Beyond End-to-End ASR: Integrating Long-Context Acoustic and Linguistic Insights

Tae Jin Park, Huck Yang, Kyu J. Han, Shinji Watanabe

Session 15:30 to 18:30
Aug 17th 2025

Scan for References

Speakers

Prof. Shinji Watanabe

Carnegie Mellon University (CMU)

Dr. Taejin Park

NVIDIA NeMo SpeechAI

Dr. Huck Yang

NVIDIA Research

Dr. Kyu J. Han

Oracle Cloud Infrastructure (OCI)

Introduction

Shinji Watanabe

Abstract:

This tutorial explores the advancements and challenges in long-context automatic speech recognition (ASR), emphasizing its critical role in improving fairness, inclusivity, and performance across diverse linguistic groups.

While modern ASR systems excel in data-rich languages such as English, they often underperform for low-resource languages, accented speech, and underrepresented communities due to limited context modeling.

(Speech Benchmarks) The tutorial also examines benchmarks such as SLUE, Speech QA, and Dynamic SUPERB, which assess ASR performance in noisy environments, spoken language understanding, and long-context comprehension.

(Evaluations for modern speech topics) This tutorial introduces robust evaluation pipelines for long-form ASR, leveraging datasets such as CHiME and LibriHeavy, and discusses metrics such as multi-talker diarization, semantic evaluation, and retrieval-augmented generation (RAG) techniques.

(Acoustic and Semantic Context modeling) We delve into acoustic and semantic context modeling, highlighting innovations in multi-speaker processing, speech-LLM integration, and RAG-based error correction.

(Long Context in ASR) By integrating long-context processing and large language models (LLMs), ASR systems can better handle disfluencies, code-switching, and speaker variability, reducing biases and improving accessibility for marginalized communities.

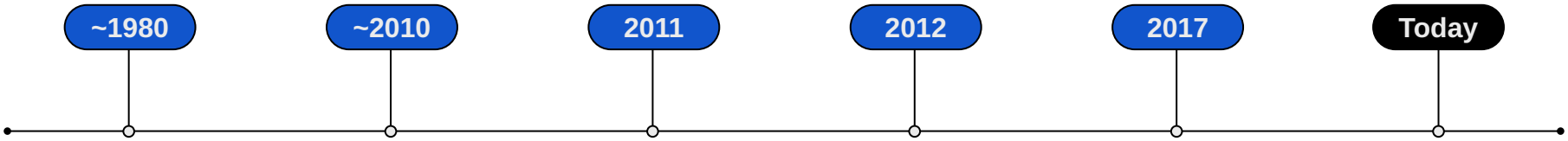
(Summary of Summary) This tutorial underscores the importance of long-context ASR in advancing speech technology toward fairer, more inclusive systems that serve all users equitably.

Table of Contents

Download Slides

Introduction (10 mins) 15:30-15:40	
Shinji (30 min) 15:40-16:10	Speech-to-Text Benchmark
Taejin (40 min) 16:10-16:50	Leveraging Long Acoustic Context
Recess (10 min) 16:50-17:00	
Huck (40 min) 17:00-17:40	Semantic Context and Speech-Language Modeling
Kyu (30 min) 17:40-18:10	Contextual Biasing and Methods Leveraging Longer Semantic Context for Speech Systems
Closing Remark (10 min) 18:10-18:20	
Q&A Session (10 min) 18:20-18:30	

A short history of speech-to-text systems



DTW

Template matching
Isolated word recognition
Speaker-dependent system

GMM-HMM

Statistical method
Maximum likelihood EM algorithm
Speaker-independent system
Combination of acoustic, lexicon, and language model WFST

DNN-HMM

Feed-forward → CNN → RNN
Longer context modeling
GPU training

RNNT

End-to-End system
No LM (or shallow fusion, etc.)
No lexicon
Streaming

Transformer

Standard architectectural in modern AI systems
Long-context modeling
Auto-regressive decoder, including prompting

