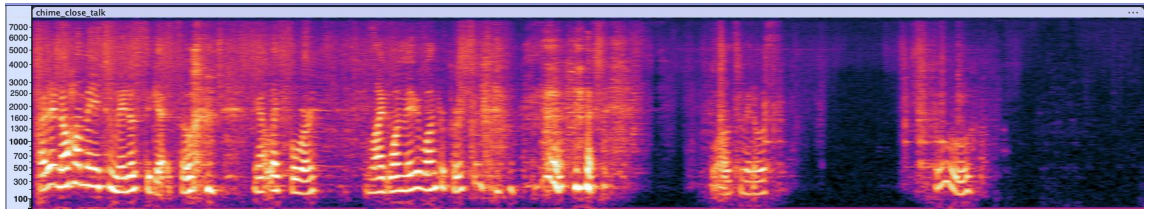
What is “Speech-in-the-Wild” ?

E.g. CHiME-6 dataset [1] («dinner-party» scenario between 4 friends)

Far-field microphone



«cleaned» close-talk on-person microphone («pseudo-oracle»)

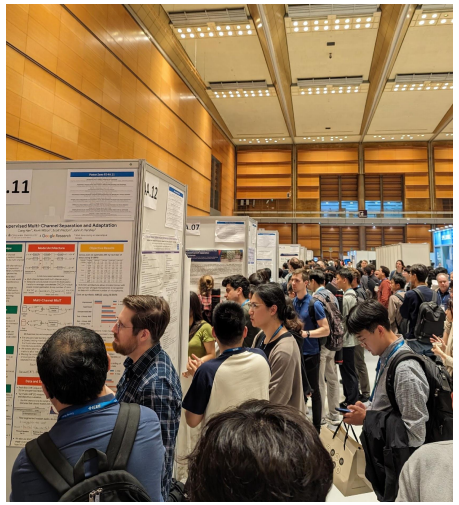


[1] Watanabe, Shinji, et al. "CHiME-6 challenge: Tackling multi-speaker speech recognition for unsegmented recordings.” CHiME Workshop. 2020

Speech Technologies “in-the-Wild”

Our ability in following one voice in challenging acoustic conditions is still unmatched by machines

- «Cocktail Party problem [1]»

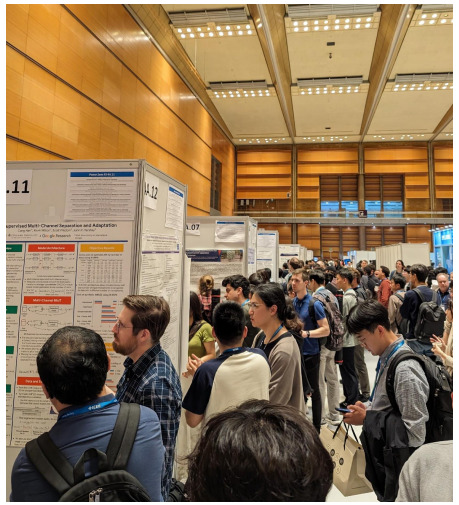


[1] Cherry EC . ”Some experiments on the recognition of speech with one and two ears” The Journal of the Acoustical Society of America. 1953

Speech Technologies “in-the-Wild”

Our ability in following one voice in challenging acoustic conditions is still unmatched by machines

- «Cocktail Party problem [1]»



Hearing-aid speech enhancement in a restaurant

Real-time translation in a busy street

[1] Cherry EC . ”Some experiments on the recognition of speech with one and two ears” The Journal of the Acoustical Society of America. 1953

Conversational speech is a fundamental problem

100,000+ years ago

Human language evolved for multi-party **spoken** interaction

→

~5,500 years ago

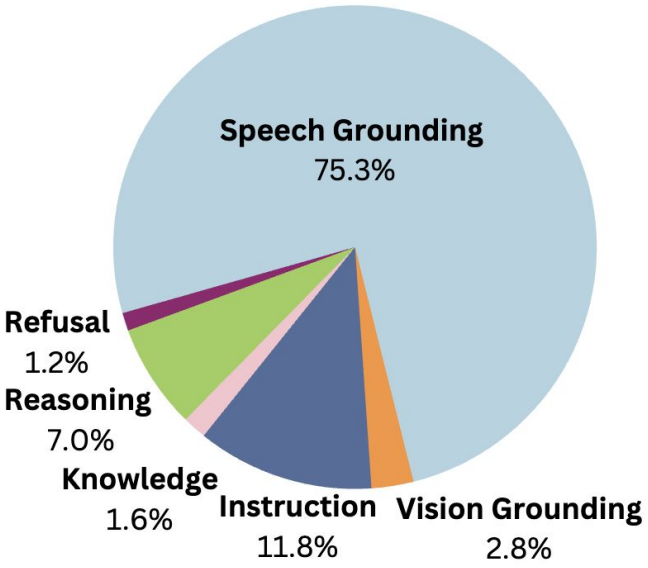
Writing was invented

→

Now

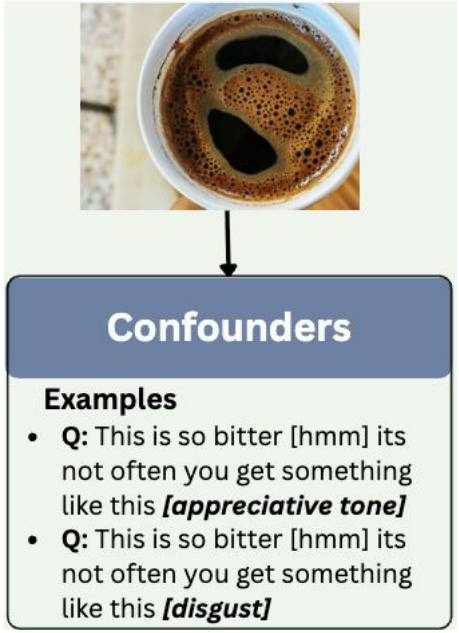
"AI" (LLMs) are mostly text-centric: speech not a "first-class citizen"

Current “AI” is still mostly “text-centric” (despite efforts)



Distribution of error types in Gemini 2.5 Pro responses for multimodal (video/image+audio) prompts

- Most errors stem from failure in speech grounding



From Selvakumar, Ramaneswaran, et al. "MULTIVOX: A Benchmark for Evaluating Voice Assistants for Multimodal Interactions.” EMNLP. 2025.

Beyond Dyadic Human-Machine Interaction (HMI)

Let’s move beyond dyadic Human-Machine Interaction (HMI) !

Dyadic HMI has its place, it is an important technology

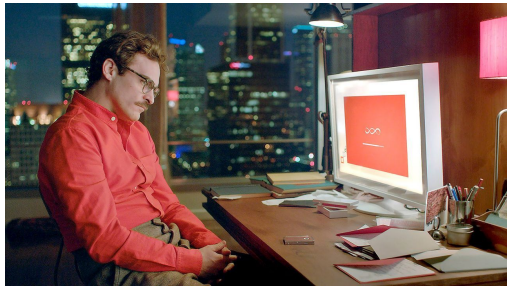
- Natural extension of LLMs chatbots
- Many applications e.g. assistive technology

Beyond Dyadic Human-Machine Interaction (HMI)

Let’s move beyond dyadic Human-Machine Interaction (HMI) !

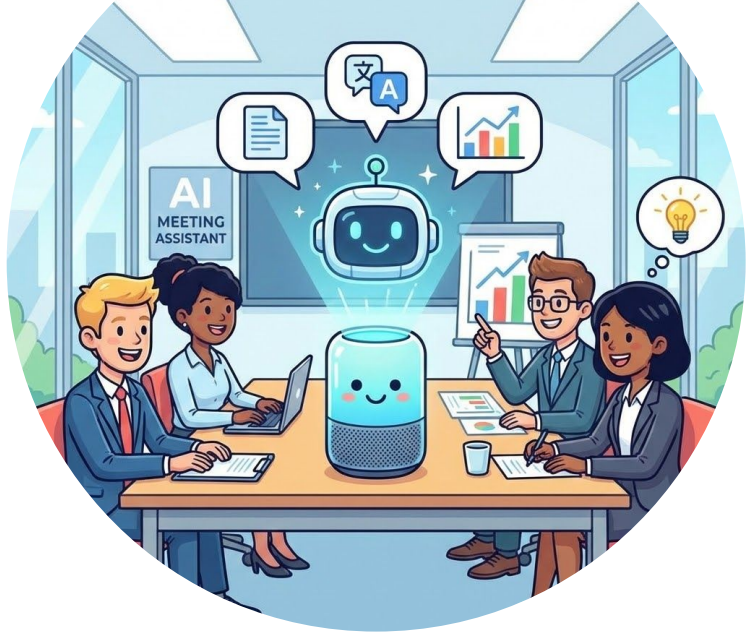
Dyadic HMI has its place, it is an important technology

- Natural extension of LLMs chatbots
- Many applications e.g. assistive technology



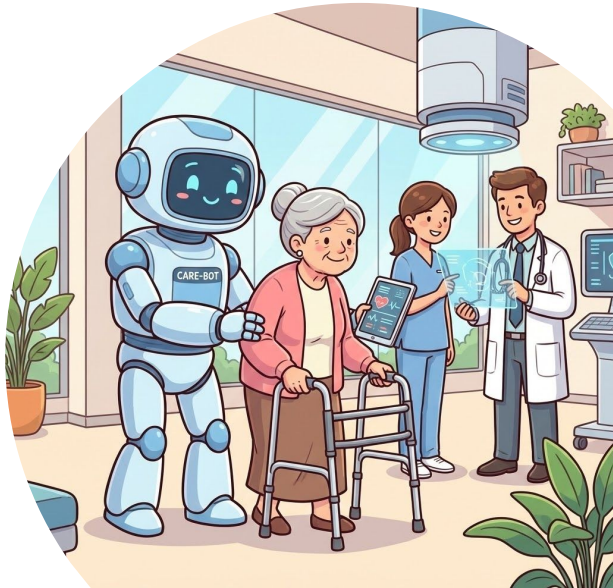
From the movie “Her”

Beyond Dyadic HMI

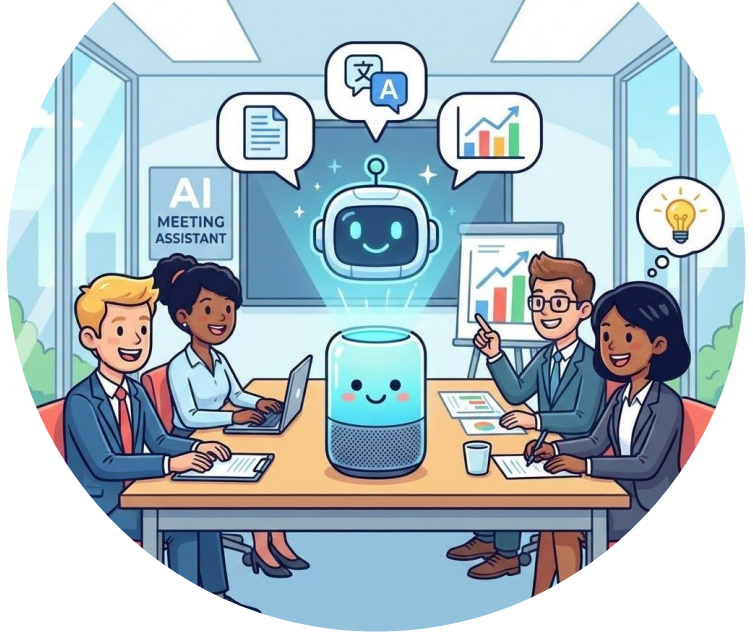


“Conversational AI” should be able to handle collaboration/conversations with multiple human beings.

- integrate into our interactions/work environments in a natural way and **handle our “messy” human to human conversations**.



Beyond Dyadic HMI

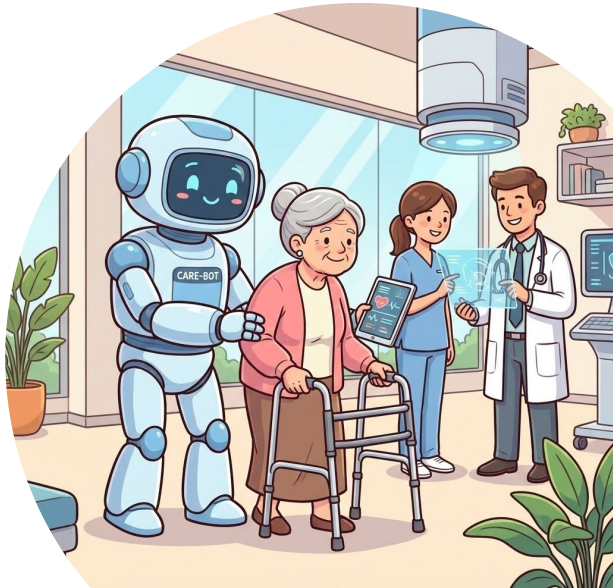


“Conversational AI” should be able to handle collaboration/conversations with multiple human beings.

- integrate into our interactions/work environments in a natural way and **handle our “messy” human to human conversations**.

Industry right-now is mostly focused on dyadic interaction

- **Great opportunity for academia** to lead and pave the way in this direction.



WER we are ? Current State-of-the-Art

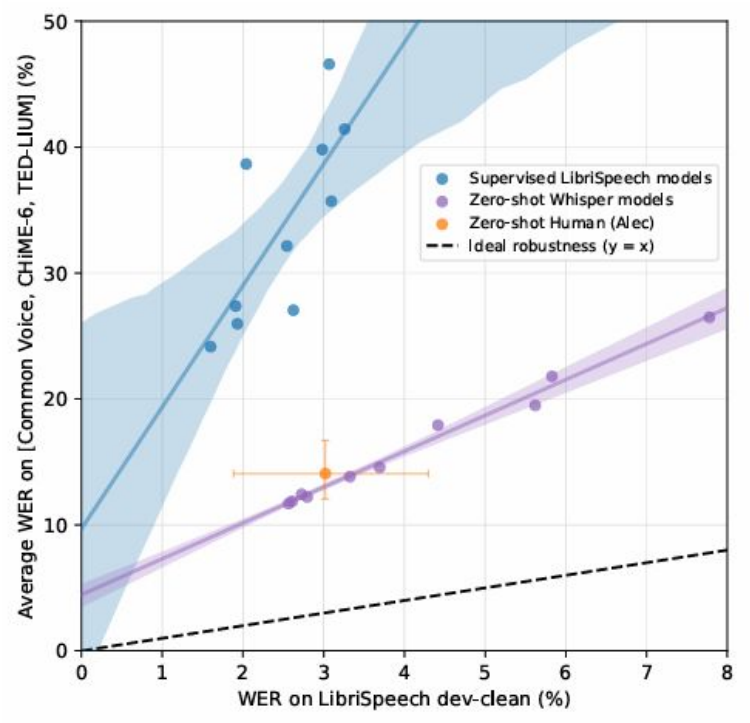


Figure 2. **Zero-shot Whisper models close the gap to human robustness.** Despite matching or outperforming a human on LibriSpeech dev-clean, supervised LibriSpeech models make roughly twice as many errors as a human on other datasets demonstrating their brittleness and lack of robustness. The estimated robustness frontier of zero-shot Whisper models, however, includes the 95% confidence interval for this particular human.

OpenAI Whisper ASR model (trained on 5M hours of weakly-labeled, probably mostly Youtube scraped data)

Word error rate (WER) **competitive or better than human** on many domains**:**

- LibriSpeech (audiobooks)
- CommonVoice (read speech)
- TEDLIUM (TED Talks)

From Radford, Alec, et al. "Robust speech recognition via large-scale weak supervision." *International conference on machine learning*. PMLR, 2023.

WER we are ? Current State-of-the-Art

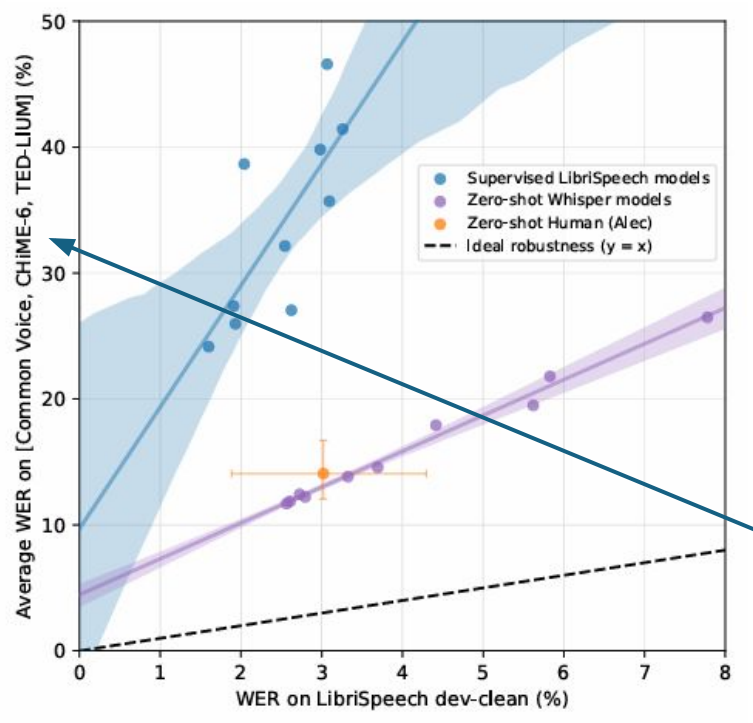


Figure 2. **Zero-shot Whisper models close the gap to human robustness.** Despite matching or outperforming a human on LibriSpeech dev-clean, supervised LibriSpeech models make roughly twice as many errors as a human on other datasets demonstrating their brittleness and lack of robustness. The estimated robustness frontier of zero-shot Whisper models, however, includes the 95% confidence interval for this particular human.

OpenAI Whisper ASR model (trained on 5M hours of weakly-labeled, probably mostly Youtube scraped data)

Word error rate (WER) **competitive or better than human** on many domains**:**

- LibriSpeech (audiobooks)
- CommonVoice (read speech)
- TEDLIUM (TED Talks)
- CHiME-6 (**“dinner-party” meeting data**)

From Radford, Alec, et al. "Robust speech recognition via large-scale weak supervision." *International conference on machine learning*. PMLR, 2023.

Thank you for this opportunity !
Any questions ?

Email: scornell@andrew.cmu.edu

“The Devil is in the Details”

Word error rate (WER) **competitive or better than human** on many domains**:**

- LibriSpeech (audiobooks)
- CommonVoice (read speech)
- TEDLIUM (TED Talks)
- CHiME-6 (**“dinner-party” meeting data**)**

****oracle segmentation & diarization**, with speech separation and beamforming

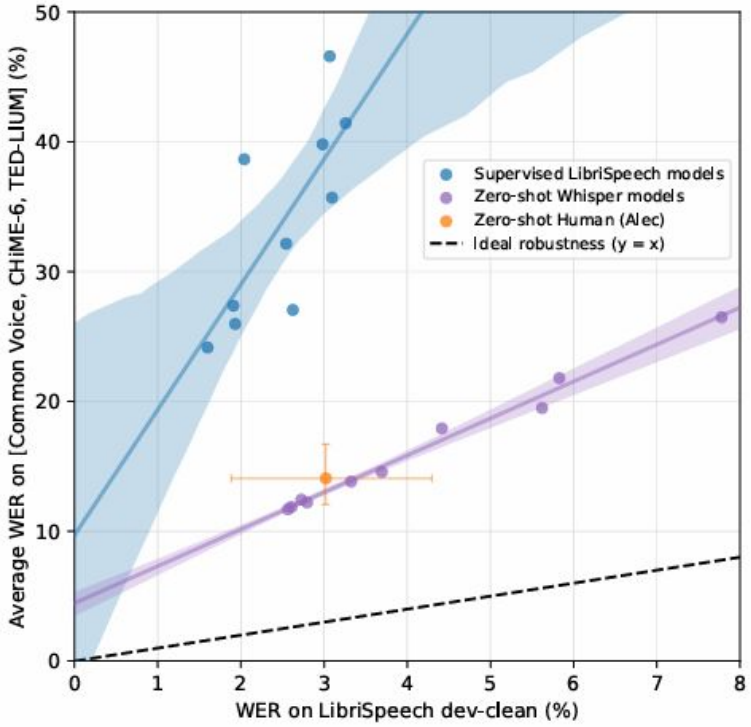
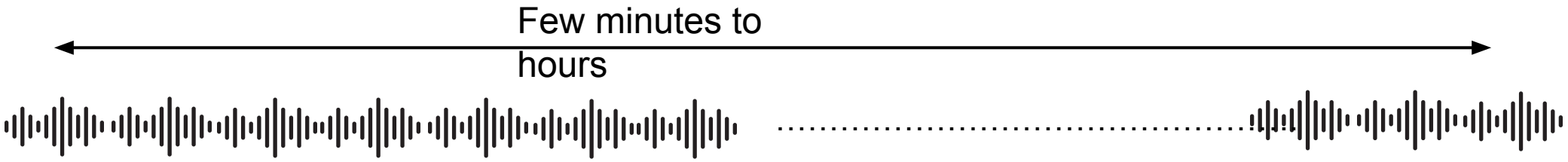


Figure 2. **Zero-shot Whisper models close the gap to human robustness.** Despite matching or outperforming a human on LibriSpeech dev-clean, supervised LibriSpeech models make roughly twice as many errors as a human on other datasets demonstrating their brittleness and lack of robustness. The estimated robustness frontier of zero-shot Whisper models, however, includes the 95% confidence interval for this particular human.

“The DER (diarization error rate) is in the Details”



<speaker_0> <0.2 s> Hi John !<1.3 s>

<speaker_1> <1.2 s> Hey ! What 'up ?

<2.1 s> ….

<speaker_0> <1200 s> Bye! 1201.2 s>

Long-term information tracking

“The DER (diarization error rate) is in the Details”

Again, AMI [1] meeting data (not particularly noisy or reverberant and also 3 to 5 participants, office meeting)
Model: Gemini 2.5 Pro (raw audio input, prompted to do diarization):

Duration (minutes)	Diarization Error Rate (DER) %
10	79.5
Full-length	98.3 (**high failure rate though)

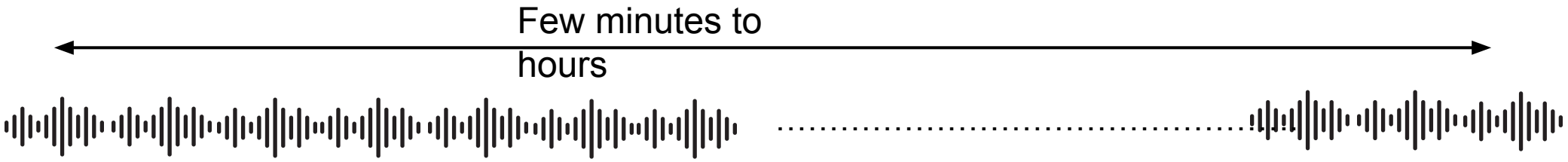
For reference: **Open-source SotA on full-length recordings (1 hour) is ~ 10% DER**

- DiariZen [2]: https://github.com/BUTSpeechFIT/DiariZen

[1] Carletta, Jean, et al. "The AMI meeting corpus: A pre-announcement." International workshop on machine learning for multimodal interaction. 2005.

[2] Han, Jiangyu, et al. "Leveraging self-supervised learning for speaker diarization." ICASSP. 2025

“The DER (diarization error rate) is in the Details”



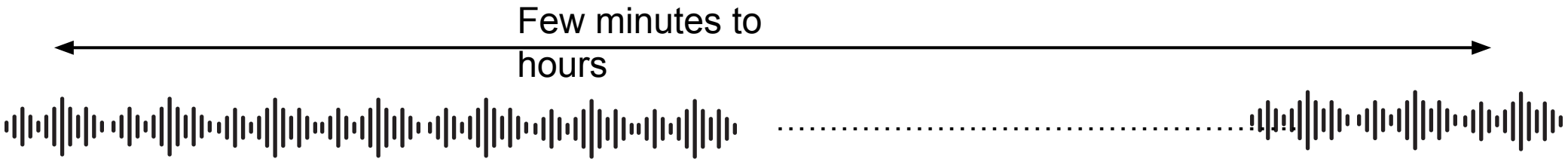
<speaker_0> <0.2 s> Hi John !<1.3 s>
<speaker_1> <1.2 s> Hey ! What 'up ?
<2.1 s> ….

<speaker_0> <1200 s> Bye! 1201.2 s>

For Gemini 2.5 Pro:

- 1 hour -> 110k tokens

“The DER (diarization error rate) is in the Details”



<speaker_0> <0.2 s> Hi John !<1.3 s>
<speaker_1> <1.2 s> Hey ! What 'up ?
<2.1 s> ….