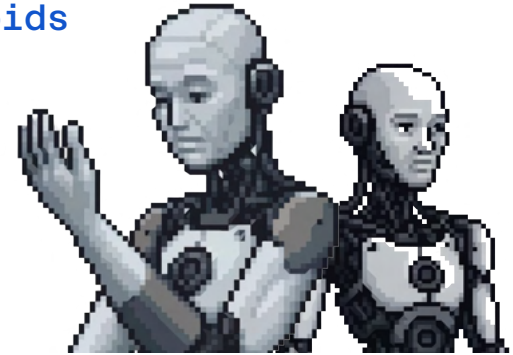
STT for Agentic AI

Multitask capabilities, e.g., ASR + X (ST, QA, etc.)
Strong contextual modeling via prompts

Modern day Speech-to-Text Applications

- Human-computer interface

- Humanoids



- Voice Assistants

- Voice Agents

What is missing?

Long-Context Acoustic and Linguistic Insights

Acoustic Insights

- How **intrinsic variability** (Physiological traits, speaker characteristics) and **extrinsic variability** (mic, codecs) construct acoustic context for ASR
- How streaming ASR application and speaker attributed ASR takes advantage of acoustic context
- **Alternative architectures** for speech recognition
- Speech encoders and **ASR endpointing** in voice agents applications.

Linguistic Insights

- **Semantic Information** Modeling in End-to-End ASR: Classical ASR-LM, Post-processing ASR corrections
- Preference-based Semantic Modeling in Post-training
- **Semantic Understanding** from ASR to Audio Contexts
- Limitation and New Evaluation of Semantic Modeling: Data Leakage via Text, Pre-training Agentic & Instruction Evaluation

Context Biasing

- **Contextual Biasing** for E2E ASR Decoder, Encoder, Hybrid Biasing
- Retrieval Toward Large-Scale Biasing
- **Retrieval Augmented Generation (RAG)** and how RAG is used for contextual Biasing

Tutorial flow

Section 1: Speech-to-Text Benchmark

Section 2: Leveraging Long Acoustic Context

Section 3: Semantic Context and Speech-Language Modeling

Section 4: Contextual Biasing and Methods Leveraging Longer Semantic Context for Speech Systems

Table of Contents

Download Slides

Introduction (10 mins) 15:30-15:40	
Shinji (30 min) 15:40-16:10	Speech-to-Text Benchmark
Taejin (40 min) 16:10-16:50	Leveraging Long Acoustic Context
Recess (10 min) 16:50-17:00	
Huck (40 min) 17:00-17:40	Semantic Context and Speech-Language Modeling
Kyu (30 min) 17:40-18:10	Contextual Biasing and Methods Leveraging Longer Semantic Context for Speech Systems
Closing Remark (10 min) 18:10-18:20	
Q&A Session (10 min) 18:20-18:30	

Speech-to-Text Benchmark

Shinji Watanabe

Speech-to-Text Benchmark

1. **Long form Beyond Utterances**
   a. Classical Speech-to-Text
   b. Long-form Conversational Speech-to-Text

2. **Evaluation Metrics**
   a. Speaker Attributed Contents
   b. Timestamp
   c. Semantics
   d. Dialog

3. **Databases**
   a. Revising existing databases

4. **Selected Benchmarks**

Long form Beyond Utterances

Classical Speech-to-Text: Utterance based

- **Utterance**
  - Speech segment corresponds to a (part of) sentence or phrase
  - A few seconds to 20 seconds (We cannot bress it so long)
  - This unit provides a lot of benefits on the **computational cost or semantic/syntactic processing** for speech

- **How to prepare a utterance?**
  - Read speech
  - Segmented speech

Classical Speech-to-Text: Read speech

- **Read a prompt**
- We can make a pair data of a prompt and corresponding audio
  - Ex) TIMIT, Wall Street Journal (WSJ), Commonvoice: https://commonvoice.mozilla.org/en,

- **Easy to collect**
  - We still need to check whether the person can correctly utter a prompt
- **Easy to anonymize**
- **Easy to cover phonetic variations**
- **But still not a real conversation** 😔

Click then read the sentence aloud

Eisenhower adviser Charles Douglas Jackson coordinated psychological warfare against Communism.

START RECORDING 1

2

3

4

5

Common Voice
moz://a

LOG IN / SIGN UP

EN

Speak | Listen | Write

Click ▶ did they accurately speak the sentence?