<speaker_0> <1200 s> Bye! 1201.2 s>

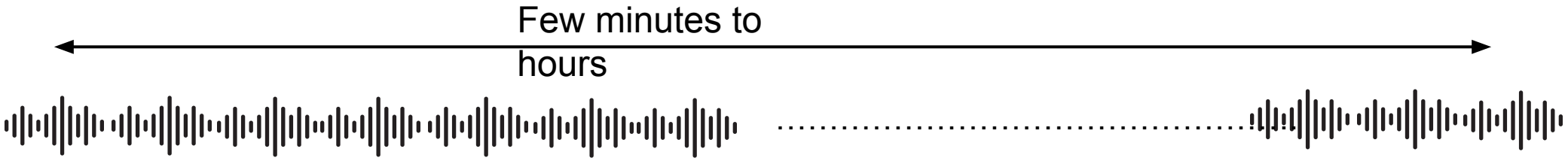
For Gemini 2.5 Pro:

- 1 hour -> 110k tokens

Challenges:

- Small amount of data on most domains
  - (~ thousands of hours)

“The DER (diarization error rate) is in the Details”



<speaker_0> <0.2 s> Hi John !<1.3 s>
<speaker_1> <1.2 s> Hey ! What 'up ?
<2.1 s> ….

<speaker_0> <1200 s> Bye! 1201.2 s>

For Gemini 2.5 Pro:

- 1 hour -> 110k tokens

Challenges:

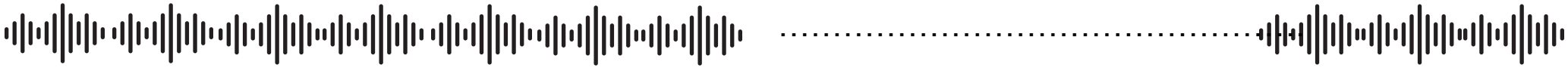
- Small amount of data on most domains
  - (~ thousands of hours)
- Autoregressive model -> more prone to error accumulation
  - Timestamps often ends up out of audio boundaries for Gemini 2.5 Pro

“The DER (diarization error rate) is in the Details”

Few minutes to hours

<speaker_0> <0.2 s> Hi John !<1.3 s>

<speaker_0> <1200 s> Bye! 1201.2 s>



<speaker_1> <1.2 s> Hey ! What 'up ? <2.1 s> ….

For telephone/virtual meeting we do not have this problem

- **Each speaker -> one stream**

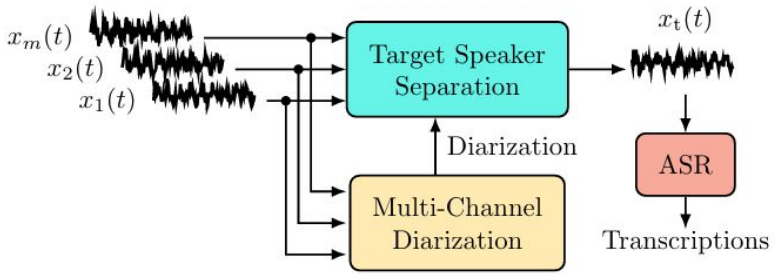
Meeting Transcription, Modular Approach (current SotA)

To handle conversational speech, right now we have to use a **modular pipeline**.

- Diarization
- Separation
- ASR

Diarization -> Guided Source Separation (GSS) [1] -> strong ASR model (e.g. Whisper)

- All participants to the past CHiME challenges (6/7/8, past 6 years) used this same scheme.



[1] Boeddeker, C at al. Front-end processing for the CHiME-5 dinner party scenario. CHiME Workshop. 2018

Current SotA: A look at CHiME-7 & 8 DASR challenges

A comprehensive benchmark [1] for **robust meeting transcription (diarization+ASR)**

Scenario	Setting	Number of Speakers	Recording Setup	Tot. Mics	Duration (minutes)
CHiME-6	dinner party	4	6 linear arrays	24	∼120 to 150
DiPCo	dinner party	4	5 circular arrays	35	∼33 to 50
Mixer 6 Speech	1-to-1 interview	2	10 heterogeneous devices	10	∼25
NOTSOFAR-1	office meeting	4-8	1 circular array	7	∼10

[1] Cornell, S., Boeddeker, C., Park, T., Huang, H., Raj, D., Wiesner, M., ... & Watanabe, S. Recent Trends in Distant Conversational Speech Recognition: A Review of CHiME-7 and 8 DASR Challenges. Computer Speech and Language. 2025

Current SotA: A look at CHiME-7 & 8 DASR challenges

A comprehensive benchmark [1] for **robust meeting transcription (diarization+ASR)**

Scenario	Setting	Number of Speakers	Recording Setup	Tot. Mics	Duration (minutes)
CHiME-6	dinner party	4	6 linear arrays	24	∼120 to 150
DiPCo	dinner party	4	5 circular arrays	35	∼33 to 50
Mixer 6 Speech	1-to-1 interview	2	10 heterogeneous devices	10	∼25
NOTSOFAR-1	office meeting	4-8	1 circular array	7	∼10

<speaker_0> <0.2 s> Hi John !<1.3 s>

<speaker_1> <1.2 s> Hey ! What 'up ?

<2.1 s> ...

Ranking based on **tcpWER** (time-constrained minimum permutation WER) [2], measures:

- segmentation
- speaker attribution
- recognition

[1] Cornell, S., Boeddeker, C., Park, T., Huang, H., Raj, D., Wiesner, M., ... & Watanabe, S. Recent Trends in Distant Conversational Speech Recognition: A Review of CHiME-7 and 8 DASR Challenges. Computer Speech and Language. 2025

[2] von Neumann, Thilo, et al. "Word Error Rate definitions and algorithms for long-form multi-talker speech recognition." IEEE Transactions on Audio, Speech and Language Processing (2025).

Current SotA: A look at CHiME-7 & 8 DASR challenges

A comprehensive benchmark [1] for **robust meeting transcription (diarization+ASR)**

Scenario	Setting	Number of Speakers	Recording Setup	Tot. Mics	Duration (minutes)
CHiME-6	dinner party	4	6 linear arrays	24	∼120 to 150
DiPCo	dinner party	4	5 circular arrays	35	∼33 to 50
Mixer 6 Speech	1-to-1 interview	2	10 heterogeneous devices	10	∼25
NOTSOFAR-1	office meeting	4-8	1 circular array	7	∼10

<speaker_0> <0.2 s> Hi John !<1.3 s>

<speaker_1> <1.2 s> Hey ! What 'up ?

<2.1 s> ….

Ranking based on **tcpWER** (time-constrained minimum permutation WER) [2], measures:

- segmentation
- speaker attribution
- recognition

[1] Cornell, S., Boeddeker, C., Park, T., Huang, H., Raj, D., Wiesner, M., ... & Watanabe, S. Recent Trends in Distant Conversational Speech Recognition: A Review of CHiME-7 and 8 DASR Challenges. Computer Speech and Language. 2025
[2]

Current SotA: A look at CHiME-7 & 8 DASR challenges

A comprehensive benchmark [1] for **robust meeting transcription (diarization+ASR)**

Scenario	Setting	Number of Speakers	Recording Setup	Tot. Mics	Duration (minutes)
CHiME-6	dinner party	4	6 linear arrays	24	∼120 to 150
DiPCo	dinner party	4	5 circular arrays	35	∼33 to 50
Mixer 6 Speech	1-to-1 interview	2	10 heterogeneous devices	10	∼25
NOTSOFAR-1	office meeting	4-8	1 circular array	7	∼10

<speaker_0> <0.2 s> Hi John !<1.3 s>

<speaker_1> <1.2 s> Hey ! What 'up ?

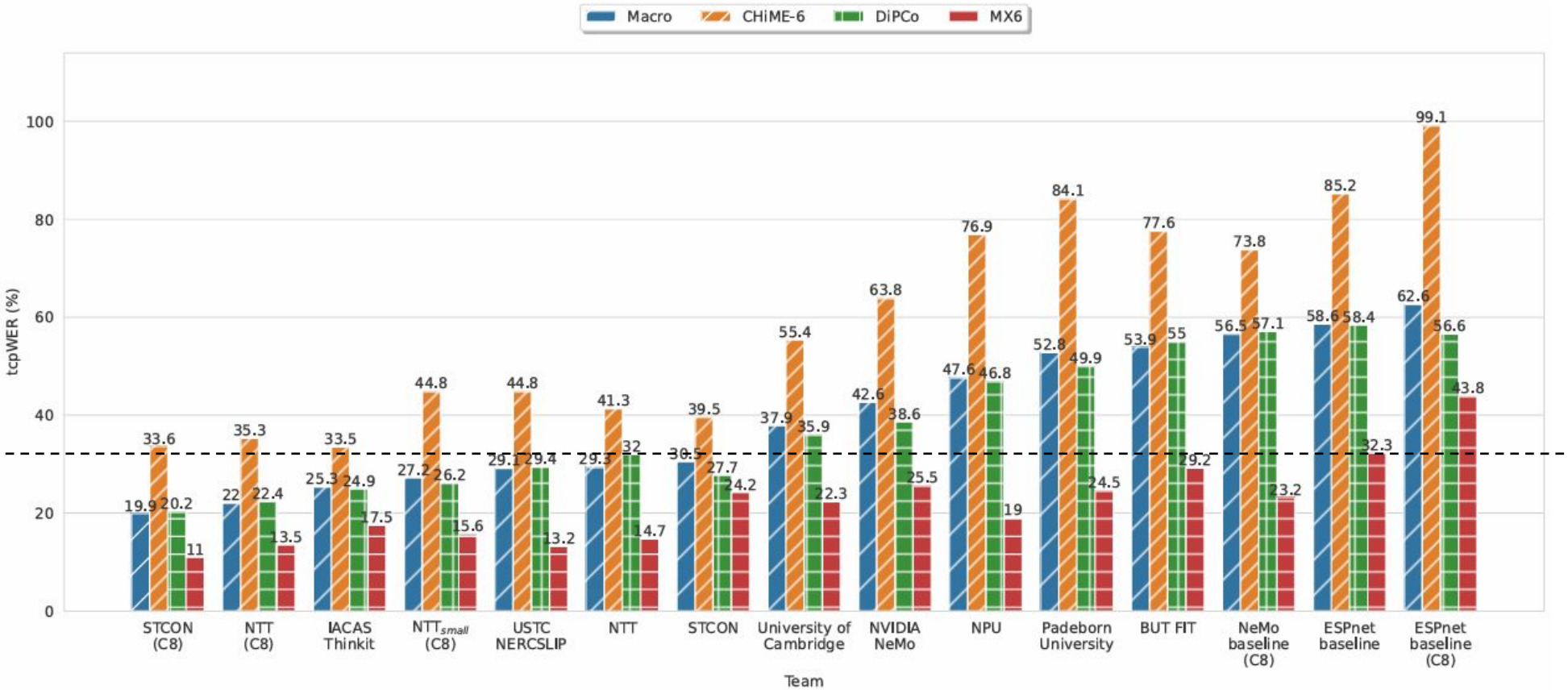
<2.1 s> ….

Ranking based on **tcpWER** (time-constrained minimum permutation WER) [2], measures:

- segmentation
- speaker attribution
- recognition

[1] Cornell, S., Boeddeker, C., Park, T., Huang, H., Raj, D., Wiesner, M., ... & Watanabe, S. Recent Trends in Distant Conversational Speech Recognition: A Review of CHiME-7 and 8 DASR Challenges. Computer Speech and Language. 2025
[2]

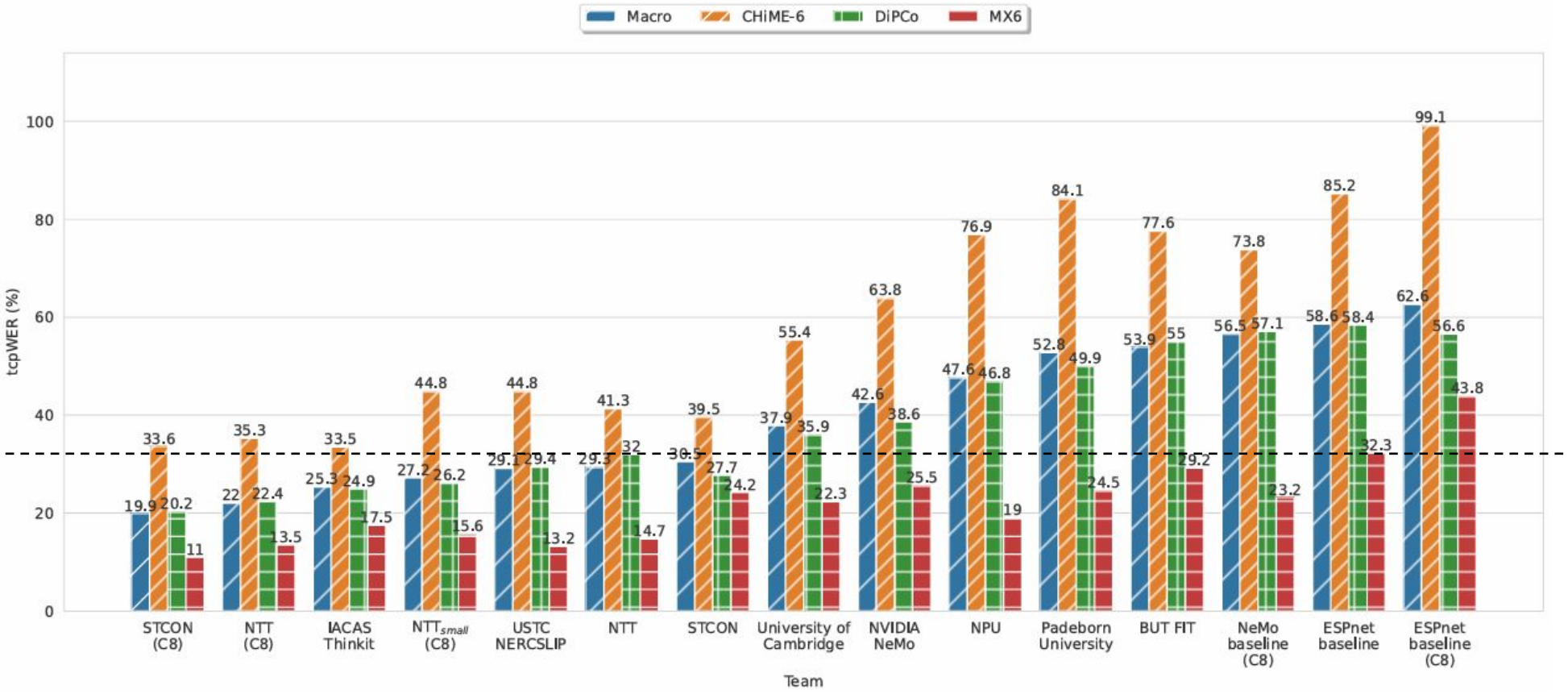
Current SotA: Joint CHiME-7 & 8 DASR Ranking



OpenAI Whisper large v3 + **oracle diarization** + Guided Source Separation [1] (GSS) (macro tcpWER % ~ 32%)

[1] Boeddeker, C at al. Front-end processing for the CHiME-5 dinner party scenario. CHiME Workshop. 2018

Current SotA: Joint CHiME-7 & 8 DASR Ranking

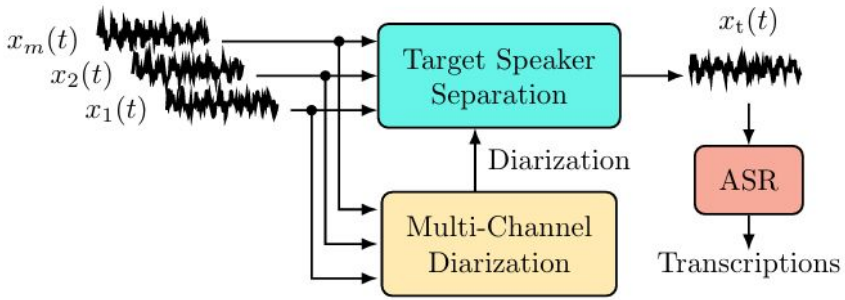


OpenAI Whisper large v3 + **oracle diarization** + Guided Source Separation [1] (GSS) (macro tcpWER % ~ 32%)

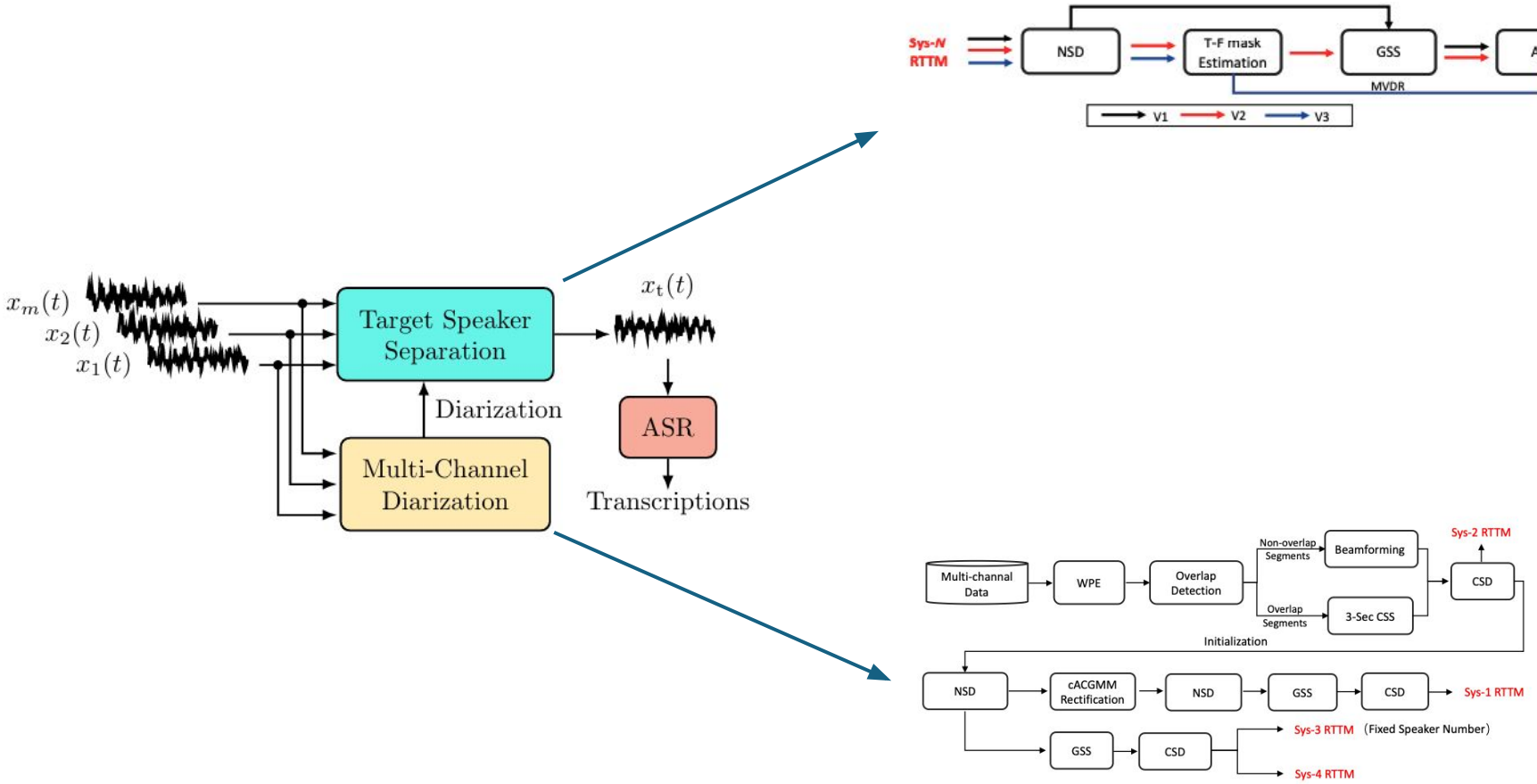
- **~ 1 out of 3 words is an error !**

[1] Boeddeker, C at al. Front-end processing for the CHiME-5 dinner party scenario. CHiME Workshop. 2018

Meeting Transcription, Modular Approach (current SotA)



Meeting Transcription, Modular Approach (current SotA)

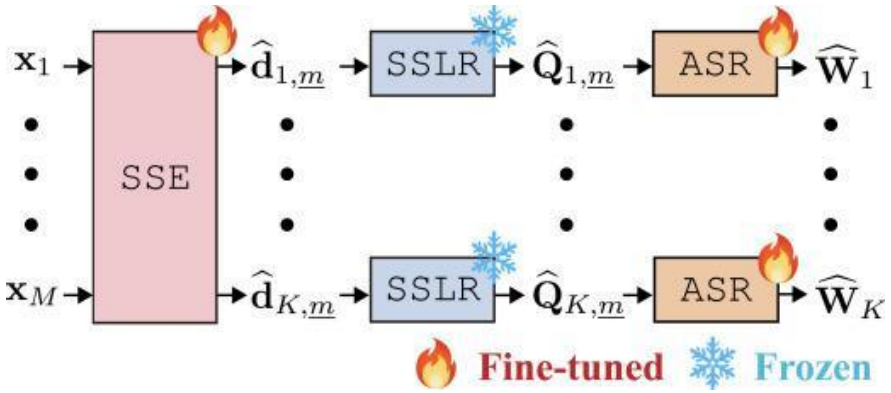


From Mitrofanov, Anton, et al. "STCON system for the CHiME-8 challenge." *CHiME Workshop* (2024).

Meeting Transcription, Modular Approach (current SotA)

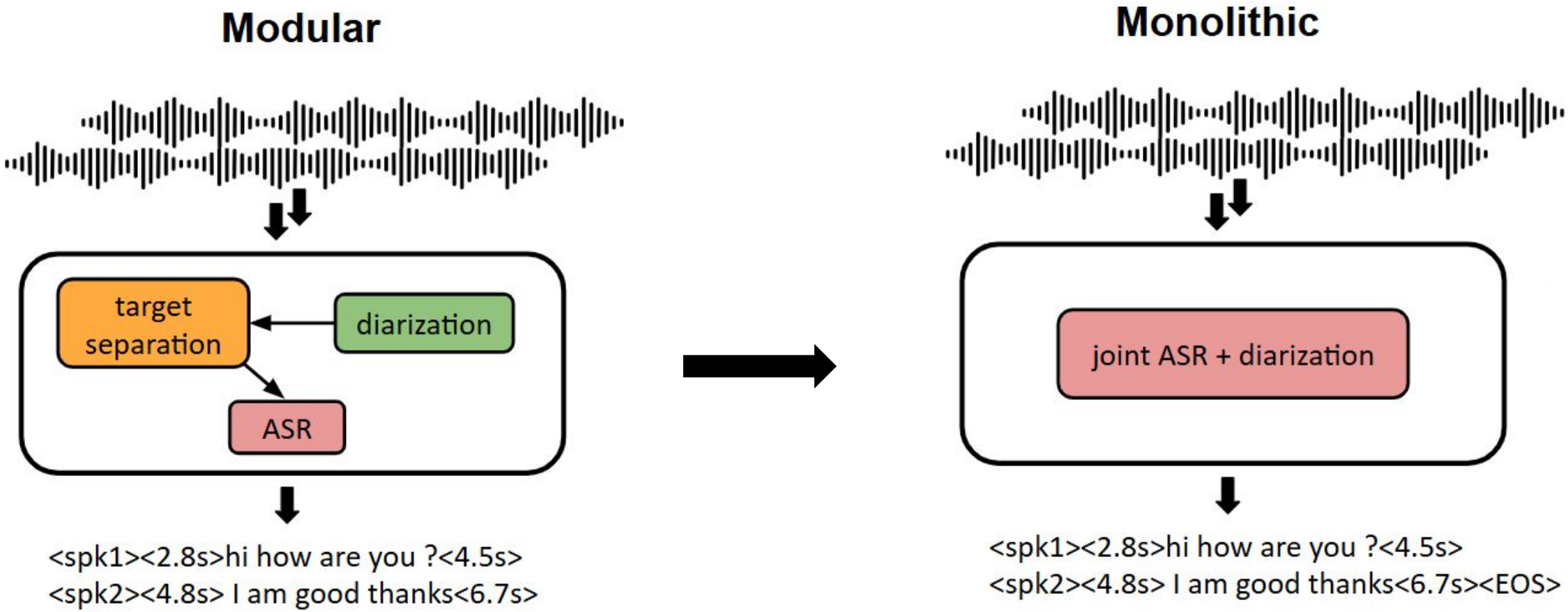
Yes ! We can optimize these modular pipelines end-to-end.

- Several prior works [1], [2] (best paper award at SLT 2022)



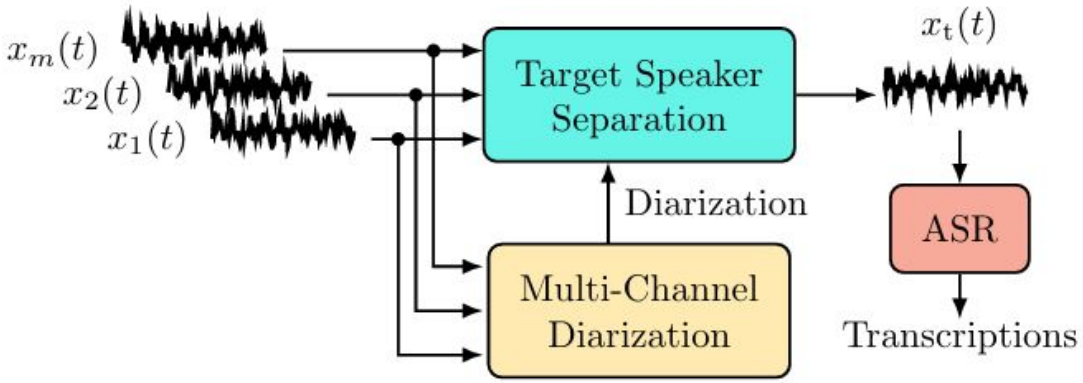
Integration of speech separation, self-supervised learning representation and ASR for multi-channel far-field recognition

[1] Masuyama, Yoshiki, et al. "End-to-end integration of speech recognition, dereverberation, beamforming, and self-supervised learning representation." SLT. 2023.
[2] Masuyama, Yoshiki, et al. "An end-to-end integration of speech separation and recognition with self-supervised learning representation." Computer Speech & Language. (2026)



Avoiding Cascading Errors: Towards e2e Meeting Transcription

e2e Meeting Transcription ?