All but thirty of Nemed's people were wiped out.

1

2

3

4

5

YES

NO

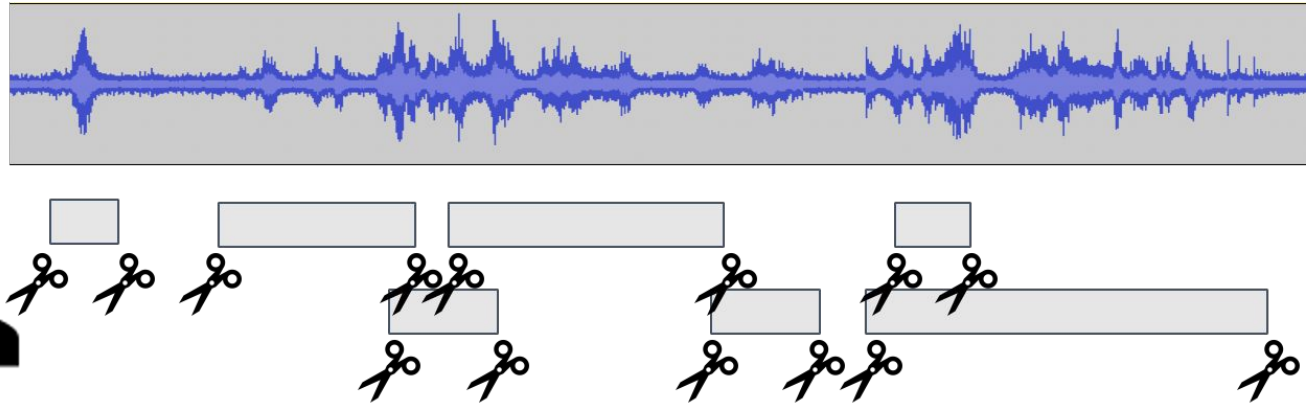
Trash

Classical Speech-to-Text: Read speech

- **Read a prompt**
- We can make a pair data of a prompt and corresponding audio
  - Ex) TIMIT, Wall Street Journal (WSJ), Commonvoice: https://commonvoice.mozilla.org/en,

- **Easy to collect**
  - We still need to check whether the person can correctly utter a prompt
- **Easy to anonymize**
- **Easy to cover phonetic variations**
- **But still not a real conversation** 😔

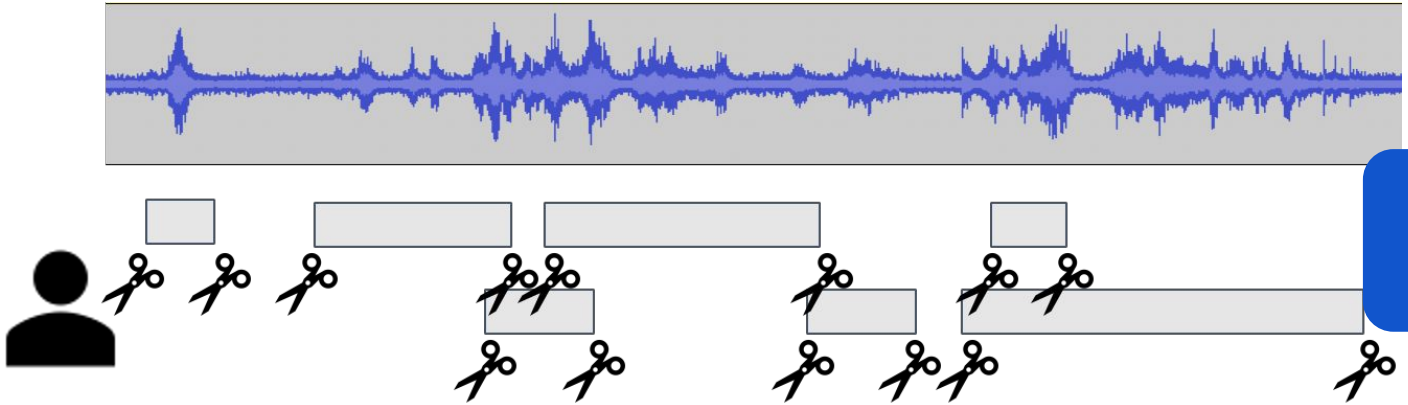
Classical Speech-to-Text: Segmented speech