Observations

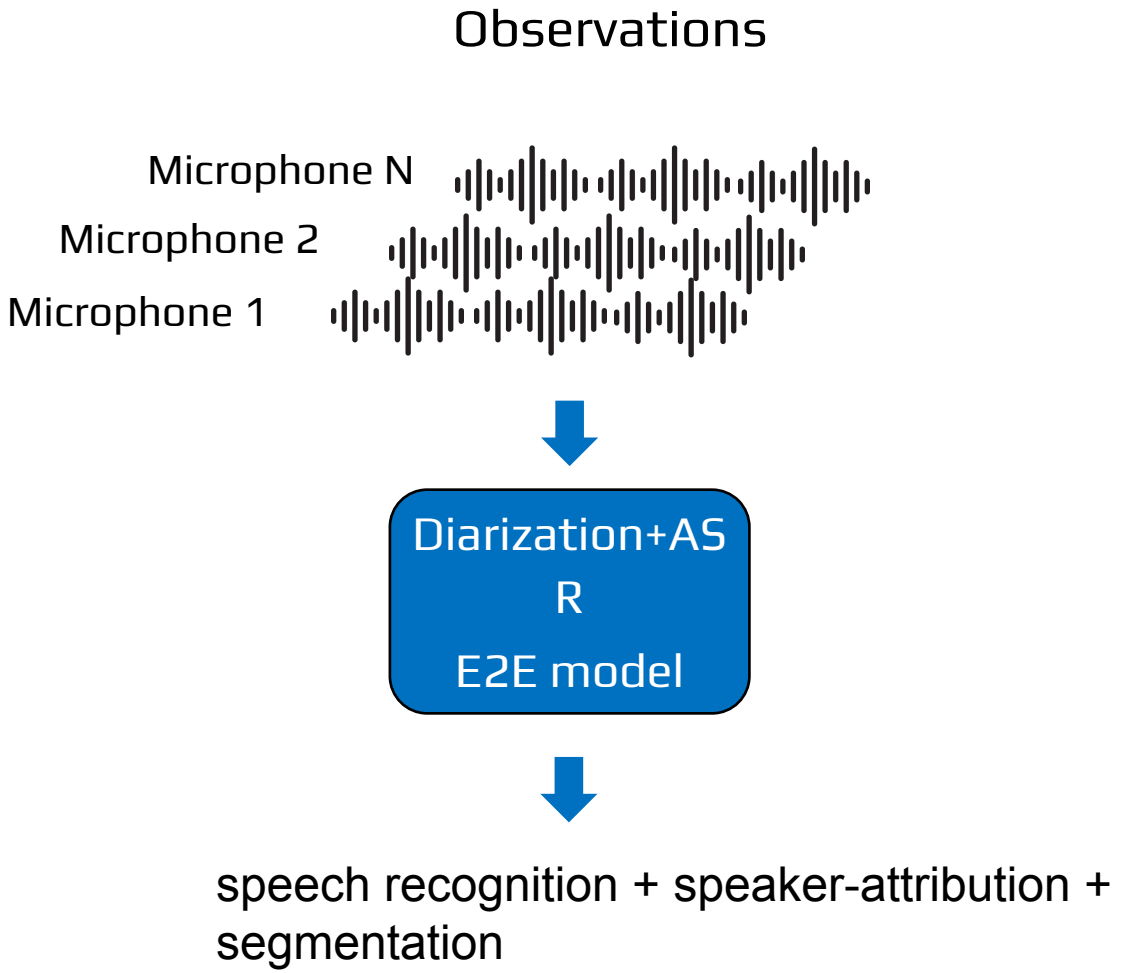
Microphone N
Microphone 2
Microphone 1

Diarization+ASR
E2E model

speech recognition + speaker-attribution + segmentation

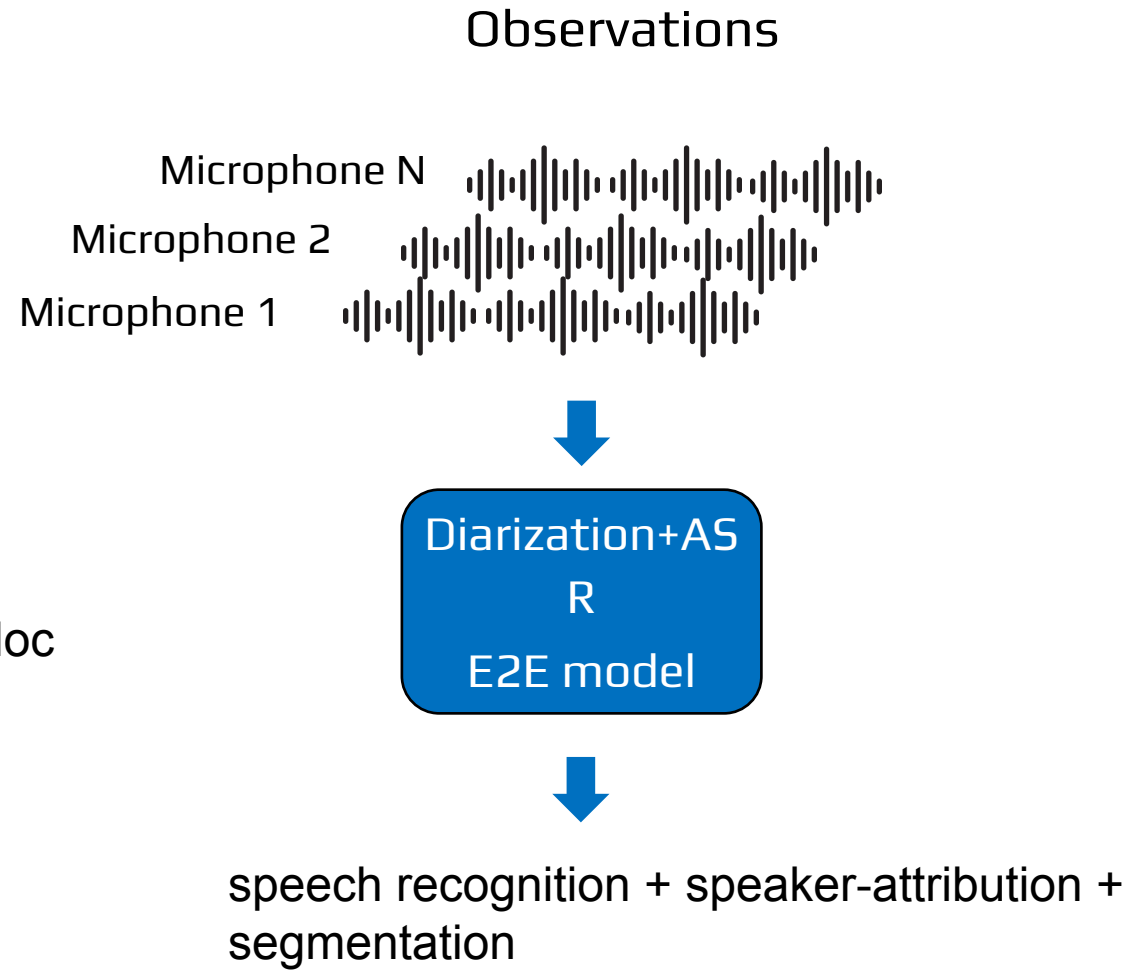
e2e Meeting Transcription ?

- ✅ No cascading errors anymore
- ✅ Optimized end-to-end for the task-at-hand
  - transcription, summarization, QA etc.



e2e Meeting Transcription ?

- ✅ No cascading errors anymore
- ✅ Optimized end-to-end for the task-at-hand
  - transcription, summarization, QA etc.

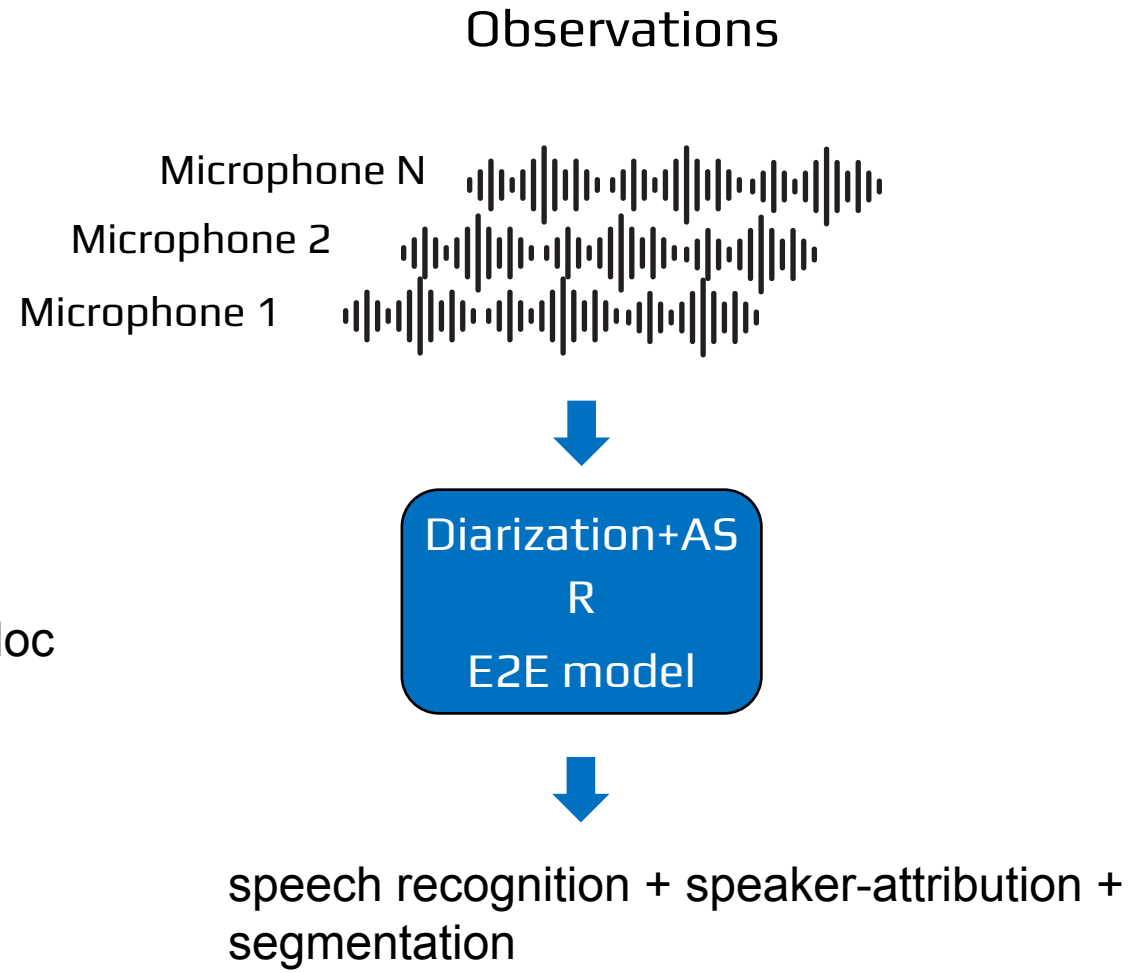


We made an initial attempt [1] at the beginning of my post-doc here

[1] Cornell, S et all. One model to rule them all? Towards end-to-end joint speaker diarization and speech recognition. ICASSP. 2023

e2e Meeting Transcription ?

- ✅ No cascading errors anymore
- ✅ Optimized end-to-end for the task-at-hand
  - transcription, summarization, QA etc.



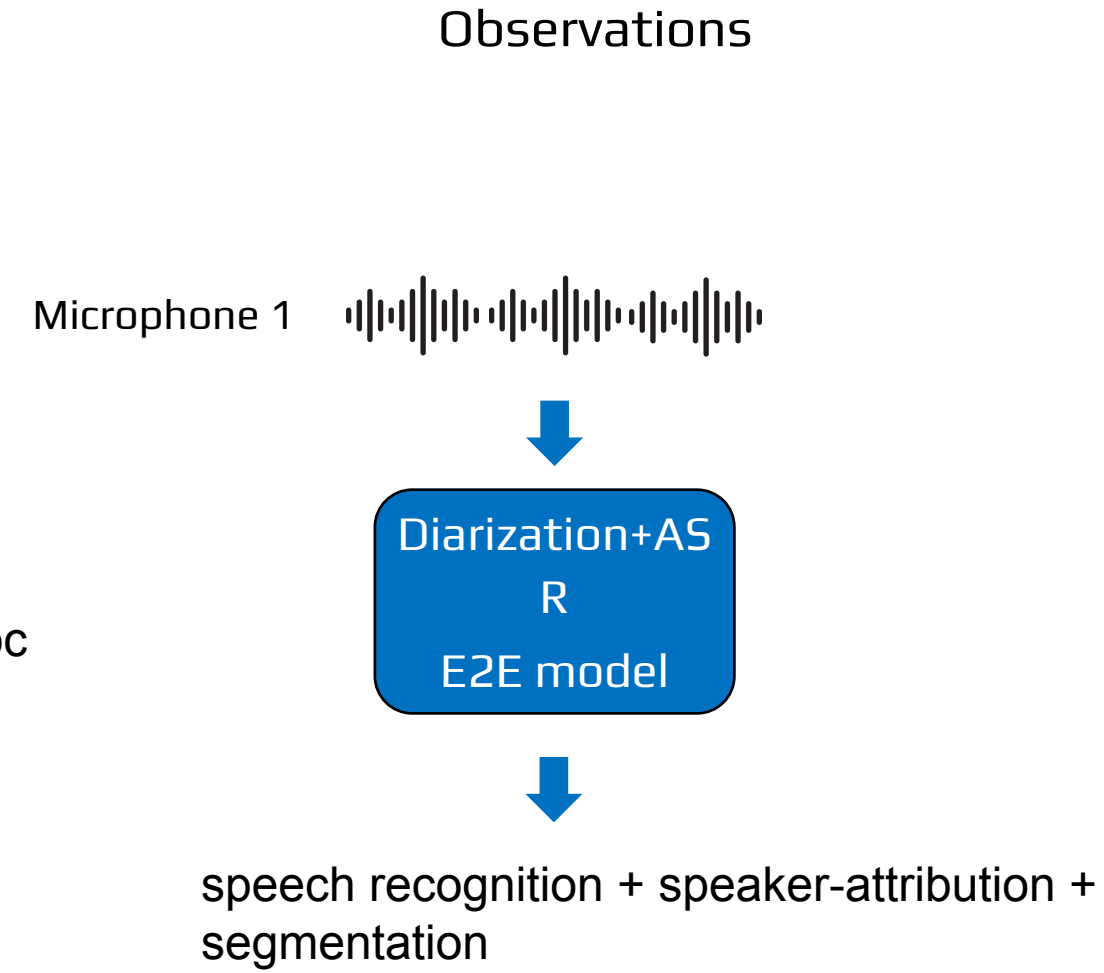
We made an initial attempt [1] at the beginning of my post-doc here

- **“Quasi” end-to-end**

[1] Cornell, S et all. One model to rule them all? Towards end-to-end joint speaker diarization and speech recognition. ICASSP. 2023

e2e Meeting Transcription ?

- ✅ No cascading errors anymore
- ✅ Optimized end-to-end for the task-at-hand
  - transcription, summarization, QA etc.

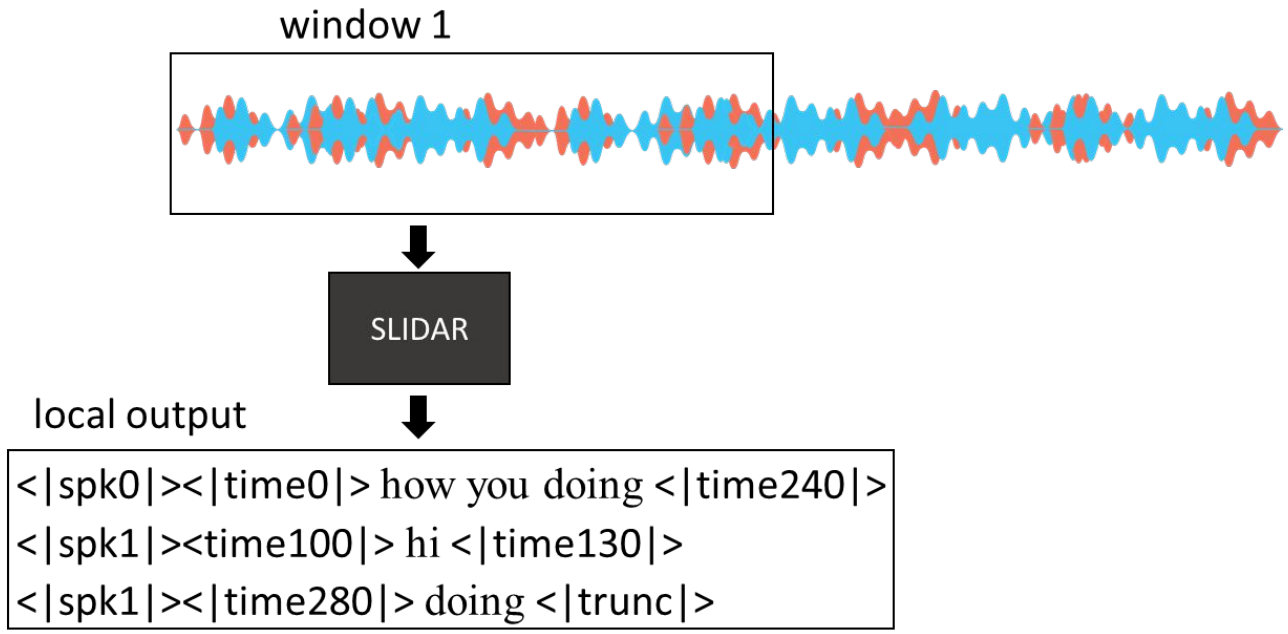


We made an initial attempt [1] at the beginning of my post-doc here

- **“Quasi” end-to-end**
- **Single-channel**

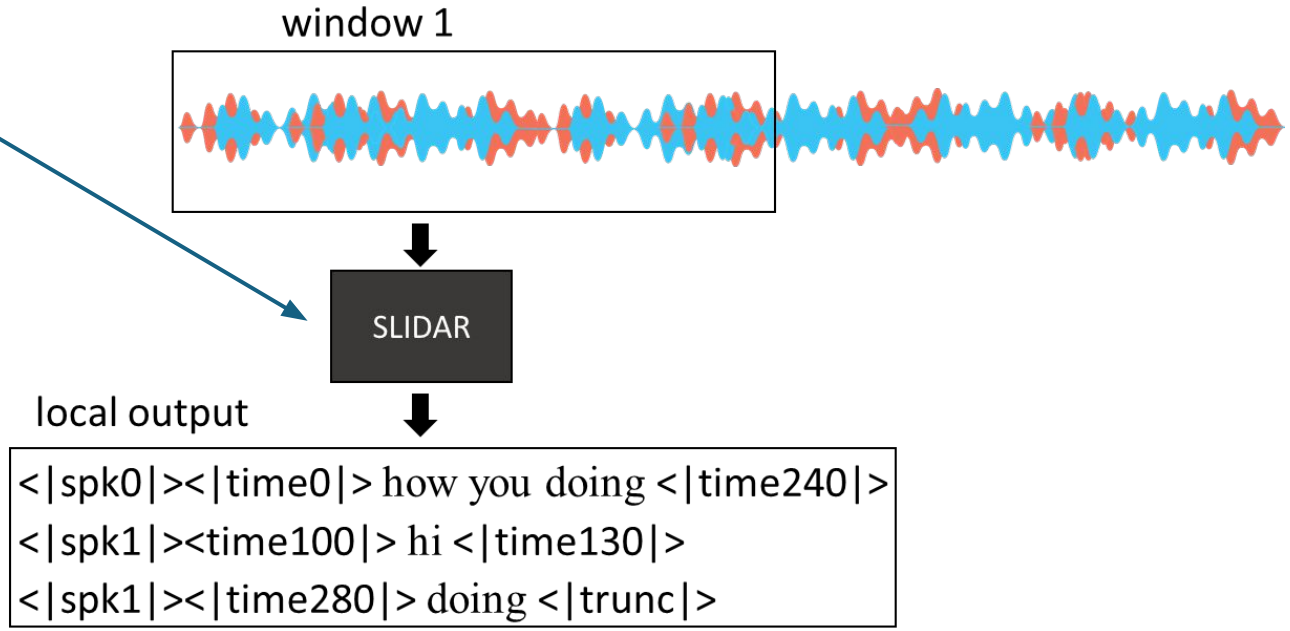
[1] Cornell, S et all. One model to rule them all? Towards end-to-end joint speaker diarization and speech recognition. ICASSP. 2023

SLIDAR: Sliding-window diarization-augmented recognitic



SLIDAR: Sliding-window diarization-augmented recognitic

We use an attention-based encoder-decoder (AED) transformer model with WavLM-large [1] as a front-end.



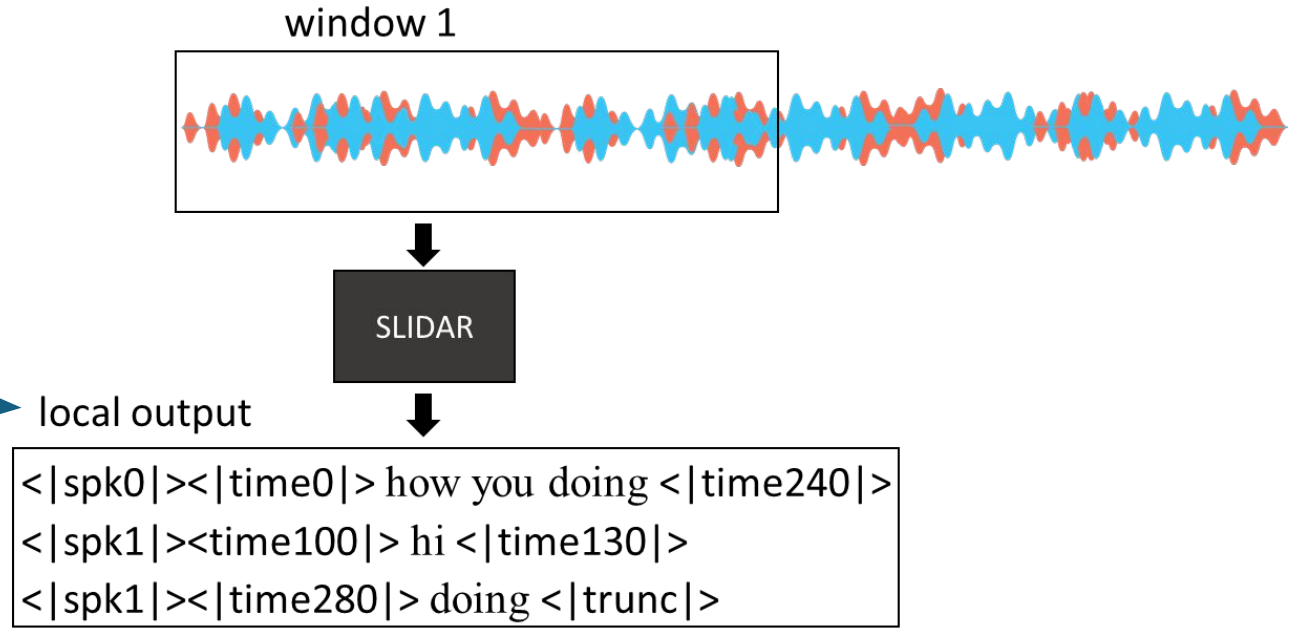
[1] Chen, Sanyuan, et al. "WavLM: Large-scale self-supervised pre-training for full stack speech processing." IEEE Journal of Selected Topics in Signal Processing (2022)

SLIDAR: Sliding-window diarization-augmented recognitio

We use an attention-based encoder-decoder (AED) transformer model with WavLM-large [1] as a front-end.

This model is trained to output, for each utterance:

1. Words
2. “Whisper [2]-style” start and stop timestamps
3. **Special speaker-id tokens**



[1] Chen, Sanyuan, et al. "WavLM: Large-scale self-supervised pre-training for full stack speech processing." IEEE Journal of Selected Topics in Signal Processing (2022)

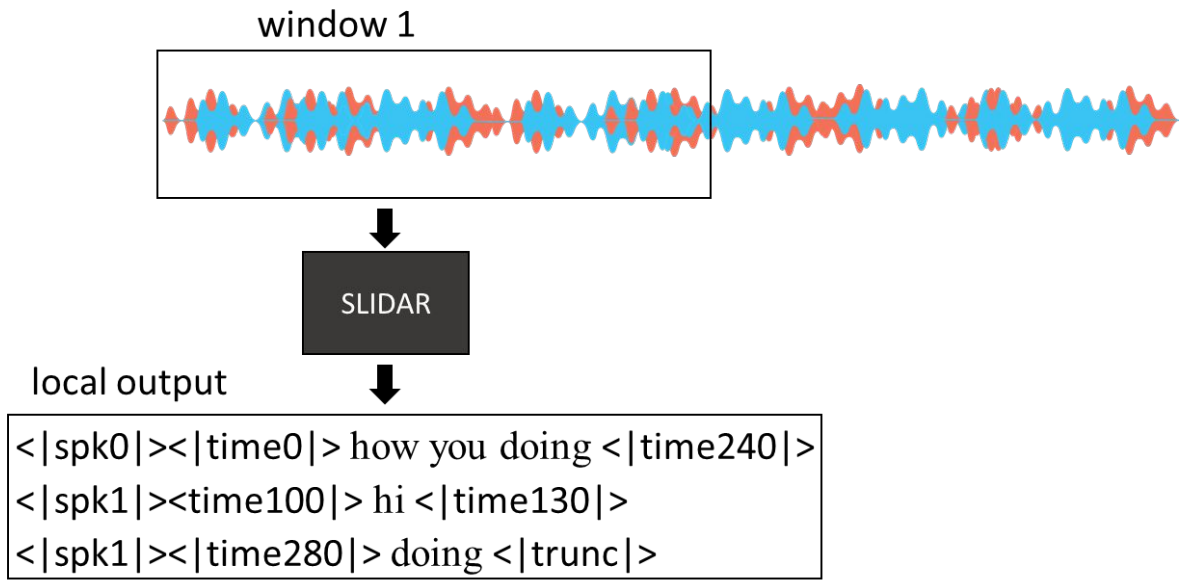
[2] Radford, Alec, et al. "Robust speech recognition via large-scale weak supervision." International conference on machine learning. PMLR, 2023.

FIFO Relative Speaker-ID Tokens

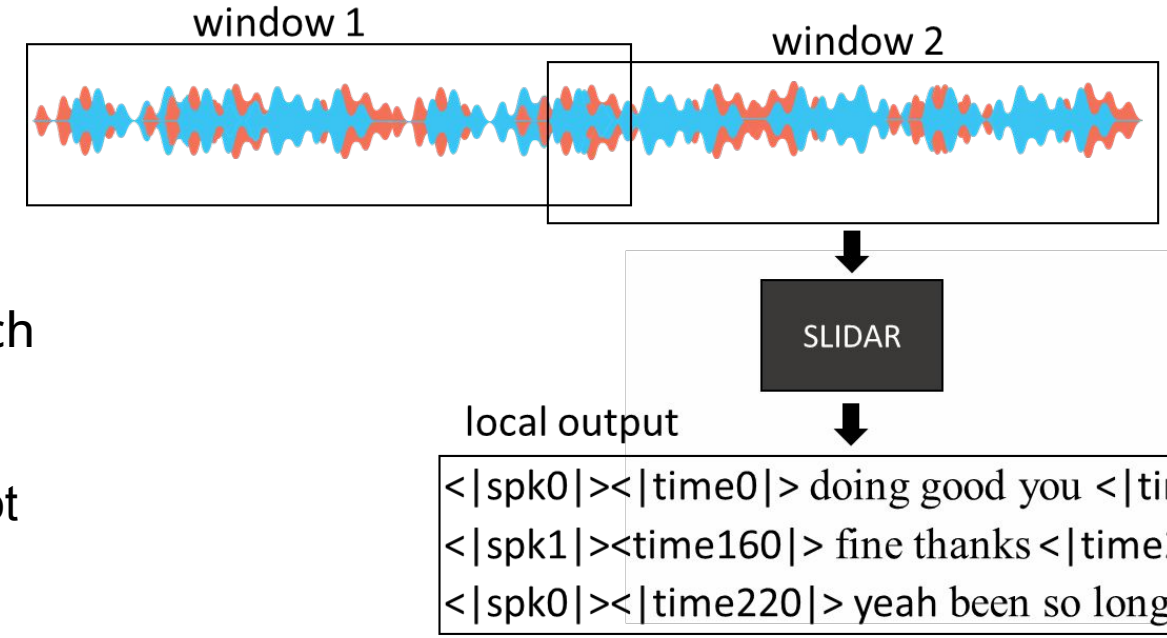
The speaker-id tokens are since the model has limited context **relative** (20 seconds).

We adopt a **first-in-first-out convention (FIFO)**:

- *spk0* is always the speaker that **speaks first** in the current chunk, *spk1* the second and so on…



FIFO Relative Speaker-ID Tokens

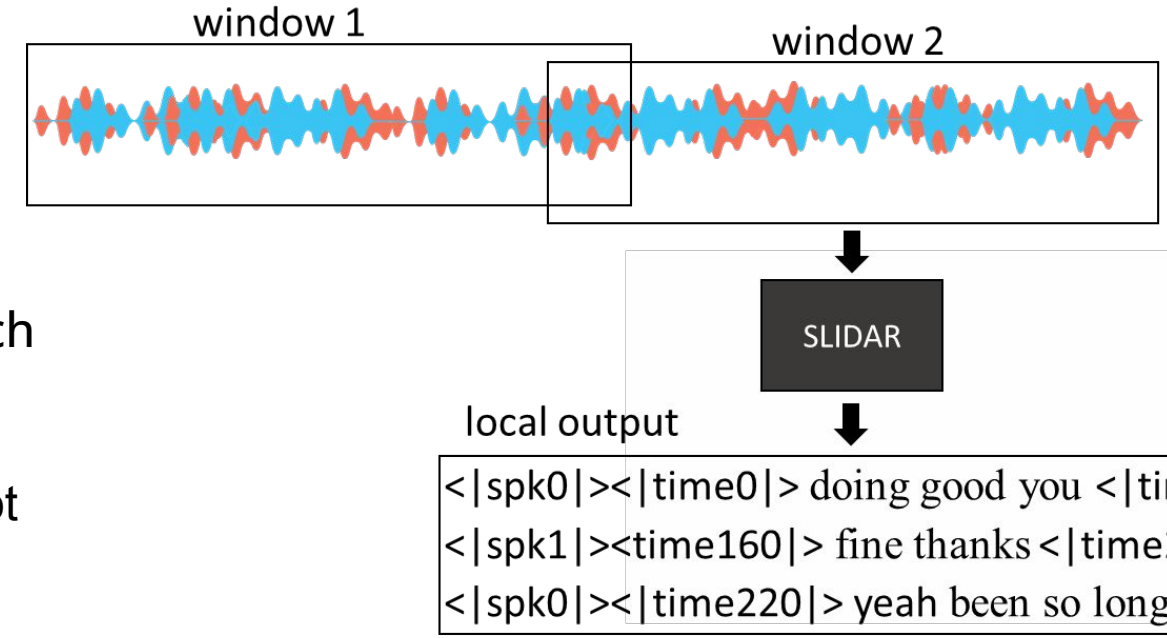


The model is applied **independently** on each window.

Between chunks the speaker labels may be not consistent !

- E.g. **<|*spk0*|> may be different.**

FIFO Relative Speaker-ID Tokens



The model is applied **independently** on each window.

Between chunks the speaker labels may be not consistent !

- E.g. **<|*spk0*|> may be different.**

*We need a **mechanism** for keeping track of the correct speaker-id assignment for the whole meeting*

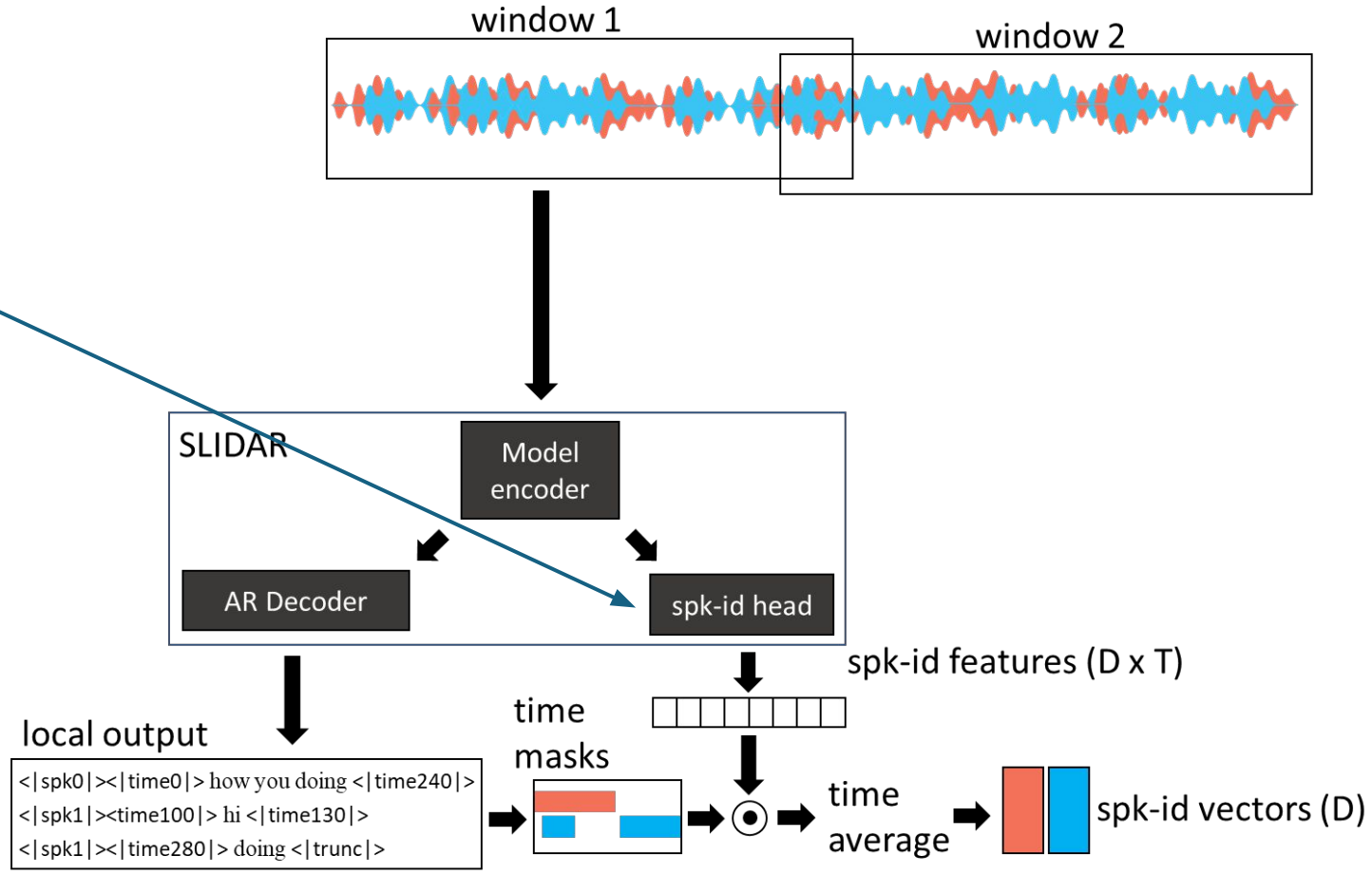
Global Speaker-ID Tracking with Embeddings

We use another output head to derive speaker-id frame-level discriminative features

and train it with a **contrastive loss [1]**:

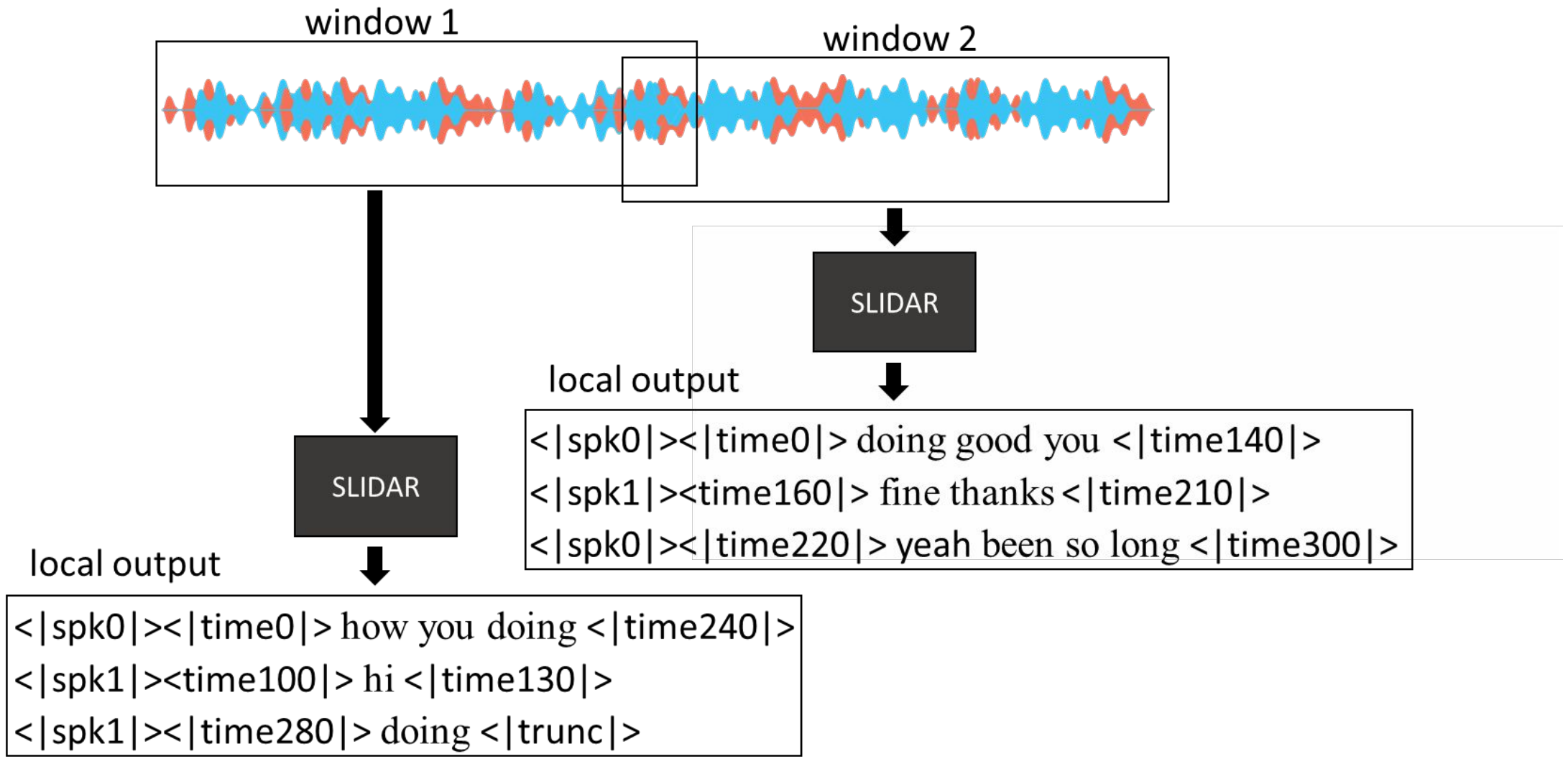
$$l_{\text{speaker}}\left(\sigma^{\star}_{i,s}, \hat{\mathbf{e}}_{i,s}\right) = -\ln\left(\frac{\exp\left(-d\left(E_{\sigma^{\star}_{i,s}}, \hat{\mathbf{e}}_{i,s}\right)\right)}{\sum_{m=1}^{M}\exp\left(-d\left(E_m, \hat{\mathbf{e}}_{i,s}\right)\right)}\right)$$

$$d\left(E_m, \hat{\mathbf{e}}_{i,s}\right) = \alpha\|E_m - \hat{\mathbf{e}}_{i,s}\|^2 + \beta,$$

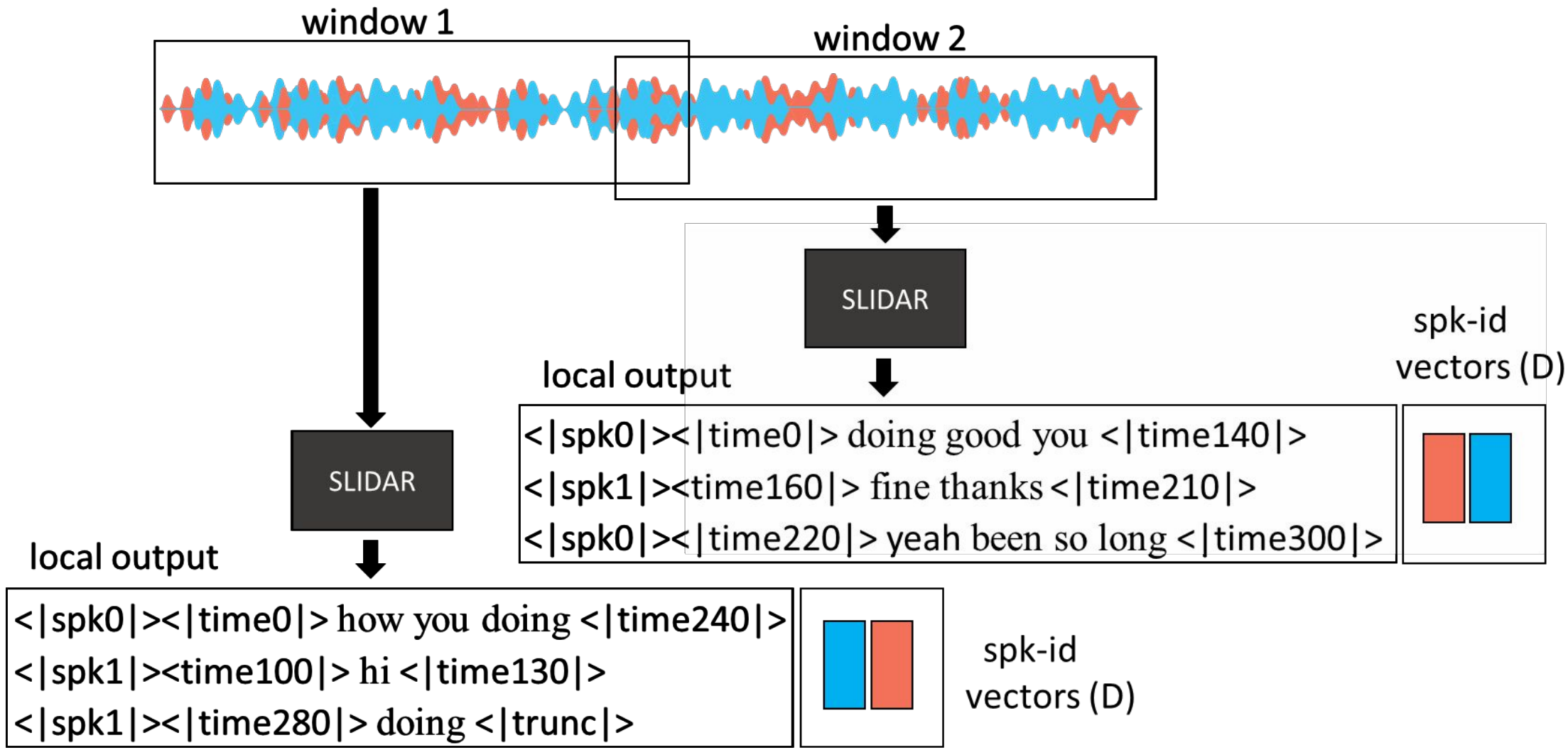


[1] Kinoshita, Keisuke, Marc Delcroix, and Naohiro Tawara. "Integrating end-to-end neural and clustering-based diarization: Getting the best of both worlds." ICASSP, 2021.

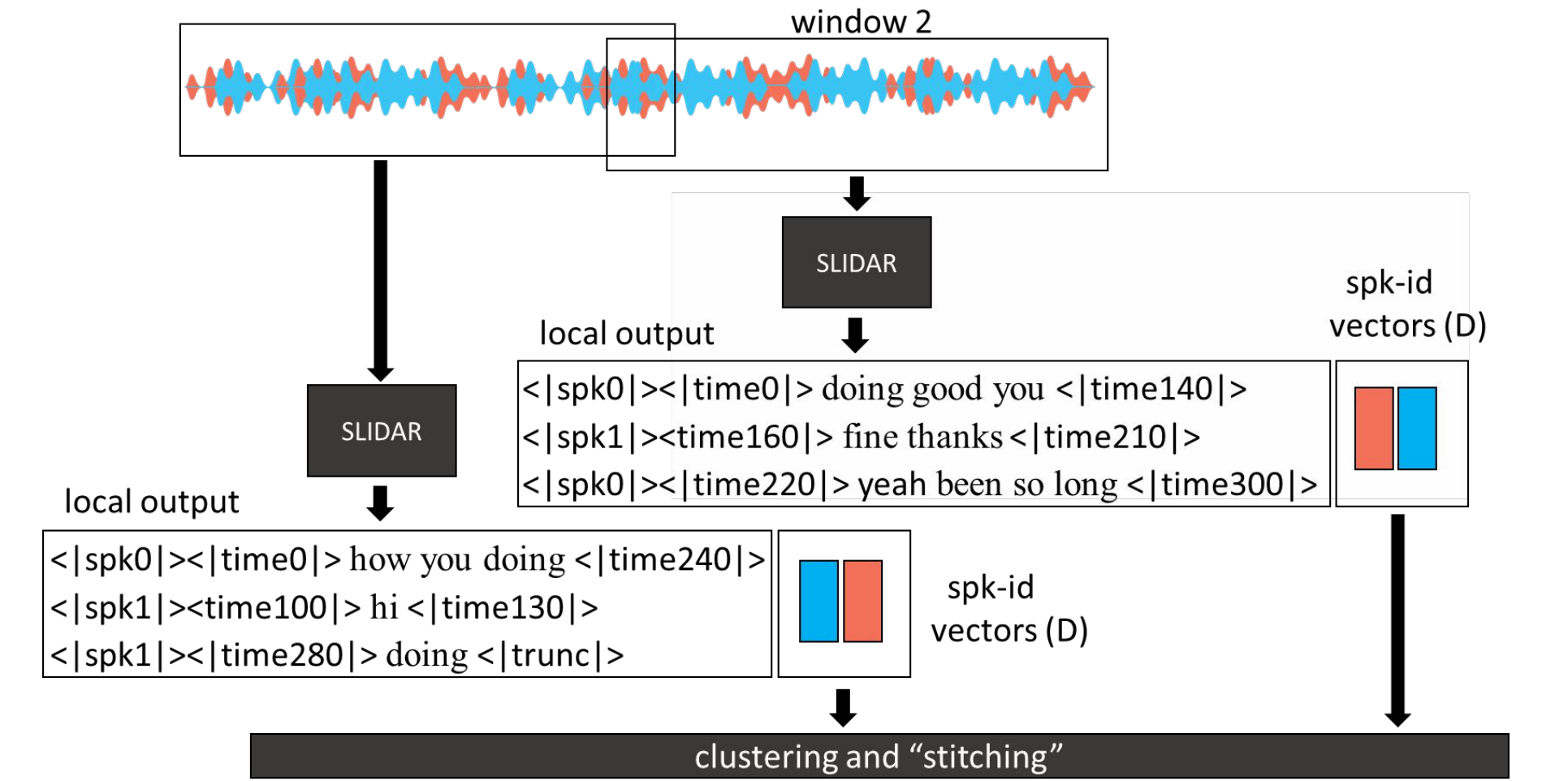
Clustering and Stitching



Clustering and Stitching



Clustering and Stitching



Training

We train the model on ~15k hours of data (but encoder is based on WavLM large):

1. Real-world long-form meeting data: AMI, Mixer 6, CHiME-6
2. Simulated meeting data (using NeMo Multi-speaker simulator [1]).

To increase data we also use **standard pre-segmented ASR data**

a. E.g. including LibriSpeech, close-talk AMI microphones and Mixer 6 weakly annotated training set.

[1] Park, Tae Jin, et al. "Property-aware multi-speaker data simulation: A probabilistic modelling technique for synthetic data generation." Interspeech. 2023

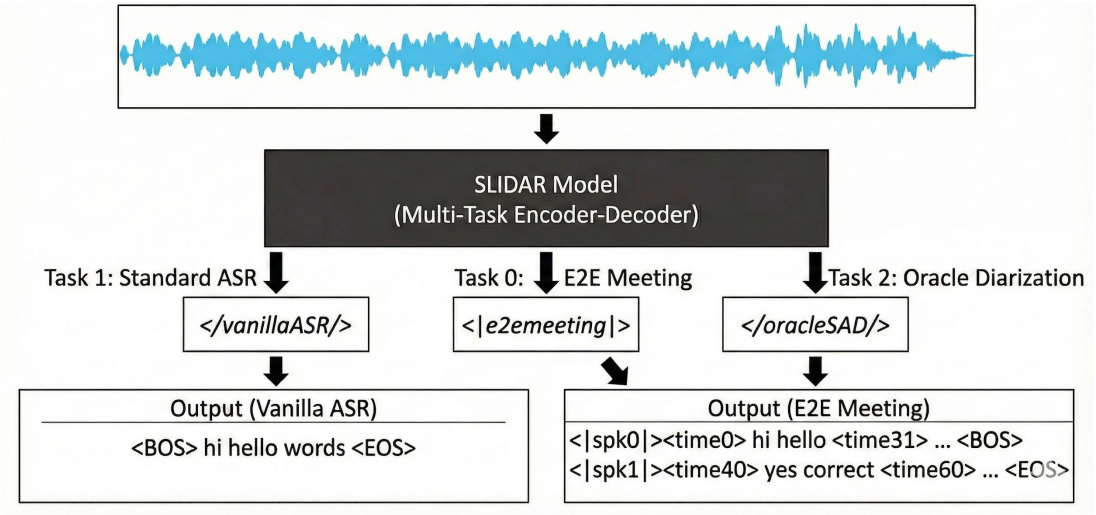
Training: Multi-Task Learning

To train the model on pre-segmented ASR data we define a multi-task prompt token (no multi-speaker): **<|*vanillaASR*|>**

We also train the model to use an **<|*oracleSAD*|>** oracle diarization prompt to speed up convergence.