Segment ID	Speaker ID	Onset	Offset	Transcription
A00011-001	A	0:01:26	0:05:04	
A00011-002	B	0:05:52	0:08:56	Oh really? Do you think you'll check out CMU?
A00011-003	A	0:09:22	0:13:04	

Given

Classical Speech-to-Text: Segmented speech

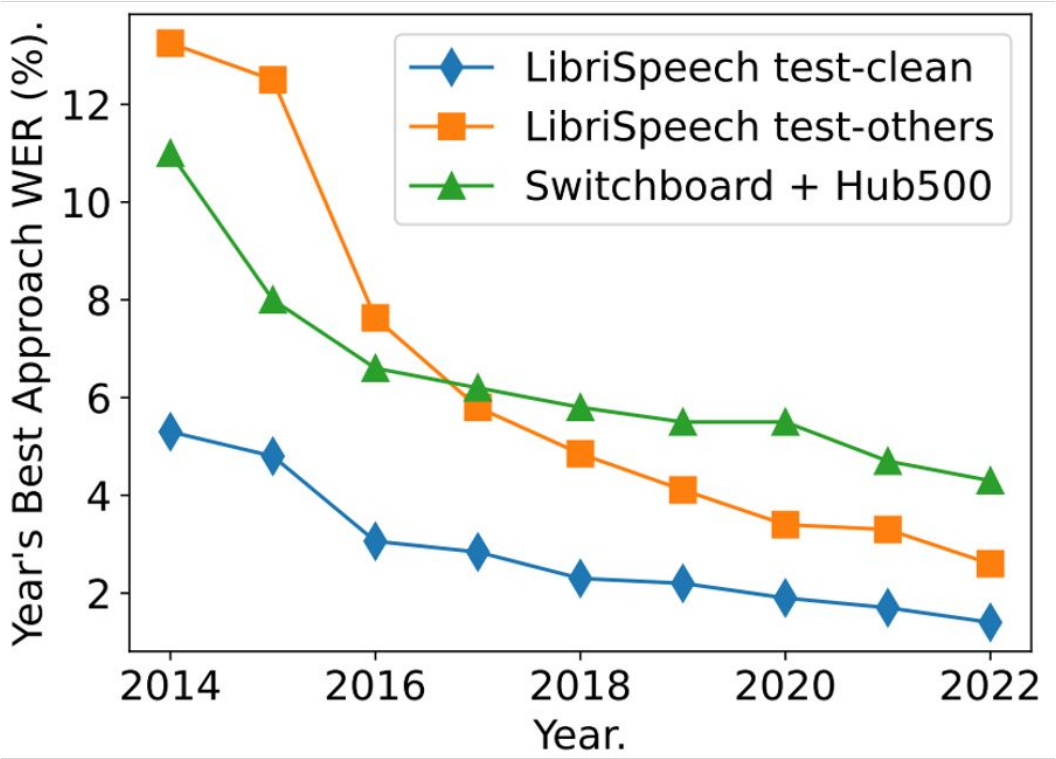


- Librispeech
- Switchboard
- etc.