Examples:

- <|*vanillaASR*|> <BOS> words <EOS>
- <|*oracleSAD*|> <|spk0|><time0><time31>… <BOS> <|spk0|><time0> hi hello <time31>… <EOS>



Experimental Results

We test the proposed system on the AMI dataset:

- 16 (eval set) meetings between 3 to 5 speakers
- Single distant microphone scenario
- Relatively quiet office meeting

Experimental Results

Performance measured in terms of:

1. **Diarization error rate (DER)**
    - Speaker-attribution + temporal localization

Reference


```
{
 "end_time": "11.370",
 "start_time": "11.000",
 "words": "so ummm",
 "speaker": "P03",
 "session_id": "S05"},
{
"end_time": "14.110",
"start_time": "12.100",
"words": "where is he?"
"speaker": "P01",
"session_id": "S05"
}
```


System Hypothesis


```
{
 "end_time": "11.350",
 "start_time": "11.010",
 "words": "so",
 "speaker": "spk1",
 "session_id": "S05" },
{
 "end_time": "14.150",
 "start_time": "12.000",
 "words": "Where is",
 "speaker": "spk2",
 "session_id": "S05"
}
```

Experimental Results

Performance measured in terms of:

1. **Diarization error rate (DER)**
   - Speaker-attribution + temporal localization
2. **Concatenated minimum permutation error rate (cpWER) [1]**
   - WER “flavor” suitable for long-form multi-speaker transcription
   - Speaker-attribution + word recognition
     - Temporal localization is not measured.

Reference


```
{
 "end_time": "11.370",
 "start_time": "11.000",
 "words": "so ummm",
 "speaker": "P03",
 "session_id": "S05"},
{
 "end_time": "14.110",
 "start_time": "12.100",
 "words": "where is he?"
 "speaker": "P01",
 "session_id": "S05"
}
```


System Hypothesis


```
{
 "end_time": "11.350",
 "start_time": "11.010",
 "words": "so",
 "speaker": "spk1",
 "session_id": "S05" },
{
 "end_time": "14.150",
 "start_time": "12.000",
 "words": "Where is",
 "speaker": "spk2",
 "session_id": "S05"
}
```

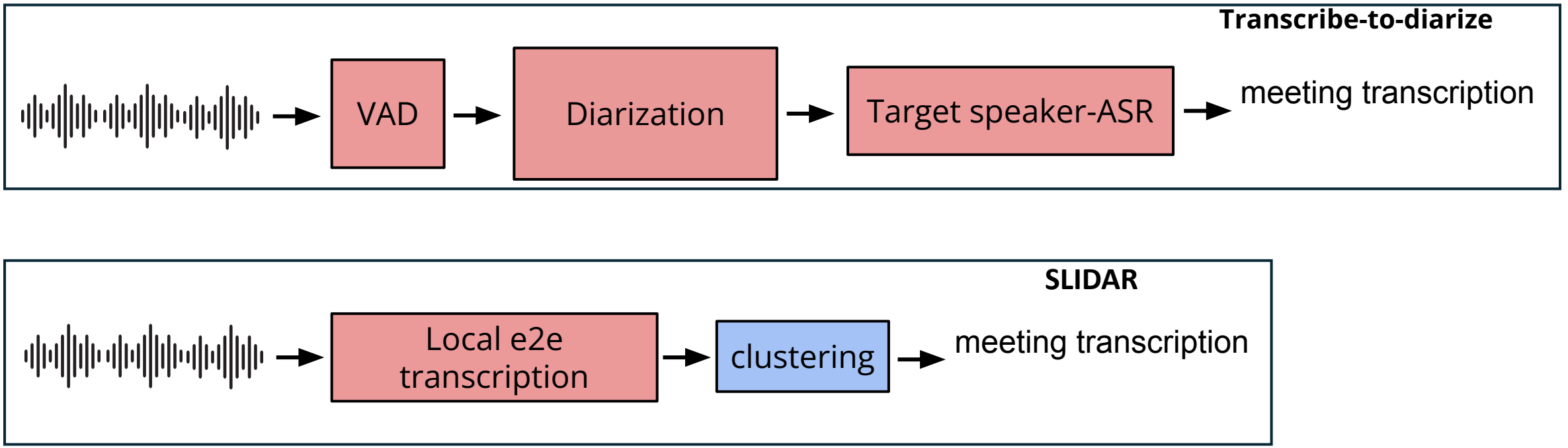

[1] Watanabe, Shinji, et al. "CHiME-6 challenge: Tackling multi-speaker speech recognition for unsegmented recordings.” CHiME Workshop. 2020

Results on AMI Meeting Distant Single Microphone

Diarization+Transcription Systems	**DER (%)**	**cpWER (%)**
Transcribe-to-Diarize (SotA) [1]	**28.12**	**24.9**
SLIDAR (proposed)	31.52	**24.5**

[1] Kanda, Naoyuki, et al. "Transcribe-to-diarize: Neural speaker diarization for unlimited number of speakers using end-to-end speaker-attributed ASR." ICASSP, 2022.

Results on AMI Distant Single Mic



Results on AMI Meeting Distant Single Microphone

Diarization+Transcription Systems	**DER (%)**	**cpWER (%)**
Transcribe-to-Diarize [1]	28.12	24.9
SLIDAR (proposed)	31.52	24.5
DiariZen + SE-DiCoW [2]	**10.4**	**18.5**

[2] Polok, Alexander, et al. "SE-DiCoW: Self-Enrolled Diarization-Conditioned Whisper." Accepted at ICASSP (2026).

[1] Kanda, Naoyuki, et al. "Transcribe-to-diarize: Neural speaker diarization for unlimited number of speakers using end-to-end speaker-attributed ASR." ICASSP, 2022.

Limitations (Let‘s be honest)

- **Is it really more robust ?**
  - ✅ no cascading errors

Limitations (Let‘s be honest)

- **Is it really more robust ?**
  - ✅ no cascading errors
  - But joint modeling of diarization and ASR can fail in the same way

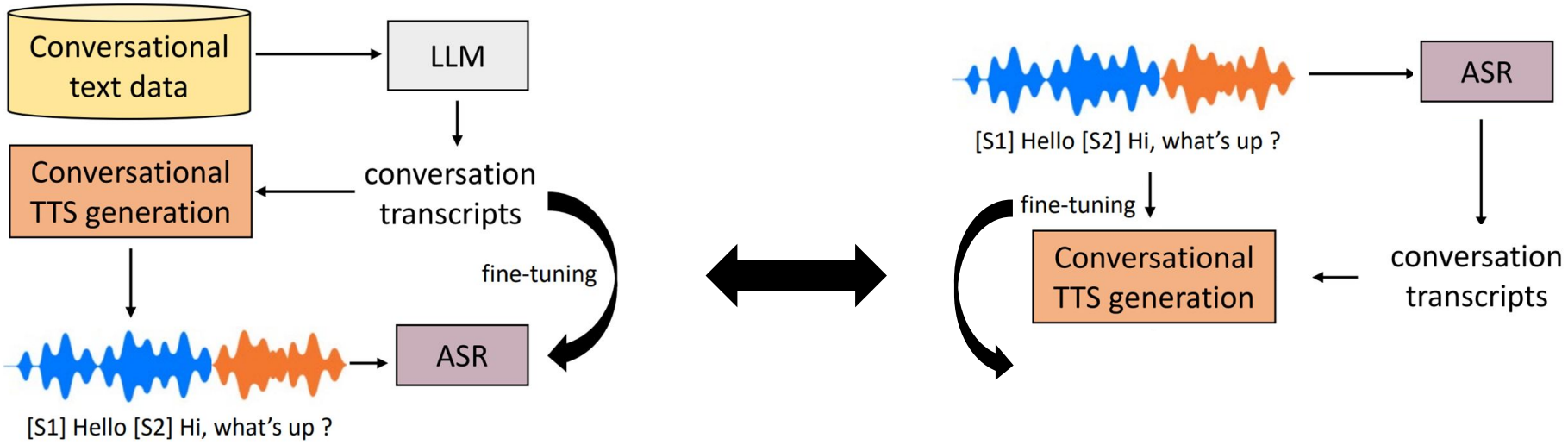
Limitations (Let‘s be honest)

- **Is it really more robust ?**
  - ✅ no cascading errors
  - But joint modeling of diarization and ASR can fail in the same way
    - E.g. ASR failures due to dialectal speech errors can affect diarization
    - Modular systems are less affected (diarization is very robust to language changes)

Limitations (Let's be honest)

- **Is it really more robust ?**
  - ✅ no cascading errors
  - But joint modeling of diarization and ASR can fail in the same way
    - E.g. ASR failures due to dialectal speech errors can affect diarization
    - Modular systems are less affected (diarization is very robust to language changes)

- **No Data, No (Cocktail) Party**
  - We need labeled multi-speaker data which is scarce (except for telephone/virtual meetings)



Addressing data scarcity:

Generating data with text-to-speech and LLMs for conversational speech recognition

Why Can't We Just Scale (in some domains) ?

Can’t we just feed **more data** to the e2e model ?

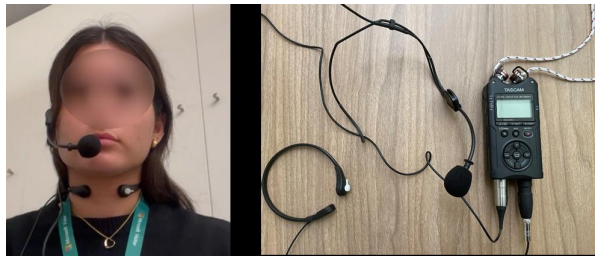
- **We cannot collect conversational data at scale for many domains**
  - Doctor-patient, business, government etc.
  - Any far-field multi-channel data

Why Can't We Just Scale (in some domains) ?

Can’t we just feed **more data** to the e2e model ?

- **We cannot collect conversational data at scale for many domains**
  - Doctor-patient, business, government etc.
  - Any far-field multi-channel data

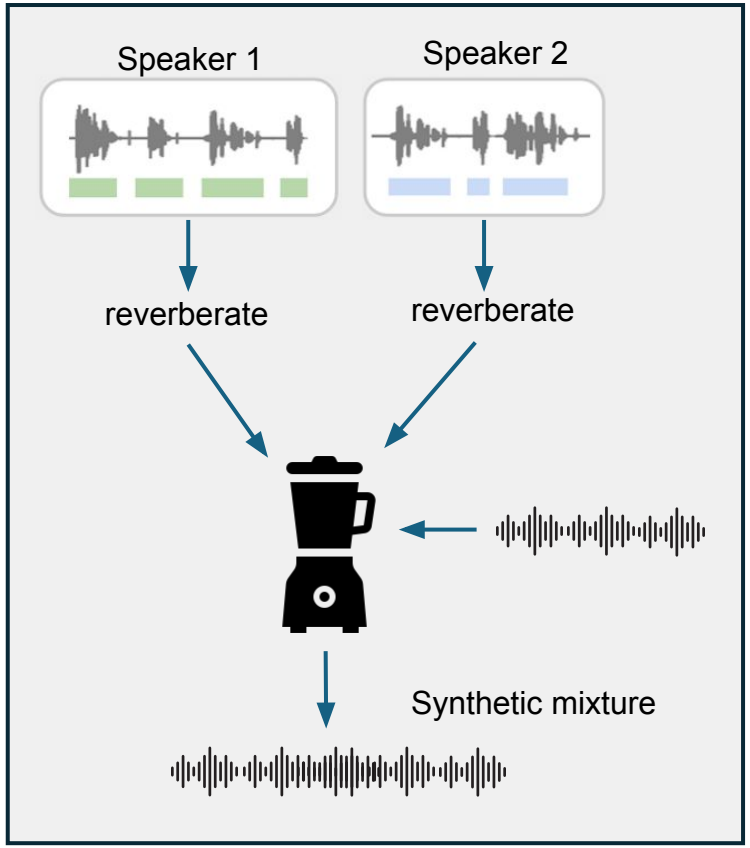
  - Cost in the annotation & collection
    - ~ few thousand hours of meeting multi-channel data



Multi-Speaker Simulation Toolkits

"**Cut, Paste, Mix**" (traditional approach):

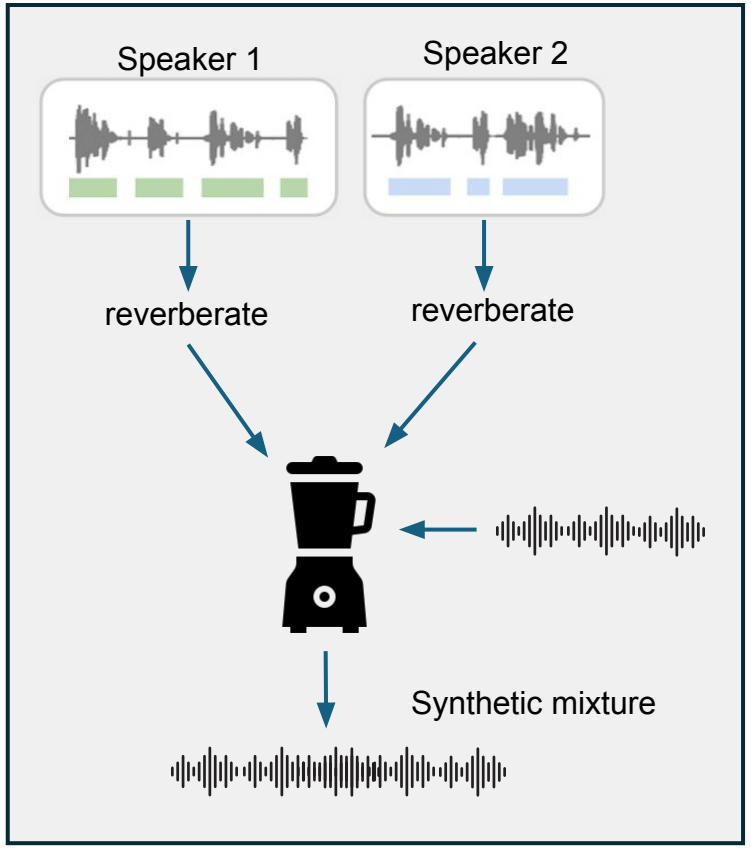
- Sample from e.g. audiobook corpora (LibriSpeech) some utterances
  - Optionally reverberate them (e.g. simulate room impulse responses with image method)
- Optionally sample some background and foreground noises
- --> Mix together



Multi-Speaker Simulation Toolkits

Pros ✅:
- Fast to run/Highly scalable
- Easy to control

Cons ❌:
- **No semantic coherence**
- **Unrealistic turn taking and multi-speaker interaction**

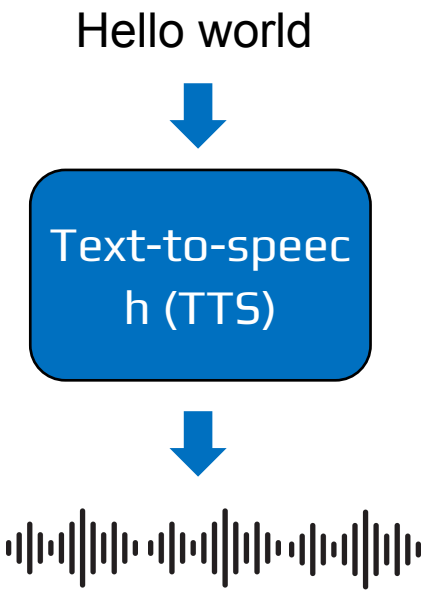


Several toolkits:
- NeMo Multi-speaker simulator [1]
- JSALT2020 simulator (github.com/jsalt2020-asrdiar/jsalt2020_simulate)
- Multipurpose Multi-Speaker Mixture Signal Generator [3]

[1] Park, Tae Jin, et al. "Property-aware multi-speaker data simulation: A probabilistic modelling technique for synthetic data generation." Interspeech 2023

[3] Cord-Landwehr, Tobias, et al. "MMS-MSG: A multi-purpose multi-speaker mixture signal generator." IWAENC 2022.

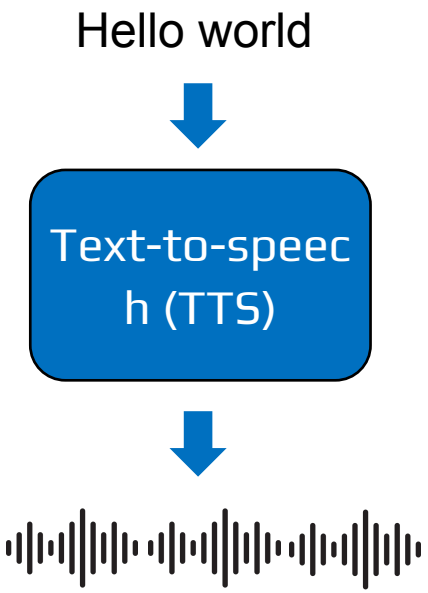
Synthetic Data with Text-to-Speech (TTS) ?



Synthetic Data with Text-to-Speech (TTS) ?

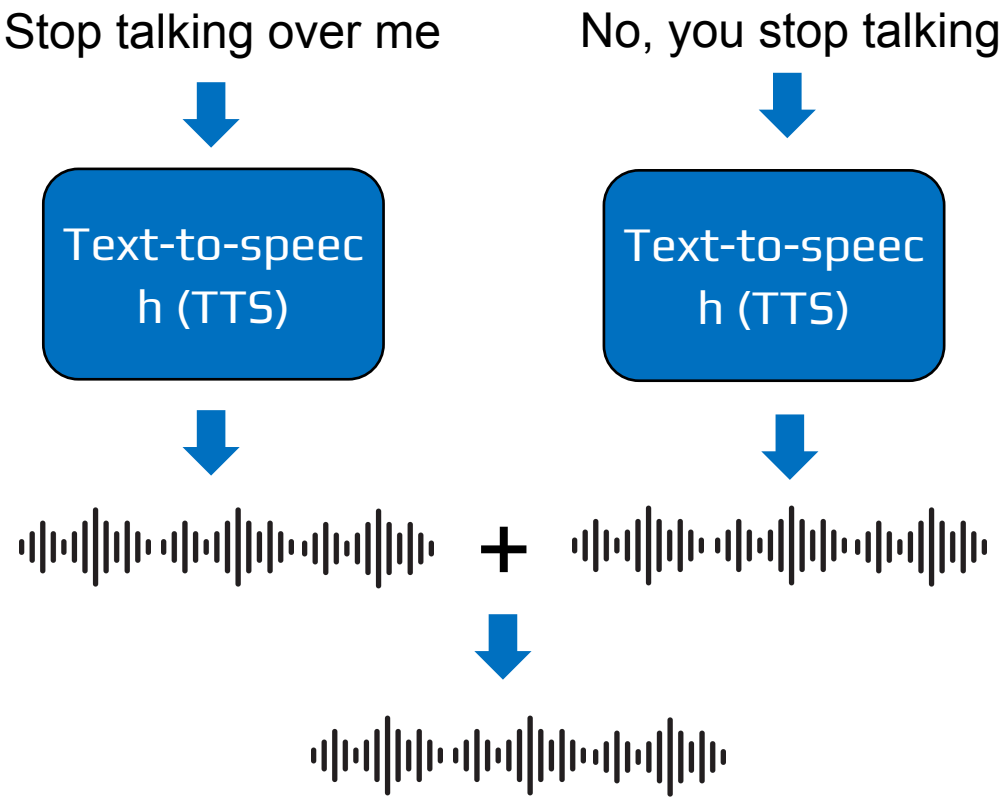
Many previous works but only in the **single speaker case**:

[1] Rosenberg, Andrew, et al. "Speech recognition with augmented synthesized speech." ASRU. 2019.
[2] Rossenbach, Nick, et al. "Generating synthetic audio data for attention-based speech recognition systems." ICASSP. 2020.
[3] Fazel, Amin, et al. "SynthASR: Unlocking synthetic data for speech recognition." Interspeech. 2021
[4] Tjandra, Andros et al. "Machine speech chain." IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2020



Synthetic Multi-Speaker Data with TTS ?

We can simply mix **two independent** TTS generations !



Synthetic Multi-Speaker Data with TTS ?

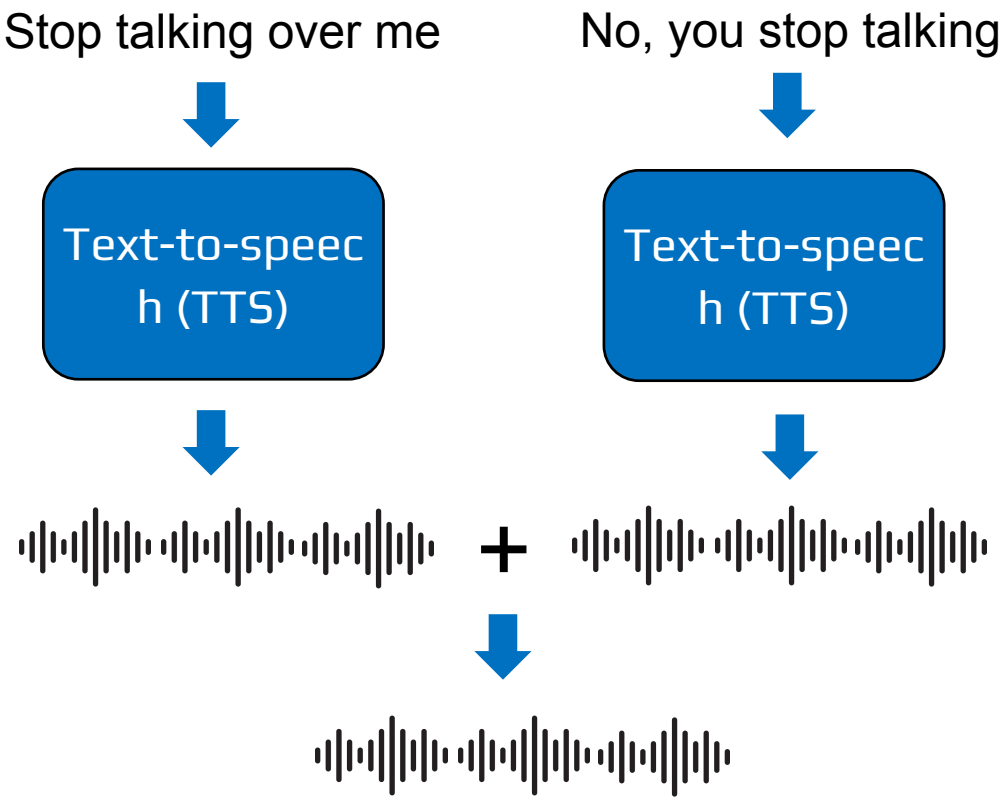
We can simply mix **two independent** TTS generations !

Pros ✅:

- Semantic coherence
  - We have full control the content

Cons ❌:

- Slower to run than naïve simulation
- **Unrealistic turn taking and multi-speaker interaction**



Synthetic Multi-Speaker Data with TTS ?

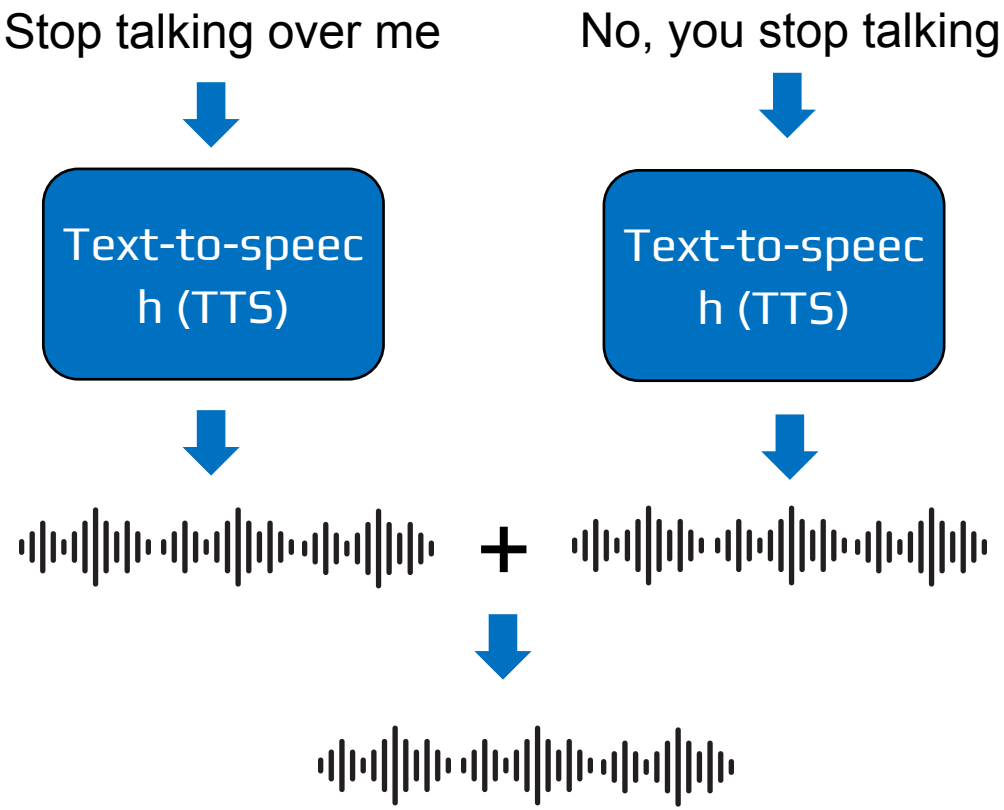
We can simply mix **two independent** TTS generations !

Pros ✅:

- Semantic coherence
  - We have full control the content

Cons ❌:

- Slower to run than naïve simulation
- **Unrealistic turn taking and multi-speaker interaction**



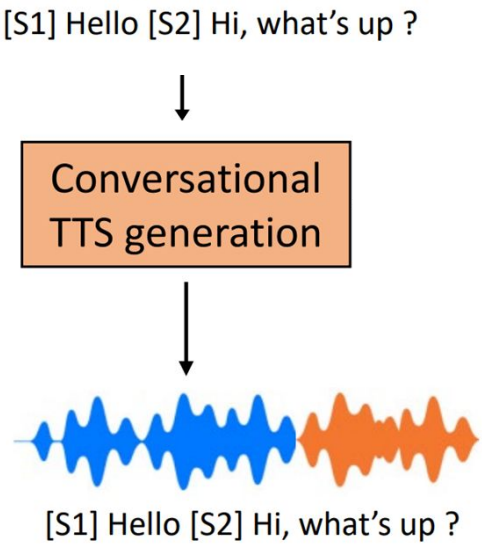
Example with CoquiAI xTTS-2 [1]

(speaker IDs randomly sampled from LibriSpeech)

[1] https://huggingface.co/coqui/XTTS-v2

Proposed Approach

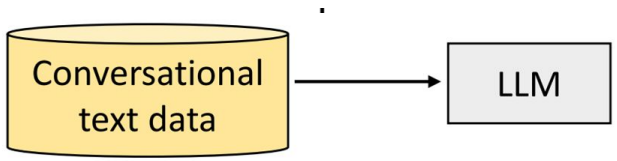
1. Train a TTS system to generate multi-speaker speech.



Cornell, S., Darefsky, J., Duan, Z., & Watanabe, S. (2024). Generating data with text-to-speech and large-language models for conversational speech recognition.

Proposed Approach

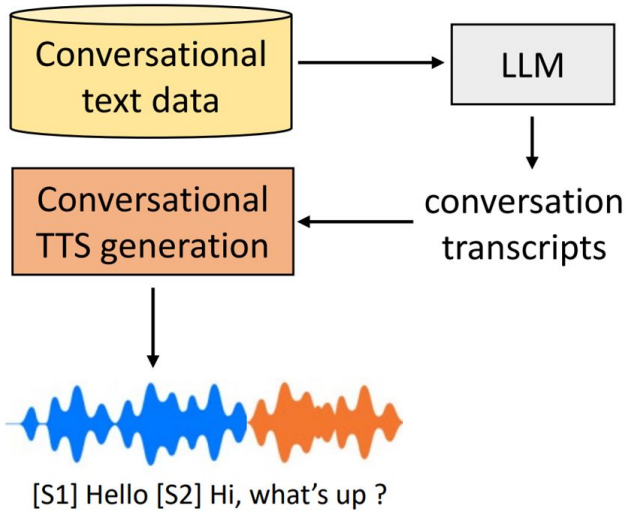
1. Train a TTS system to generate multi-speaker speech.
2. Use an LLM to augment or generate conversational speech transcripts.
   - Should be reasonably effective as they are trained on “conversational text”



Cornell, S., Darefsky, J., Duan, Z., & Watanabe, S. (2024). Generating data with text-to-speech and large-language models for conversational speech recognition.

Proposed Approach

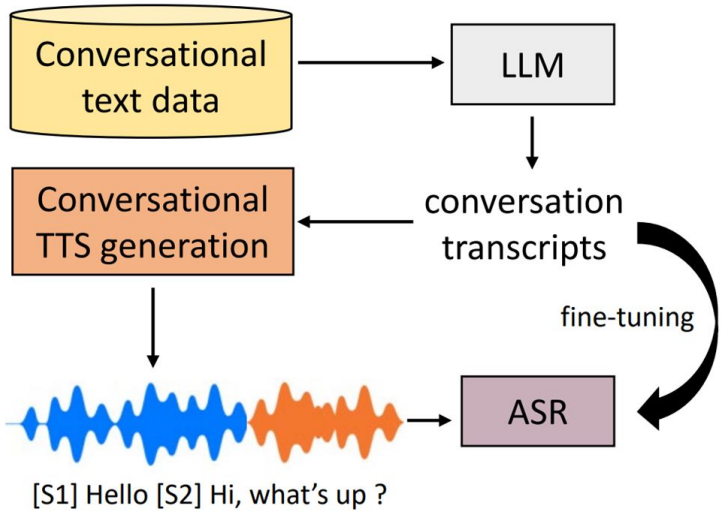
1. Train a TTS system to generate multi-speaker speech.
2. Use an LLM to augment or generate conversational speech transcripts.
   - Should be reasonably effective as they are trained on “conversational text”
3. Generate multi-speaker TTS data for the target domain



Cornell, S., Darefsky, J., Duan, Z., & Watanabe, S. (2024). Generating data with text-to-speech and large-language models for conversational speech recognition.

Proposed Approach

1. Train a TTS system to generate multi-speaker speech.
2. Use an LLM to augment or generate conversational speech transcripts.
   - Should be reasonably effective as they are trained on “conversational text”
3. Generate multi-speaker TTS data for the target domain
4. Fine-tune ASR (e.g. Whisper) on the target domain to do multi-speaker recognition

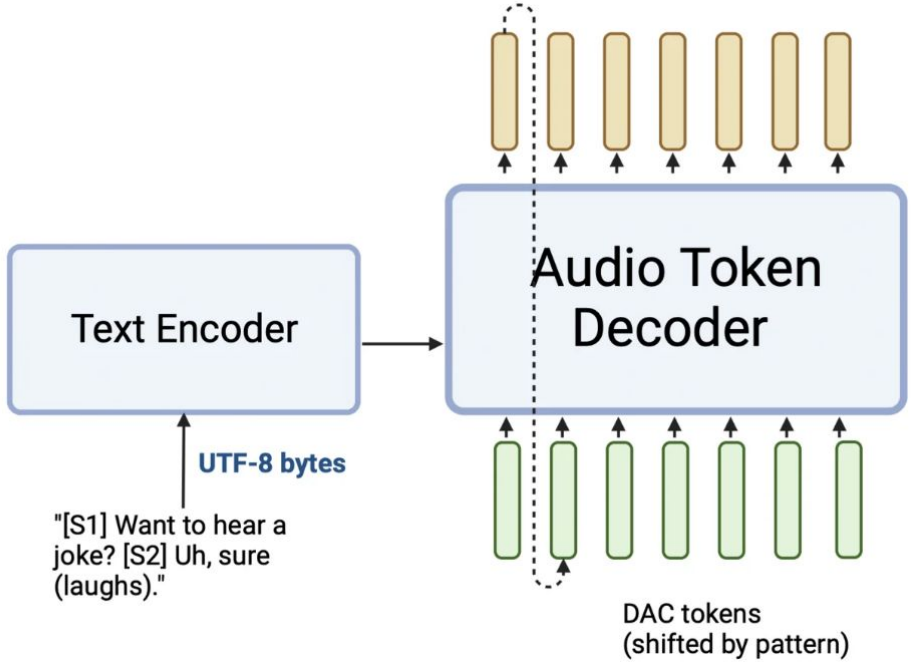


Cornell, S., Darefsky, J., Duan, Z., & Watanabe, S. (2024). Generating data with text-to-speech and large-language models for conversational speech recognition.

Parakeet Conversational TTS model

TTS model that supports the generation of
2-speakers conversations of 20 seconds (@44kHz)

- Encoder-decoder model based on DAC neural codec representation.
- DIA [1] is an open source reproduction of Parakeet



Parakeet: https://jordandarefsky.com/blog/2024/parakeet/

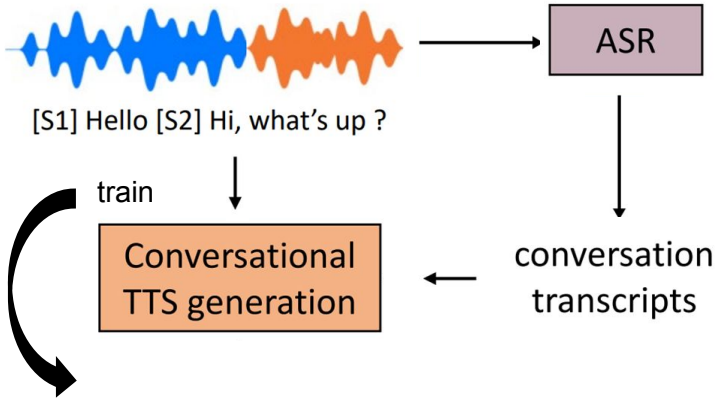
[1] https://github.com/nari-labs/dia

Parakeet Conversational TTS model

Trained on 60k hours of **Spotify Podcasts data [1]**.

- We used Whisper large v2 to pseudo label this corpus.
  - Whisper was fine-tuned to perform serialized output training on **200 utterances** which were manually annotated in this way:

[S1]: blah blah [S2]: yeah



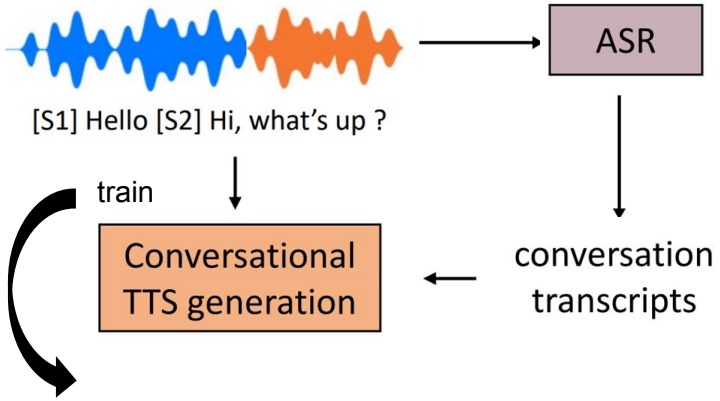
[1] Clifton, Ann, et al. "The spotify podcast dataset." *arXiv preprint arXiv:2004.04270* (2020).

Parakeet Conversational TTS model

Trained on 60k hours of **Spotify Podcasts data [1]**.

- We used Whisper large v2 to pseudo label this corpus.
  - Whisper was fine-tuned to perform serialized output training on **200 utterances** which were manually annotated in this way:

[S1]: blah blah [S2]: yeah



Very data efficient because it probably matches Whisper training data.

[1] Clifton, Ann, et al. "The spotify podcast dataset." *arXiv preprint arXiv:2004.04270* (2020).

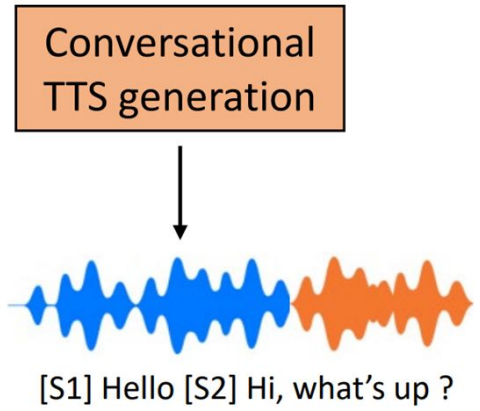
Fully Conversational TTS !

Pros ✅:

- Semantic coherence
  - We have full control the content
- **Realistic turn taking and multi-speaker interaction**

Cons ❌:

- Slower to run than naïve simulation

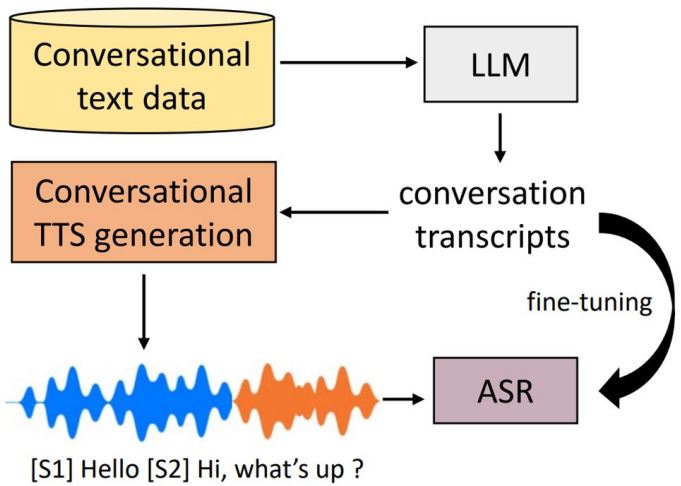


Example of a Generated Conversation (no 🍒 picked)

LLM generated transcript: [S1] I'm like, "What's going on?" [S2] Yeah, like, we're sitting there, and you're like, "What's going on?" And I'm like, "I don't know, man. I think we're just...I don't know. Like, what are we doing, man?" [S1] Yeah, like, we're just...we're just sitting there, man

Demo

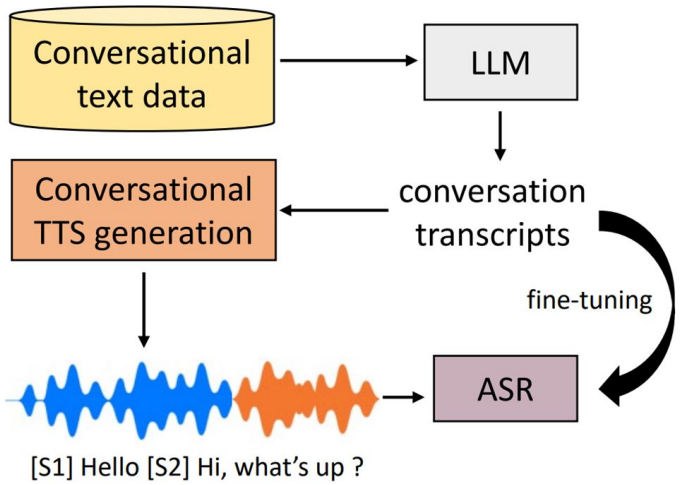
Experimental Results: Fisher



Adaptation Data	amount (hours)	cpWER (%)
-	0	44.94
Fisher	1960	13.76
Fisher	80	15.43
NeMo MSS	80	34.37
xTTS (Fisher)	80	24.88
xTTS (LLM_{rnd})	80	34.65
Parakeet (Fisher)	80	21.44
Parakeet (LLM_{rnd})	80	20.41
Parakeet (LLM_{rnd})	160	19.93

Experimental Results: Fisher

Topline: Fisher in-domain data

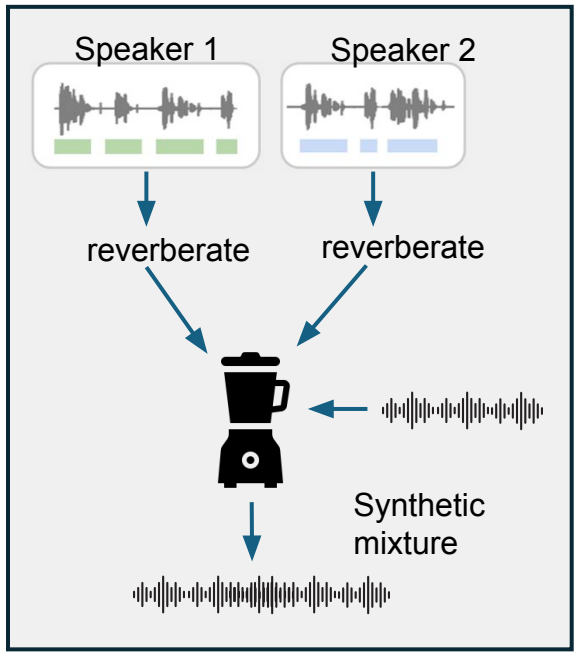


Adaptation Data	amount (hours)	cpWER (%)
-	0	44.94
Fisher	1960	13.76
Fisher	80	15.43
NeMo MSS	80	34.37
xTTS (Fisher)	80	24.88
xTTS (LLM_{rnd})	80	34.65
Parakeet (Fisher)	80	21.44
Parakeet (LLM_{rnd})	80	20.41
Parakeet (LLM_{rnd})	160	19.93

Experimental Results: Fisher

Topline: Fisher in-domain data
Baselines:

- NeMo Multi-speaker simulator

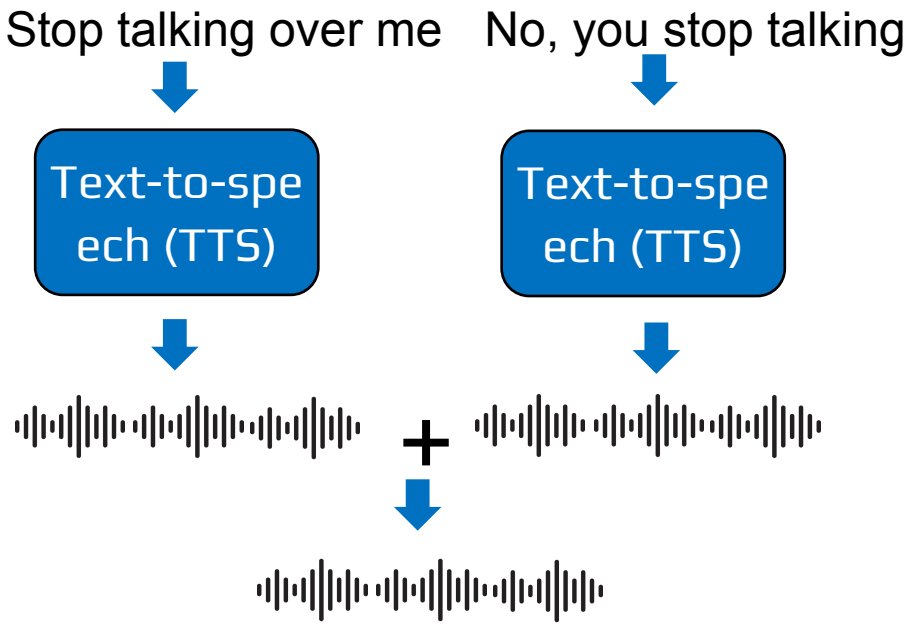


Adaptation Data	amount (hours)	cpWER (%)
-	0	44.94
Fisher	1960	13.76
Fisher	80	15.43
NeMo MSS	80	34.37
xTTS (Fisher)	80	24.88
xTTS (LLM_{rnd})	80	34.65
Parakeet (Fisher)	80	21.44
Parakeet (LLM_{rnd})	80	20.41
Parakeet (LLM_{rnd})	160	19.93

Experimental Results: Fisher

Topline: Fisher in-domain data
Baselines:

- NeMo Multi-speaker simulator
- xTTS (disjoint TTS simulation for each speaker)



Adaptation Data	amount (hours)	cpWER (%)
-	0	44.94
Fisher	1960	13.76
Fisher	80	15.43
NeMo MSS	80	34.37
xTTS (Fisher)	80	24.88
xTTS (LLM_{rnd})	80	34.65
Parakeet (Fisher)	80	21.44
Parakeet (LLM_{rnd})	80	20.41
Parakeet (LLM_{rnd})	160	19.93

Experimental Results: Fisher

Topline: Fisher in-domain data
Baselines:

- NeMo Multi-speaker simulator
- xTTS (disjoint TTS simulation for each speaker)

LLM-generated vs original transcripts

Adaptation Data	amount (hours)	cpWER (%)
-	0	44.94
Fisher	1960	13.76
Fisher	80	15.43
NeMo MSS	80	34.37
xTTS (Fisher)	80	24.88
xTTS (LLM_{rnd})	80	34.65
Parakeet (Fisher)	80	21.44
Parakeet (LLM_{rnd})	80	20.41
Parakeet (LLM_{rnd})	160	19.93

Experimental Results: Fisher

No degradation on the Parakeet model

Adaptation Data	amount (hours)	cpWER (%)
-	0	44.94
Fisher	1960	13.76
Fisher	80	15.43
NeMo MSS	80	34.37
xTTS (Fisher)	80	24.88
xTTS (LLM_{rnd})	80	34.65
Parakeet (Fisher)	80	21.44
Parakeet (LLM_{rnd})	80	20.41
Parakeet (LLM_{rnd})	160	19.93

Experimental Results: Fisher

Significantly surpasses baseline methods

Adaptation Data	amount (hours)	cpWER (%)
-	0	44.94
Fisher	1960	13.76
Fisher	80	15.43
NeMo MSS	80	34.37
xTTS (Fisher)	80	24.88
xTTS (LLM_{rnd})	80	34.65
Parakeet (Fisher)	80	21.44
Parakeet (LLM_{rnd})	80	20.41
Parakeet (LLM_{rnd})	160	19.93

Experimental Results: Fisher

But diminishing returns….

Adaptation Data	amount (hours)	cpWER (%)
-	0	44.94
Fisher	1960	13.76
Fisher	80	15.43
NeMo MSS	80	34.37
xTTS (Fisher)	80	24.88
xTTS (LLM_{rnd})	80	34.65
Parakeet (Fisher)	80	21.44
Parakeet (LLM_{rnd})	80	20.41
Parakeet (LLM_{rnd})	160	19.93

Experimental Results: Mixer 6

Topline: Mixer 6 in-domain data

Adaptation Data	amount (hours)	cpWER (%)
-	0	43.67
Mixer6	2.30	20.36
Fisher	1960	20.83
Fisher	80	22.12
NeMo MSS	80	36.71
xTTS (Mixer6)	2.30	25.99
xTTS (LLM_{rnd})	80	35.65
Parakeet (Mixer6)	2.30	23.52
Parakeet (LLM_{rnd})	2.30	23.70
Parakeet (LLM_{rnd})	80	21.25

Experimental Results: Mixer 6

Synthetic data **as effective as using real-world conversational data** from another domain.

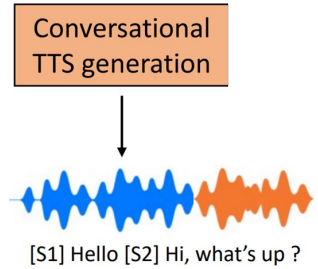
- But we have full control on the content !