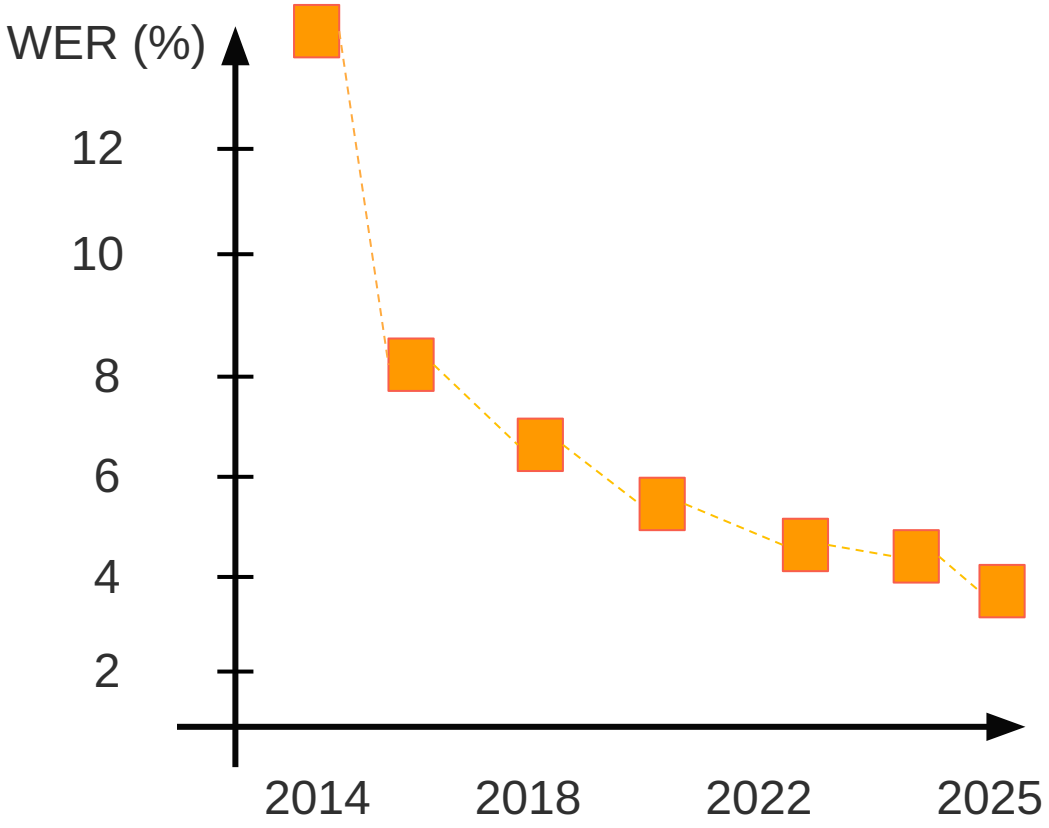
Segment ID	Speaker ID	Onset	Offset	Transcription
A00011-001	A	0:01:26	0:05:04	Yeah, um, I'm kinda thinking about going to Pittsburgh
A00011-002	B	0:05:52	0:08:56	Oh really? Do you think you'll check out CMU?
A00011-003	A	0:09:22	0:13:04	Hmm, well, I've actually never been there.

Do short-form ASR

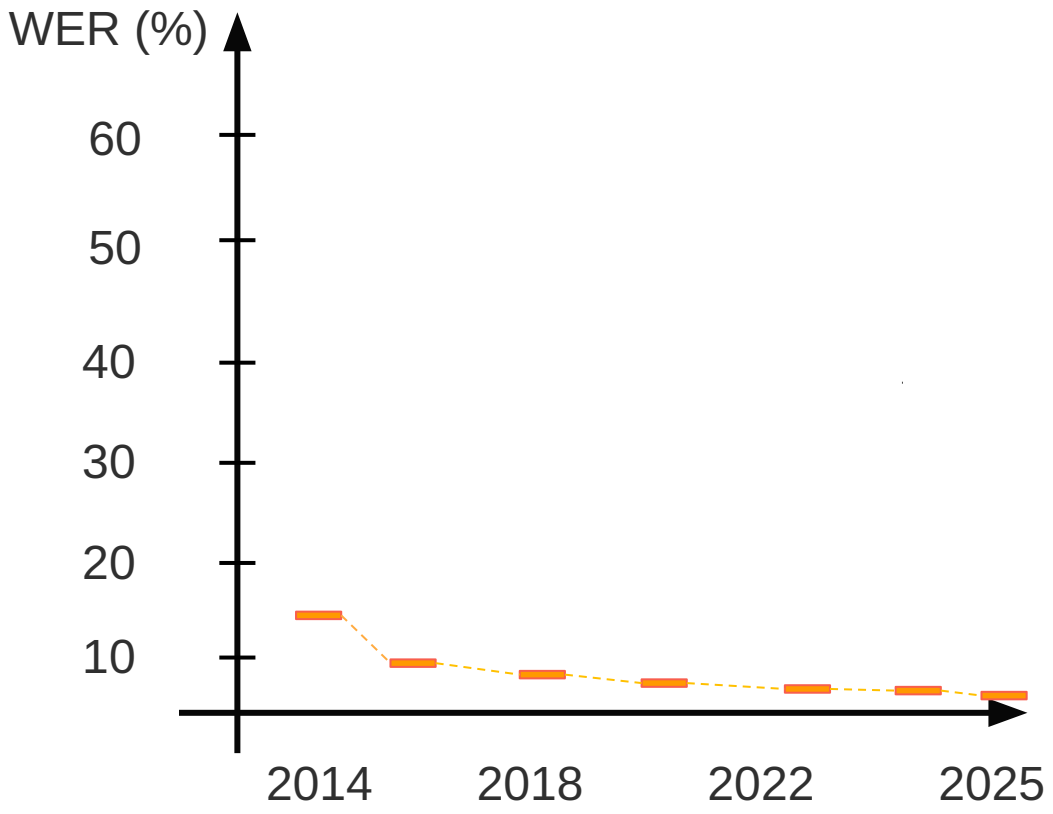
ASR is solved?