Adaptation Data	amount (hours)	cpWER (%)
-	0	43.67
Mixer6	2.30	20.36
Fisher	1960	20.83
Fisher	80	22.12
NeMo MSS	80	36.71
xTTS (Mixer6)	2.30	25.99
xTTS (LLM_{rnd})	80	35.65
Parakeet (Mixer6)	2.30	23.52
Parakeet (LLM_{rnd})	2.30	23.70
Parakeet (LLM_{rnd})	80	21.25

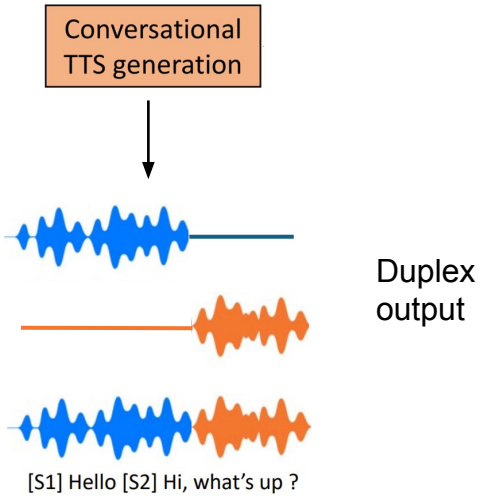
Limitations

- **Only suitable for ASR training** (inverse task of TTS)
  - We need duplex for diarization, speech separation, spoken dialog systems…

What we got:



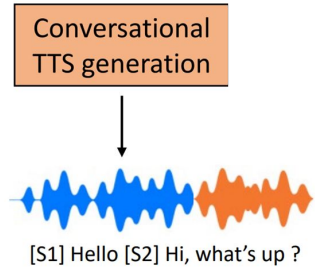
What we really wanted:



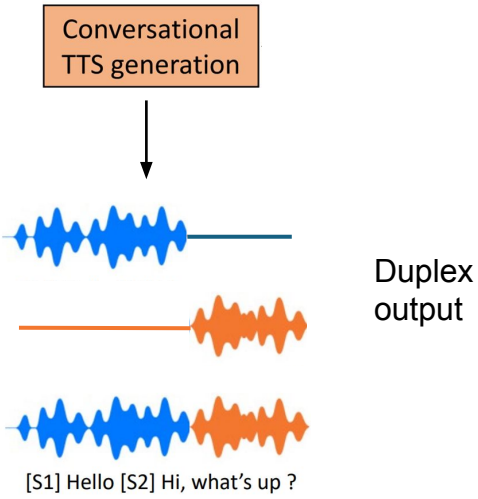
Limitations

- **Only suitable for ASR training** (inverse task of TTS)
  - We need duplex for diarization, speech separation, spoken dialog systems…
- Trained on Podcast data -> clean speech only

What we got:

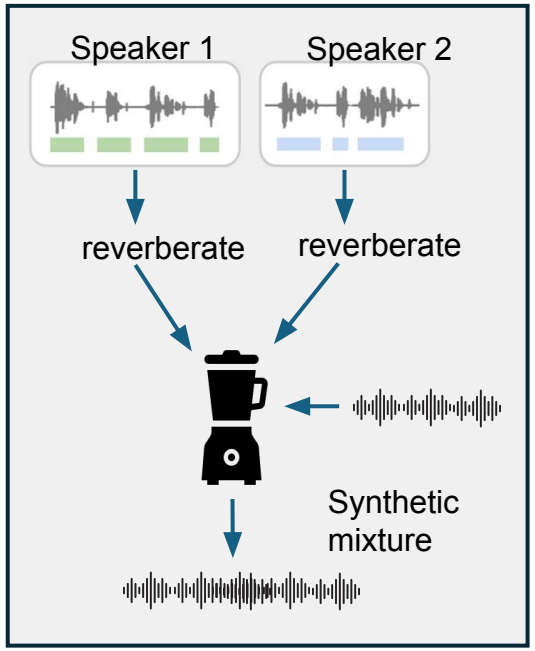


What we really wanted:

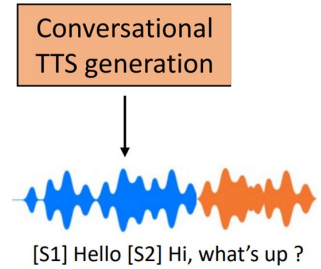


Limitations

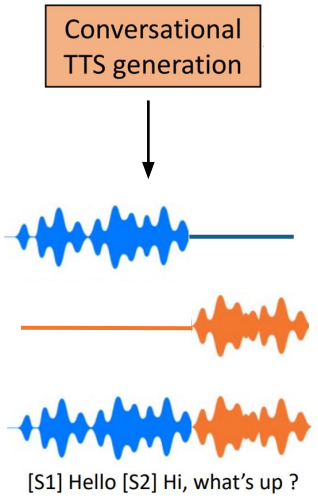
- **Only suitable for ASR training** (inverse task of TTS)
  - We need duplex for diarization, speech separation, spoken dialog systems…
- Trained on Podcast data -> clean speech only



What we got:



What we really wanted:

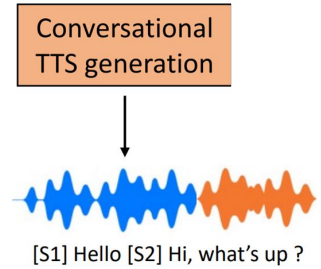


Duplex output

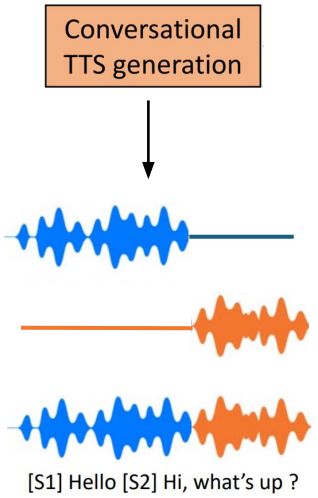
Limitations

- **Only suitable for ASR training** (inverse task of TTS)
  - We need duplex for diarization, speech separation, spoken dialog systems…
- Trained on Podcast data -> clean speech only
- What about **multi-channel** ?

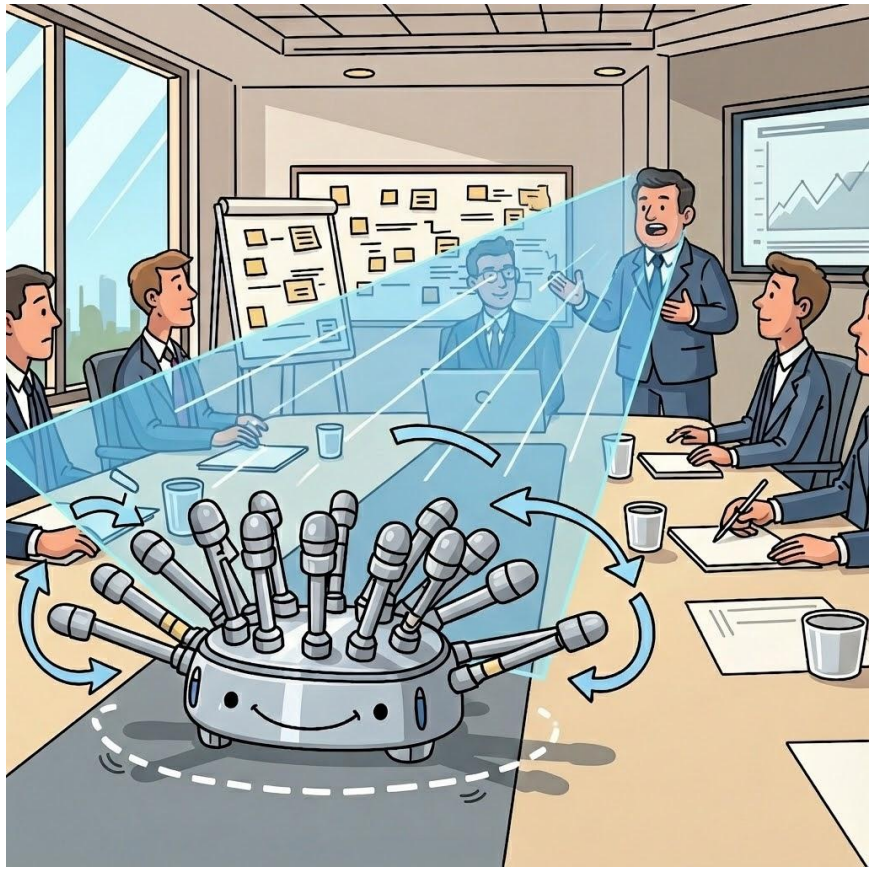
What we got:

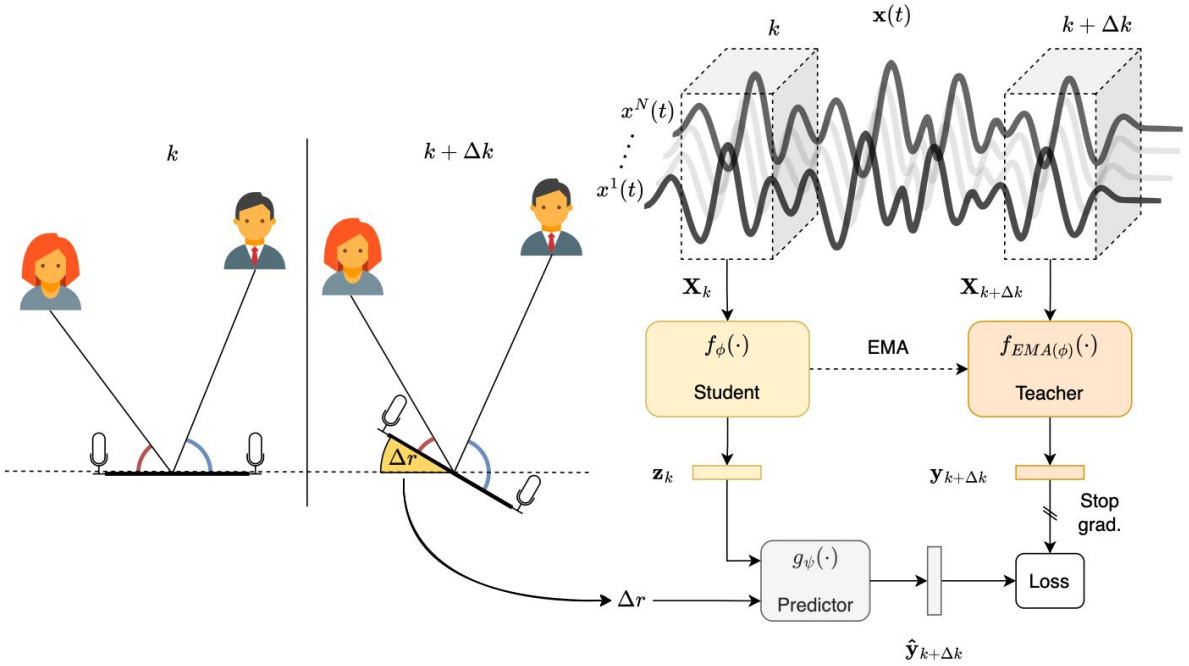


What we really wanted:



Duplex output





The Cocktail Party Isn't Over: Future & Current Directions

Data is scarce, so let’s learn on-device !

In-domain data for real-world multi-speaker speech is scarce and difficult to acquire (privacy):

- **Especially multi-channel**

This is a significant problems for e.g. the development of better hearing-aid.

Data is scarce, so let’s learn on-device !

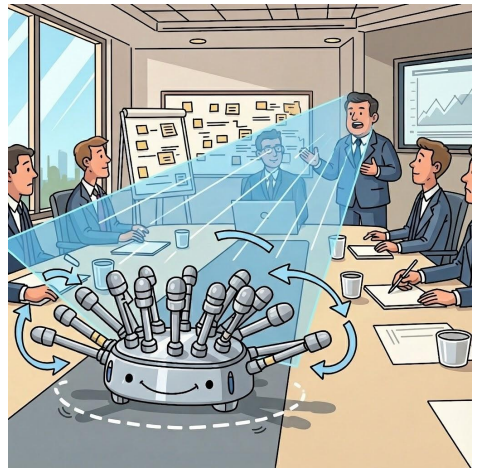
In-domain data for real-world multi-speaker speech is scarce and difficult to acquire (privacy):

- **Especially multi-channel**

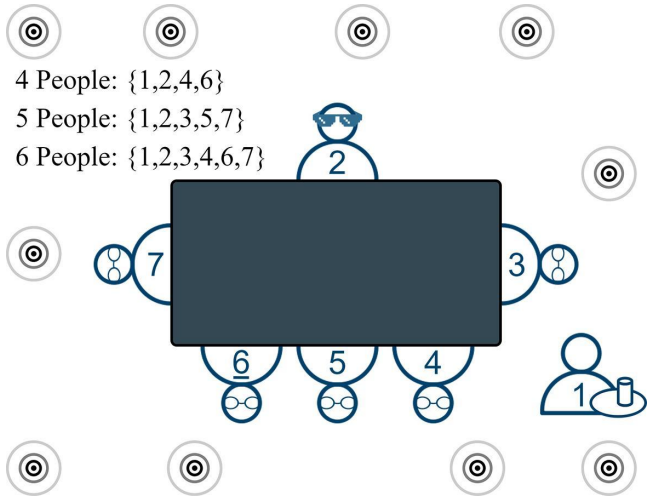
This is a significant problems for e.g. the development of better hearing-aid.

What however if we design a training framework that can learn **on device** ?

- Self supervised learning (SSL)
- **Light enough** that can be performed on-device

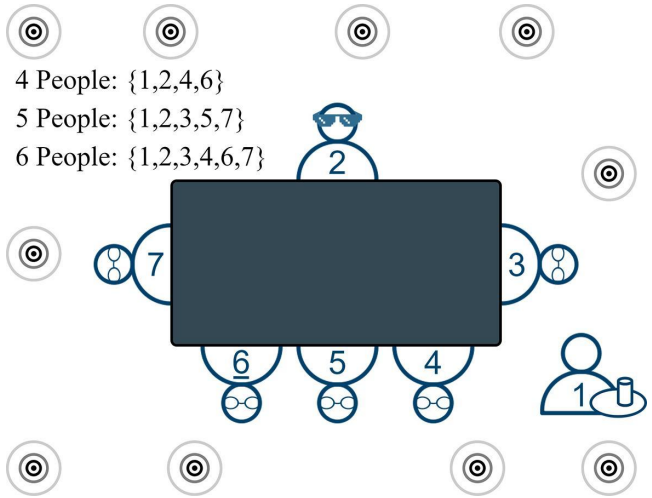


Spatial Joint-Embedding Predictive Architecture (S-JEPA)

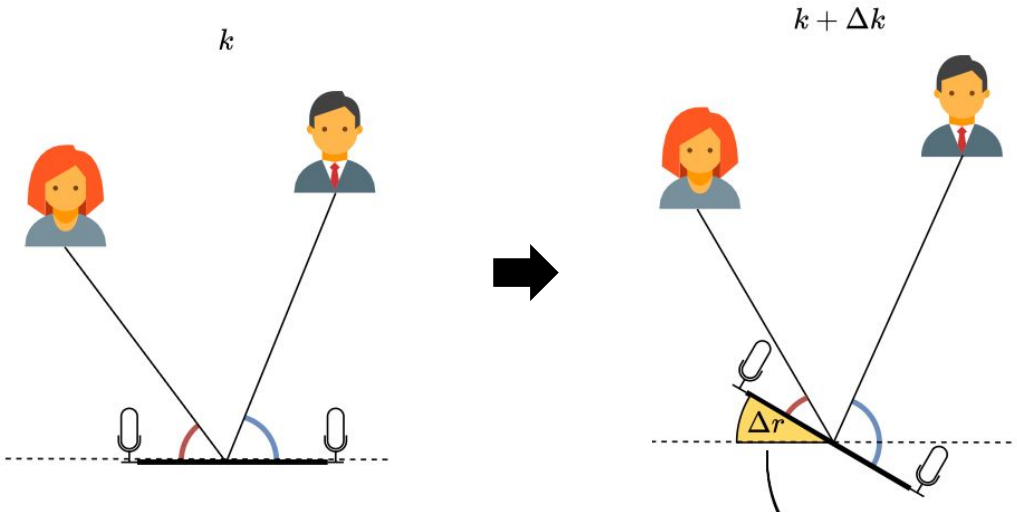


From Donley, Jacob, et al. "EasyCom: An augmented reality dataset to support algorithms for easy communication in noisy environments." arXiv preprint arXiv:2107.04174 (2021).

Spatial Joint-Embedding Predictive Architecture (S-JEPA)



From Donley, Jacob, et al. "EasyCom: An augmented reality dataset to support algorithms for easy communication in noisy environments." arXiv preprint arXiv:2107.04174 (2021).

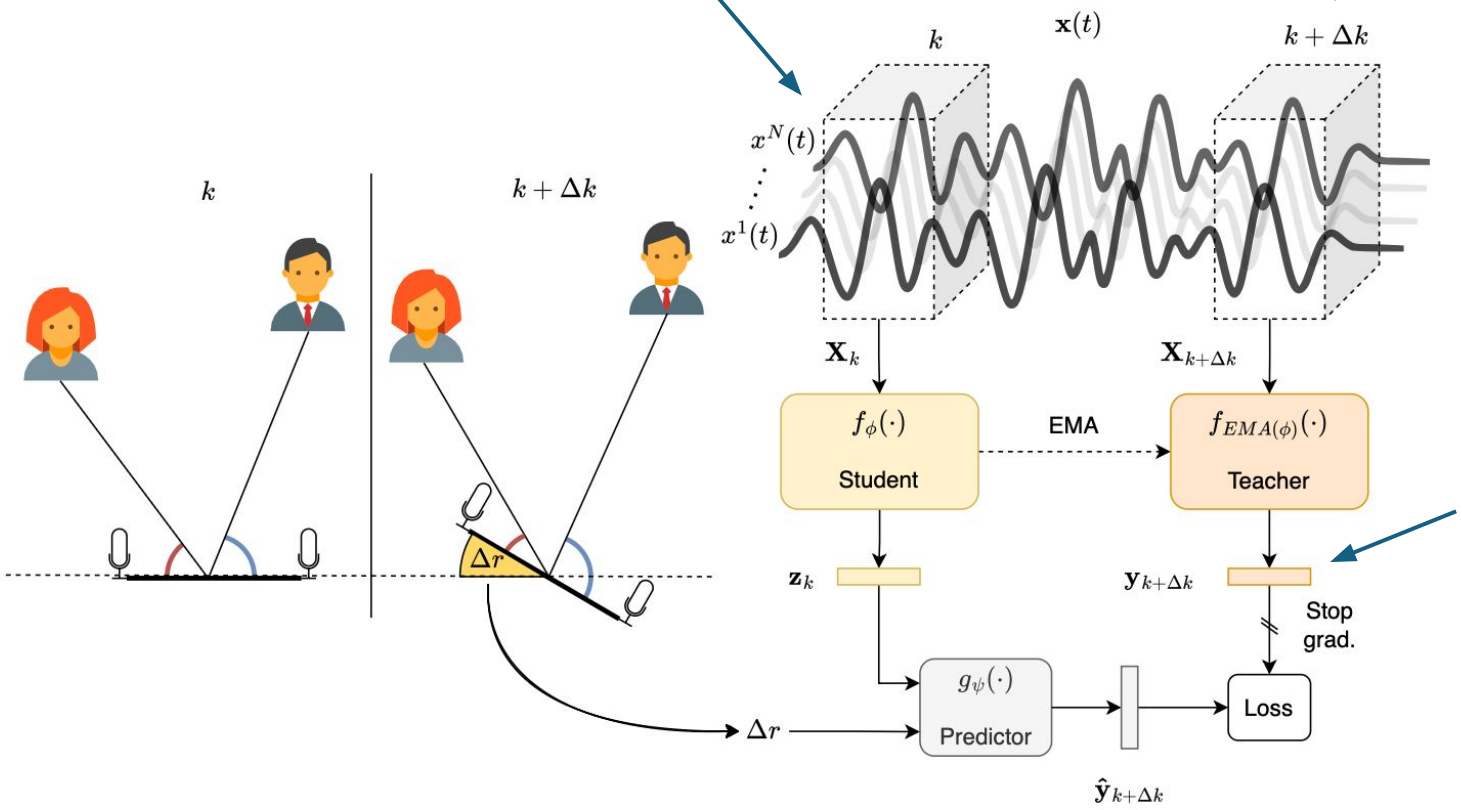


Most devices have inertial measurement unit (IMU) sensor (e.g. Airpods)

Spatial Joint-Embedding Predictive Architecture (S-JEPA)

Spatial features
e.g. interchannel phase difference

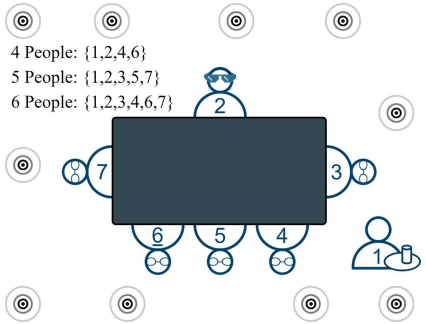
Spatial features after the array moved
- Delta_k ~ 100ms

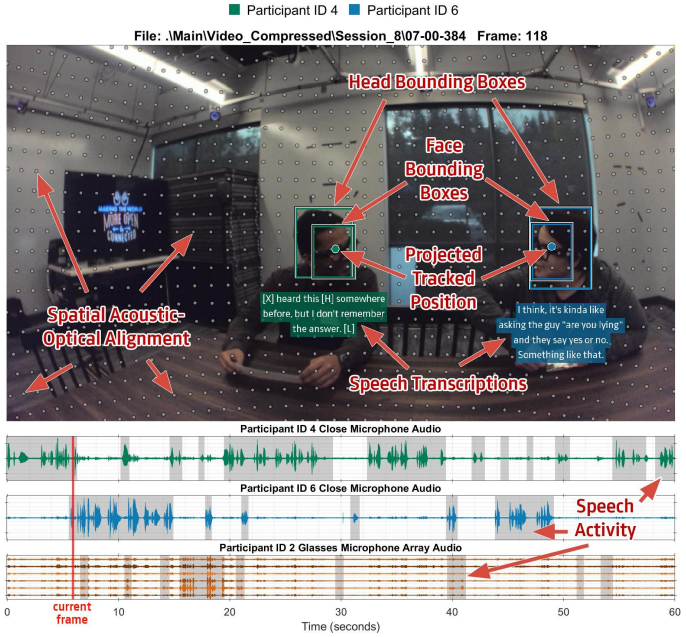


Representation from the teacher after the array moved

Preliminary Results

Scenario: EasyCom

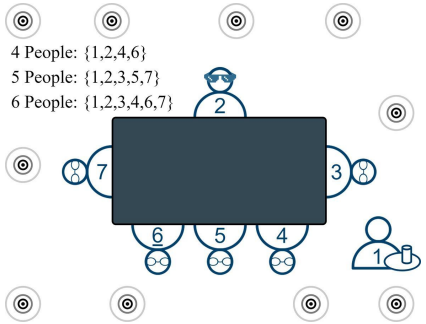


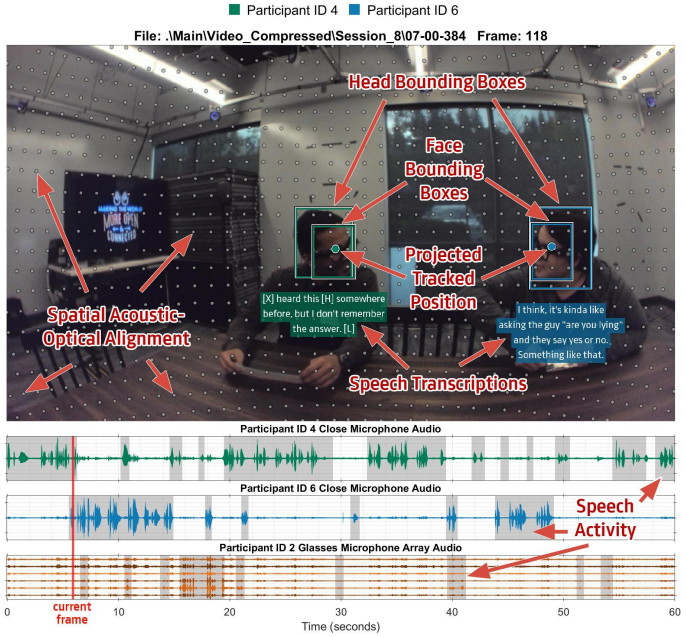


Yang, Yufeng, et al. "M-best-rq: A multi-channel speech foundation model for smart glasses." ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2025.

Preliminary Results

Scenario: EasyCom





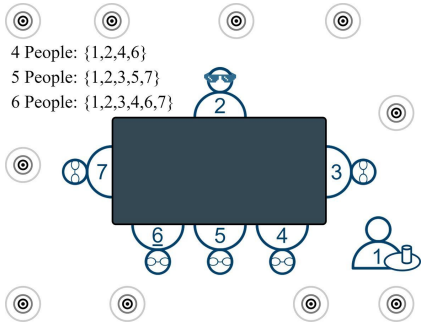
- Spherical Acoustic Sound Localization (S-ASL)
  - Mean absolute error (MAE)
- Glass-wearer Voice activity detection (W-VAD)
  - Mean average precision (mAP)

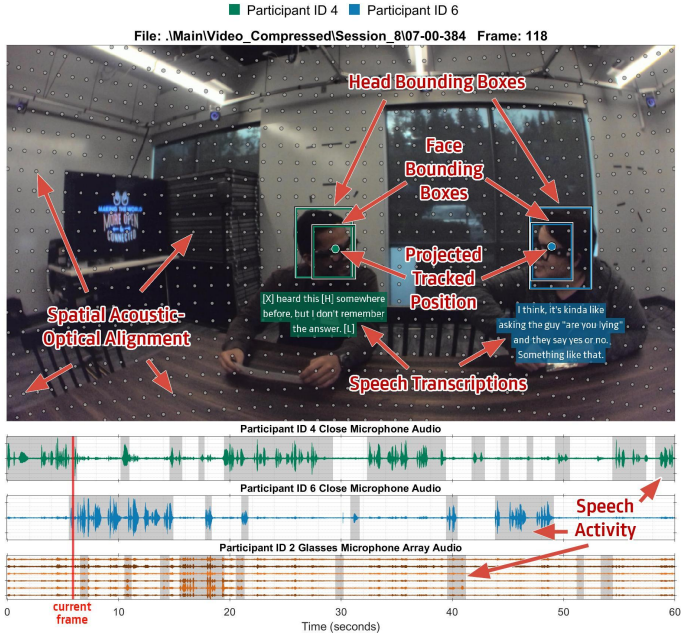
Model	AV	Size (M)	S-ASL			W-VAD
			$MAE_{p\to g}$	$MAE_{g\to p}$	mMAE	mAP (%)
Yang et al. (2025) M-BEST-RQ (weighted comb)	×	99.7	24.0	4.8	14.4	87.7
+ full fine-tune	×	99.7	**4.9**	7.0	<u>6.0</u>	90.7
Supervised	×	0.2	6.7	6.4	6.5	93.2
S-JEPA	×	0.2	9.7	10.8	10.2	91.1
+ full fine-tune	×	0.2	<u>5.7</u>	**4.4**	**5.1**	**94.6**

Yang, Yufeng, et al. "M-best-rq: A multi-channel speech foundation model for smart glasses." ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2025.

Preliminary Results

Scenario: EasyCom





- Spherical Acoustic Sound Localization (S-ASL)
  - Mean absolute error (MAE)
- Glass-wearer Voice activity detection (W-VAD)
  - Mean average precision (mAP)

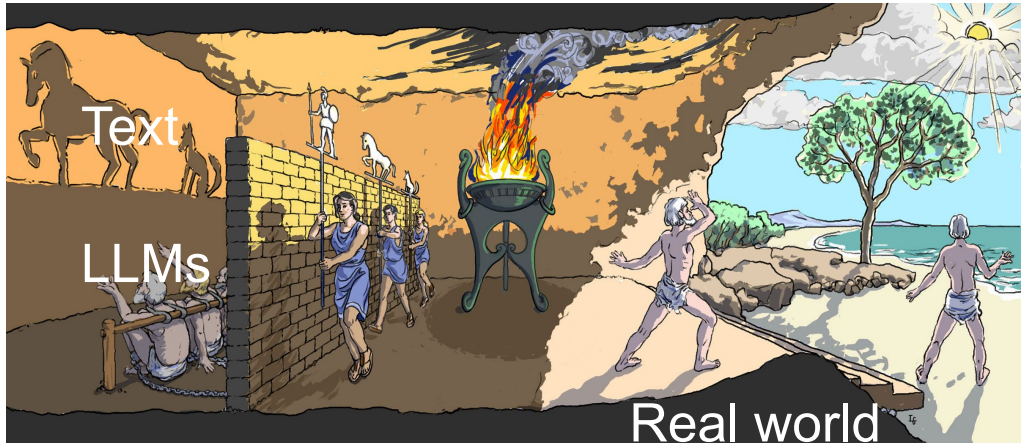
Model	AV	Size (M)	S-ASL			W-VAD
			$MAE_{p\to g}$	$MAE_{g\to p}$	mMAE	mAP (%)
Yang et al. (2025) M-BEST-RQ (weighted comb)	×	99.7	24.0	4.8	14.4	87.7
+ full fine-tune	×	99.7	**4.9**	7.0	<u>6.0</u>	90.7
Supervised	×	0.2	6.7	6.4	6.5	93.2
S-JEPA	×	0.2	9.7	10.8	10.2	91.1
+ full fine-tune	×	0.2	<u>5.7</u>	**4.4**	**5.1**	**94.6**

Yang, Yufeng, et al. "M-best-rq: A multi-channel speech foundation model for smart glasses." ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2025.

Escaping Plato's Cave

We should use more **multi-sensory information for pretraining**

- **Enforce inter multi-sensory constraints/connections**
  - This is especially important for multi-party spontaneous interaction.



E.g. **audio and movement** but it can be audio and video and movement and so on…

SLIDAR Results... Continued

Diarization+Transcription Systems	**DER (%)**	**cpWER (%)**
Transcribe-to-Diarize	28.1	24.9
SLIDAR (proposed)	31.5	24.5
DiariZen + SE-DiCoW [1]	**10.4**	**18.5**

[1] Polok, Alexander, et al. "SE-DiCoW: Self-Enrolled Diarization-Conditioned Whisper." Accepted at ICASSP (2026).

Omni vs Task Specific

Diarization+Transcription Systems	**DER (%)**	**cpWER (%)**
Transcribe-to-Diarize	28.1	24.9
SLIDAR (proposed)	31.5	24.5
DiariZen + SE-DiCoW [1]	**10.4**	**18.5**

Omni LLMs	**DER (%)**	**cpWER (%)**
Gemini 2.5 Pro (SLIDAR-style chunked inference [2])	23.8	34.8
Gemini 3.0 Pro (SLIDAR-style chunked inference [2])	43.0	26.9
LLM-based Systems		
TagSpeech [3] (this month, fully e2e)	24.8	42.5
VibeVoice-ASR [2] (this month, fully e2e)	13.4	28.4

[1] Polok, Alexander, et al. "SE-DiCoW: Self-Enrolled Diarization-Conditioned Whisper." Accepted at ICASSP (2026).

[2] Peng, Zhiliang, et al. "VIBEVOICE-ASR Technical Report." arXiv preprint arXiv:2601.18184 (2026).

[3] Huo, Mingyue, Yiwen Shao, and Yuheng Zhang. "TagSpeech: End-to-End Multi-Speaker ASR and Diarization with Fine-Grained Temporal Grounding." arXiv preprint arXiv:2601.06896 (2026).

Jack of All Trades, Master of None?

We **want e2e for flexibility**
However **we are forced to use chunking (for now ?)**.

- Theoretically it should be better (e.g. paralinguistic should help for summarization.
  - Is it only a lack of data the issue ?

Omni LLMs	**DER (%)**	**cpWER (%)**
Gemini 2.5 Pro (SLIDAR-style chunked inference [2])	23.8	34.8
Gemini 3.0 Pro (SLIDAR-style chunked inference [2])	43.0	26.9
LLM-based Systems		
TagSpeech [3] (this month, fully e2e)	24.8	42.5
VibeVoice-ASR [2] (this month, fully e2e)	13.4	28.4

[2] Peng, Zhiliang, et al. "VIBEVOICE-ASR Technical Report." arXiv preprint arXiv:2601.18184 (2026).
[3] Huo, Mingyue, Yiwen Shao, and Yuheng Zhang. "TagSpeech: End-to-End Multi-Speaker ASR and Diarization with Fine-Grained Temporal Grounding." arXiv preprint arXiv:2601.06896 (2026).

Jack of All Trades, Master of None?

We **want e2e for flexibility**
However **we are forced to use chunking (for now ?)**.

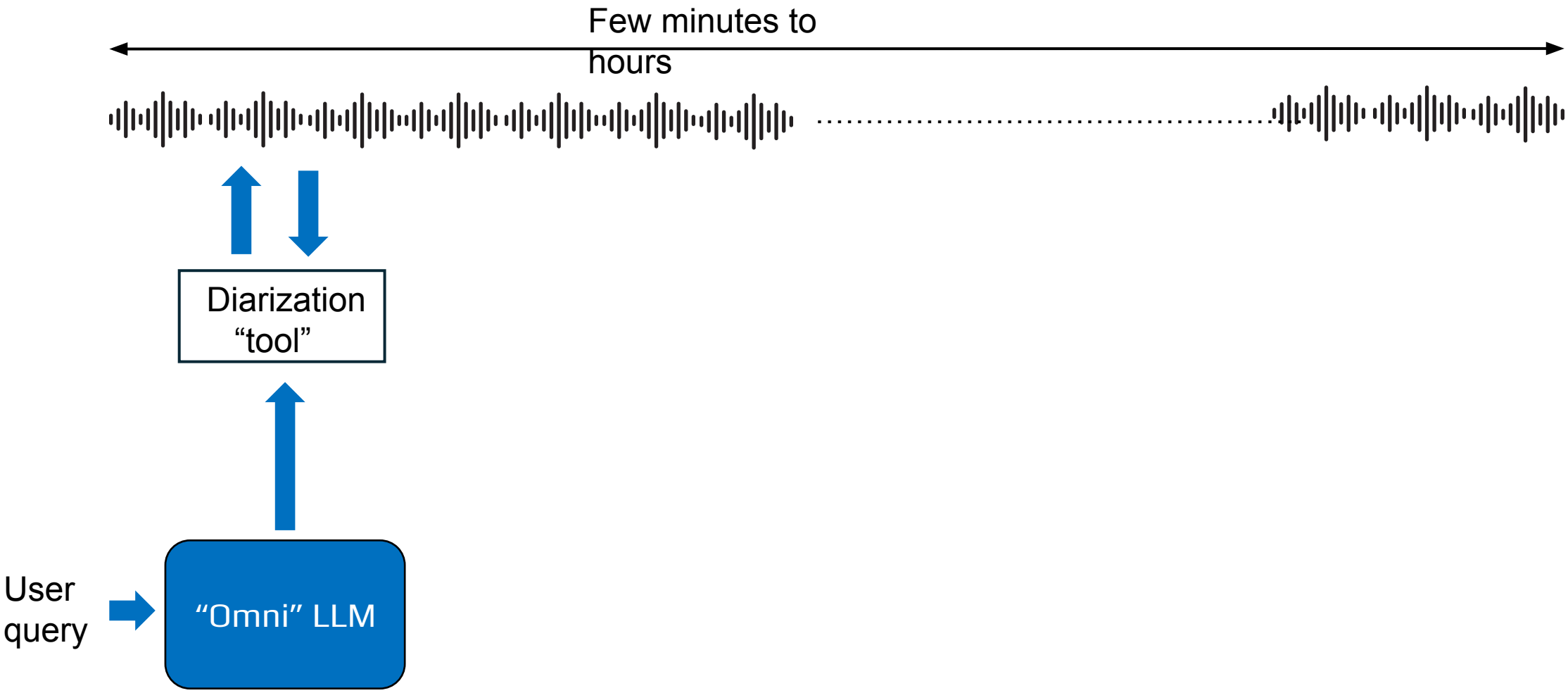
- Theoretically it should be better (e.g. paralinguistic should help for summarization.
  - Is it only a lack of data the issue ?

- **All specialized models (LLM-based) lose their "LLM abilities"**
  - Tension between "generative tasks" and verbatim long-form diarization+transcription ?

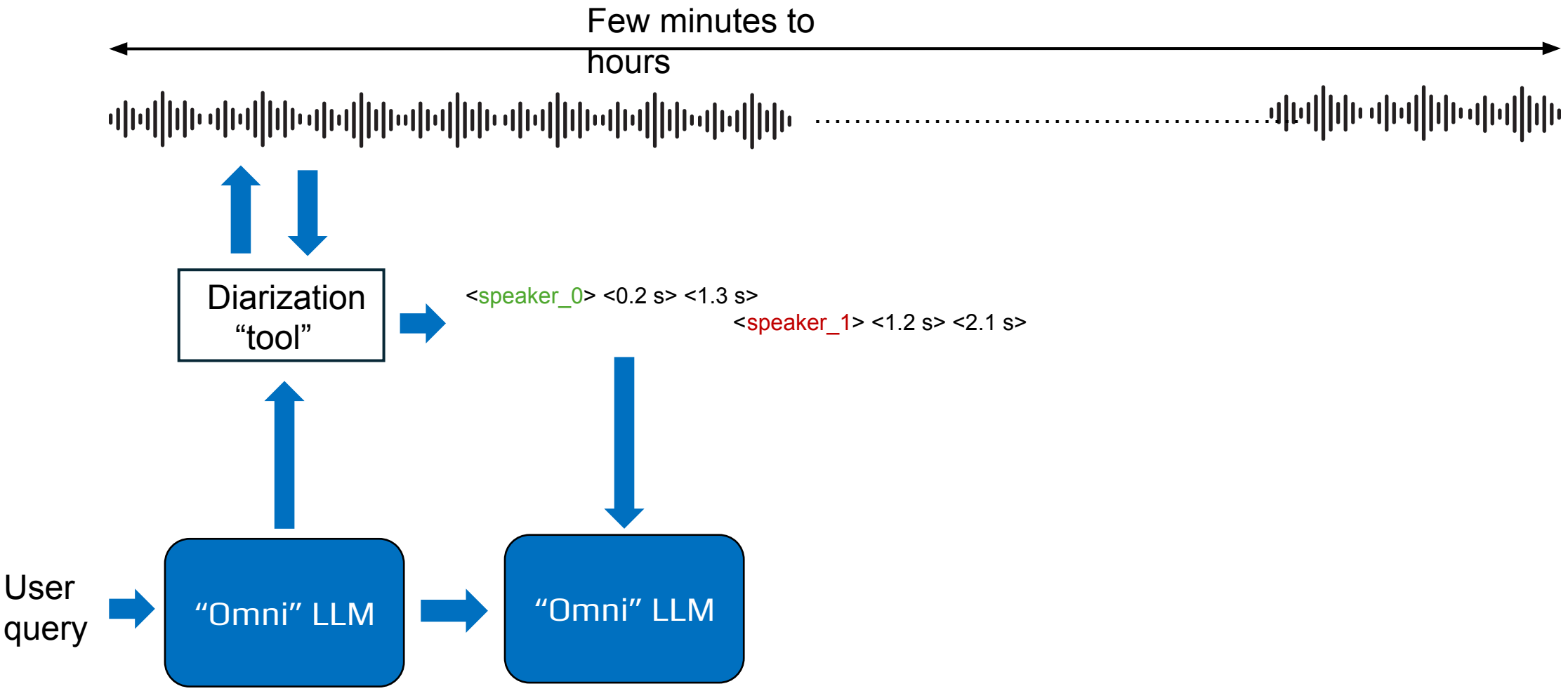
Omni LLMs	**DER (%)**	**cpWER (%)**
Gemini 2.5 Pro (SLIDAR-style chunked inference [2])	23.8	34.8
Gemini 3.0 Pro (SLIDAR-style chunked inference [2])	43.0	26.9
LLM-based Systems		
TagSpeech [3] (this month, fully e2e)	24.8	42.5
VibeVoice-ASR [2] (this month, fully e2e)	13.4	28.4

[2] Peng, Zhiliang, et al. "VIBEVOICE-ASR Technical Report." arXiv preprint arXiv:2601.18184 (2026).
[3] Huo, Mingyue, Yiwen Shao, and Yuheng Zhang. "TagSpeech: End-to-End Multi-Speaker ASR and Diarization with Fine-Grained Temporal Grounding." arXiv preprint arXiv:2601.06896 (2026).

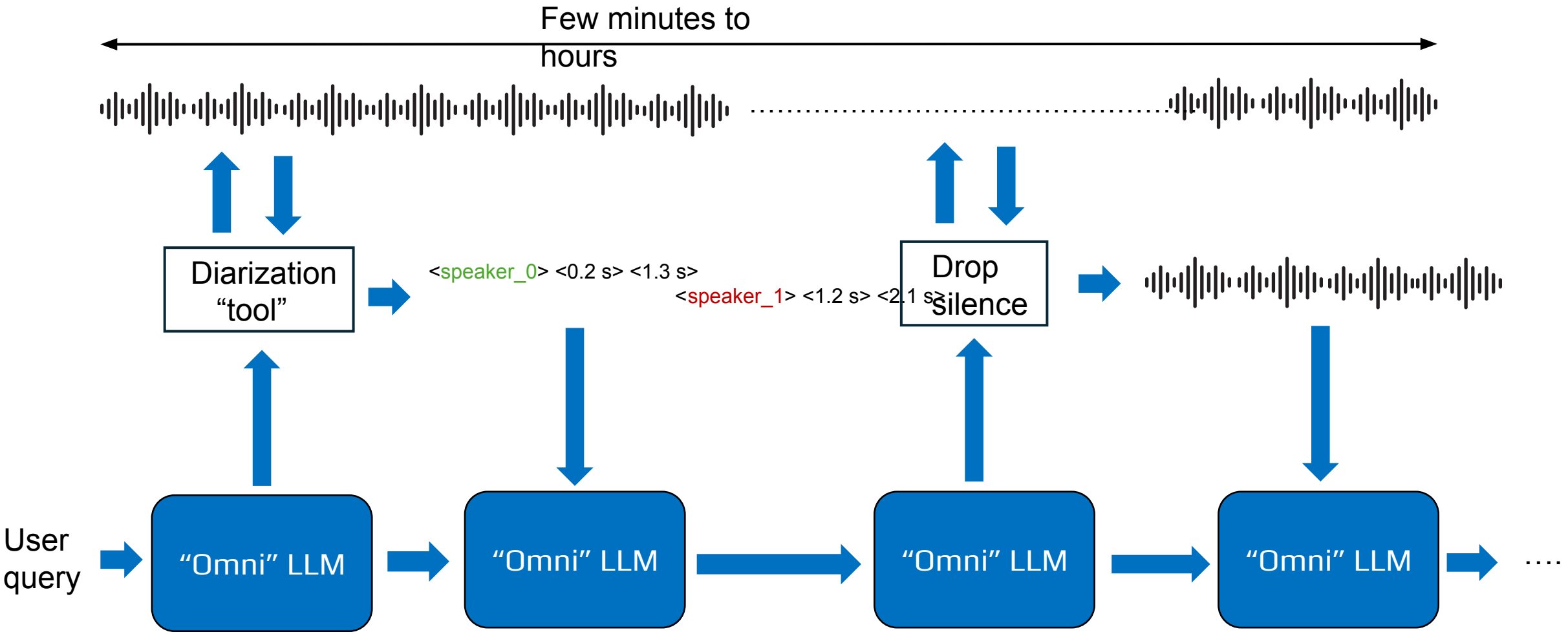
“Agent” for long form audio understanding



“Agent” for long form audio understanding



“Agent” for long form audio understanding



Thank you for this opportunity !
Any questions ?

Email: scornell@andrew.cmu.edu

Link to these slides

What about Speech Separation ?

I also contributed to SotA speech separation with SepFormer first and then TF-GridNet.

- I extended it to the online scenario for target speaker extraction with **5 ms latency** for the hearing-aid enhancement Clarity challenge 2 and won the first place.

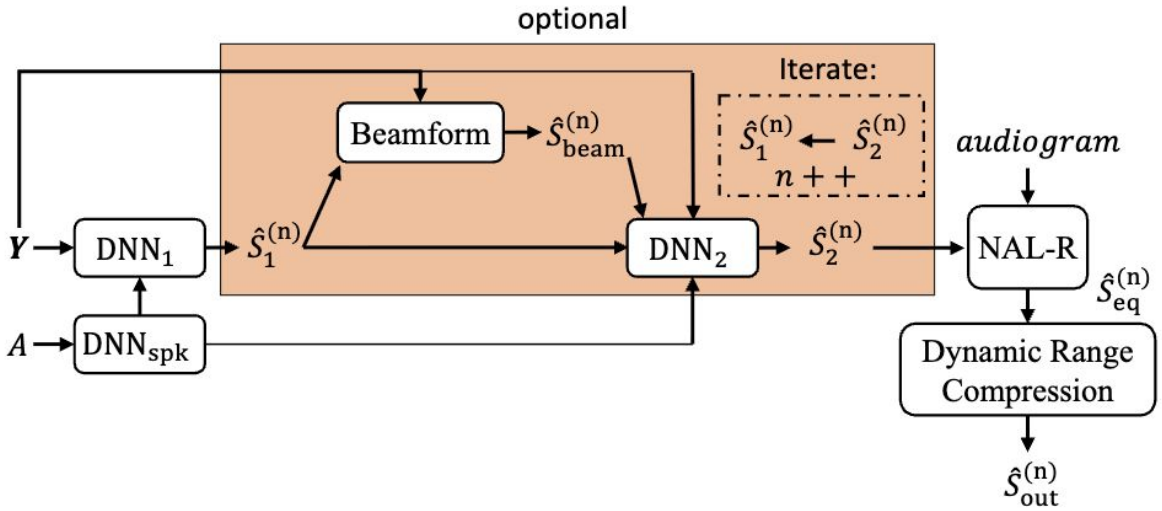


Fig. 1: Overview of proposed iNeuBe-X framework. We employ a causal multi-channel Wiener filter beamformer between DNN_1 and DNN_2.

Surpasses human-performance on extremely difficult noisy/reverberant conditions.

See demo: https://popcornell.github.io/WAVLab_CEC2_demo/

Also here: “Super-human Results”

I also contributed to SotA speech separation with SepFormer first and then TF-GridNet.

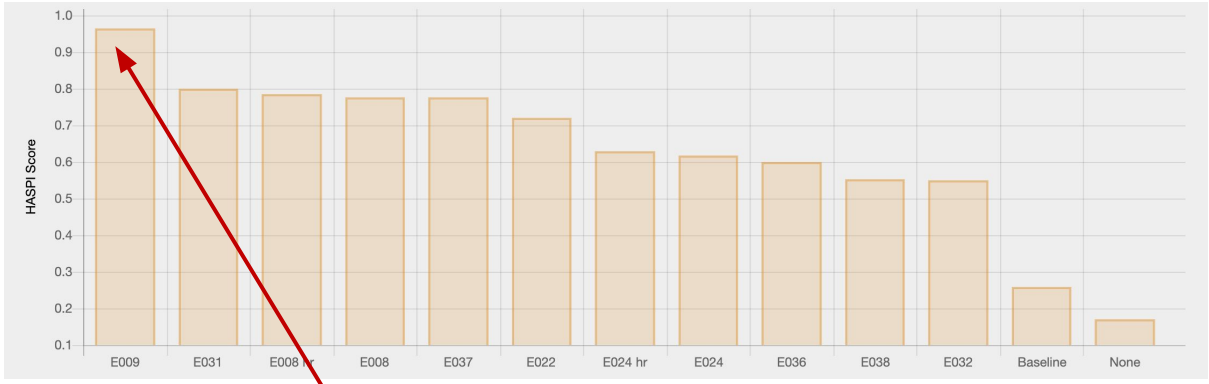
- I extended it to the online scenario for target speaker extraction with **5 ms latency** for the hearing-aid enhancement Clarity challenge and won the first place.

Mixture

Enhanced

Surpasses human-performance on extremely difficult noisy/reverberant conditions.

Demo available:
https://popcornell.github.io/WAVLab_CEC2_demo/

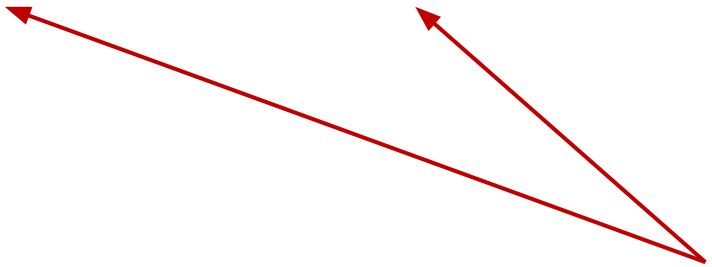


Our system: almost perfect HASPI (intelligibility metric for hearing-aid)

Synthetic to Real Generalization Gap

This «super-human» hearing-aid enhancement» does not generalize to the real world.
Results on synthetic vs. real-world data evaluation on the subsequent Clarity ICASSP Grand Challenge:

Approaches	Synth		Real	
	HASPI	HASQI	HASPI	HASQI
mixture CH1 Left	0.26	0.13	0.18	0.12
ours (submitted)	0.80	0.41	0.29	0.11



Significant drop in performance on real-data

- Lots of audible artifacts and distortions

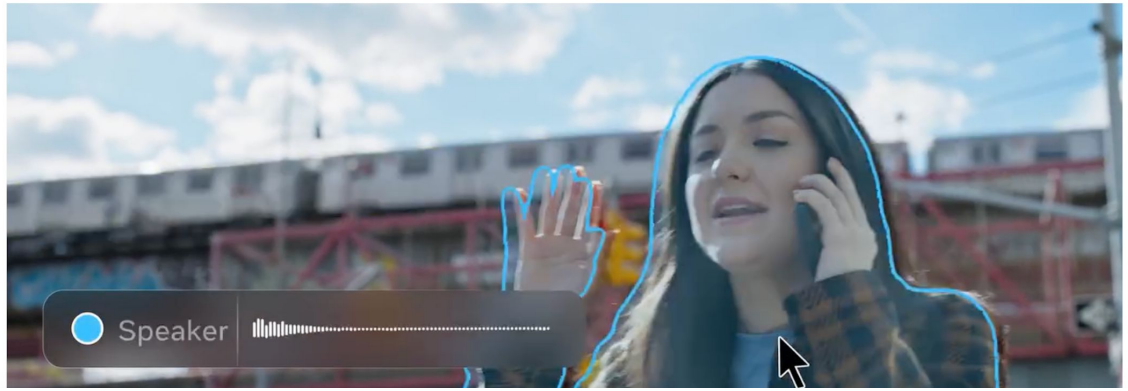
WER we are ? Current State-of-the-Art

Meta | Meta AI | AI Research | The Latest | About | Get Llama | Try Meta AI

Computer Vision

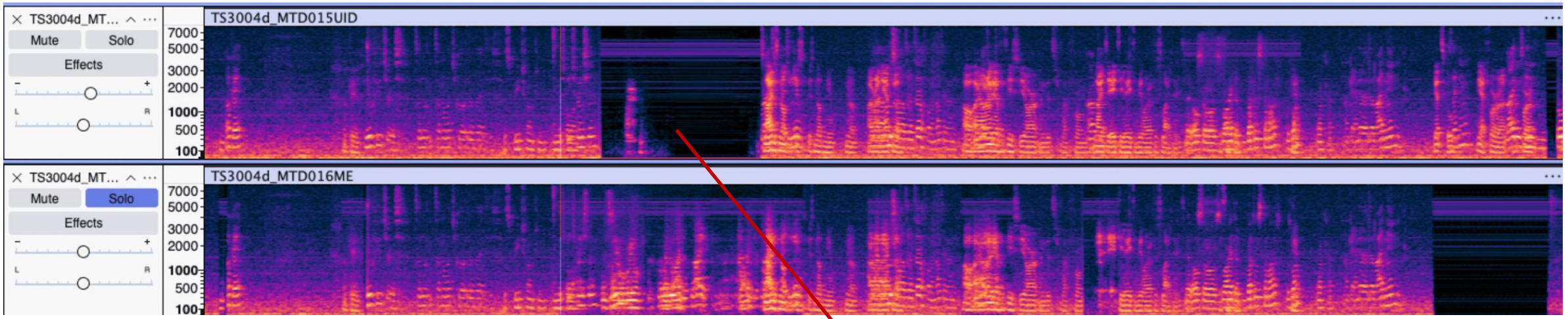
Introducing SAM Audio: The First Unified Multimodal Model for Audio Separation

December 16, 2025 • 11 minute read



Poor Generalization ?

And BTW, the same for Meta SAM-Audio when you apply it to real-world far-field meeting data (not even particularly noisy)



Should be speech but instead it is some weird metallic noise.