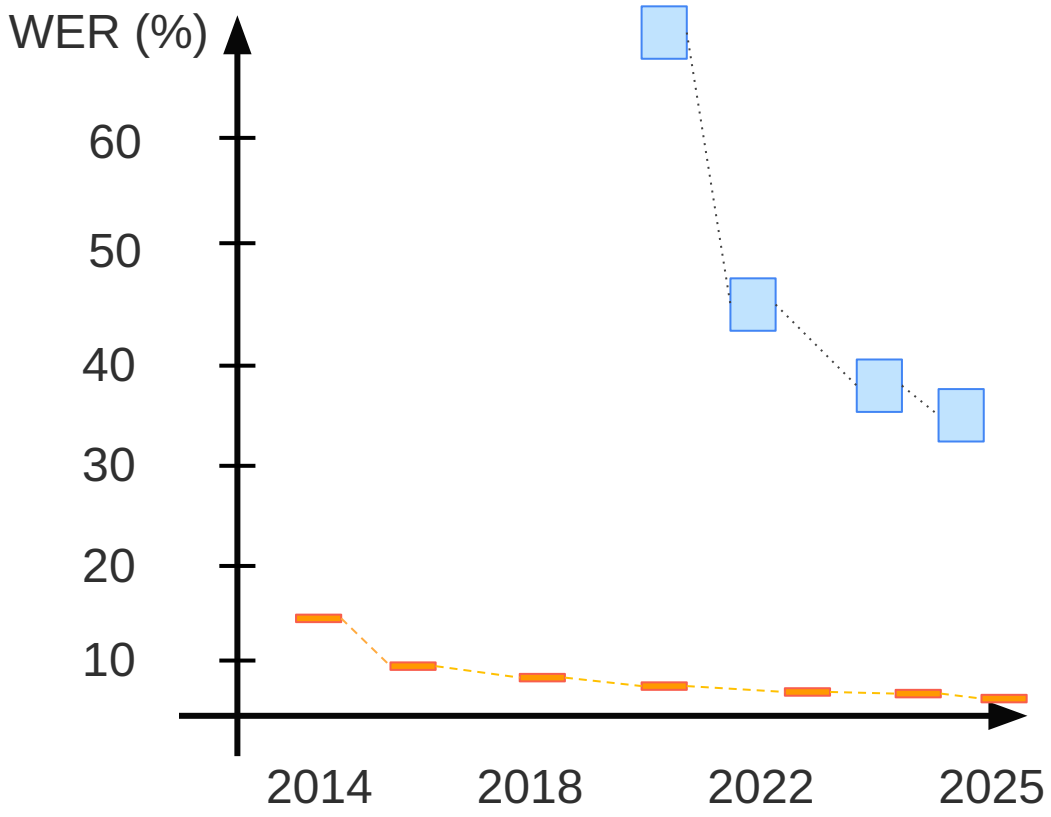
Is ASR solved (Librispeech etc.)



Is ASR solved (Librispeech etc.)



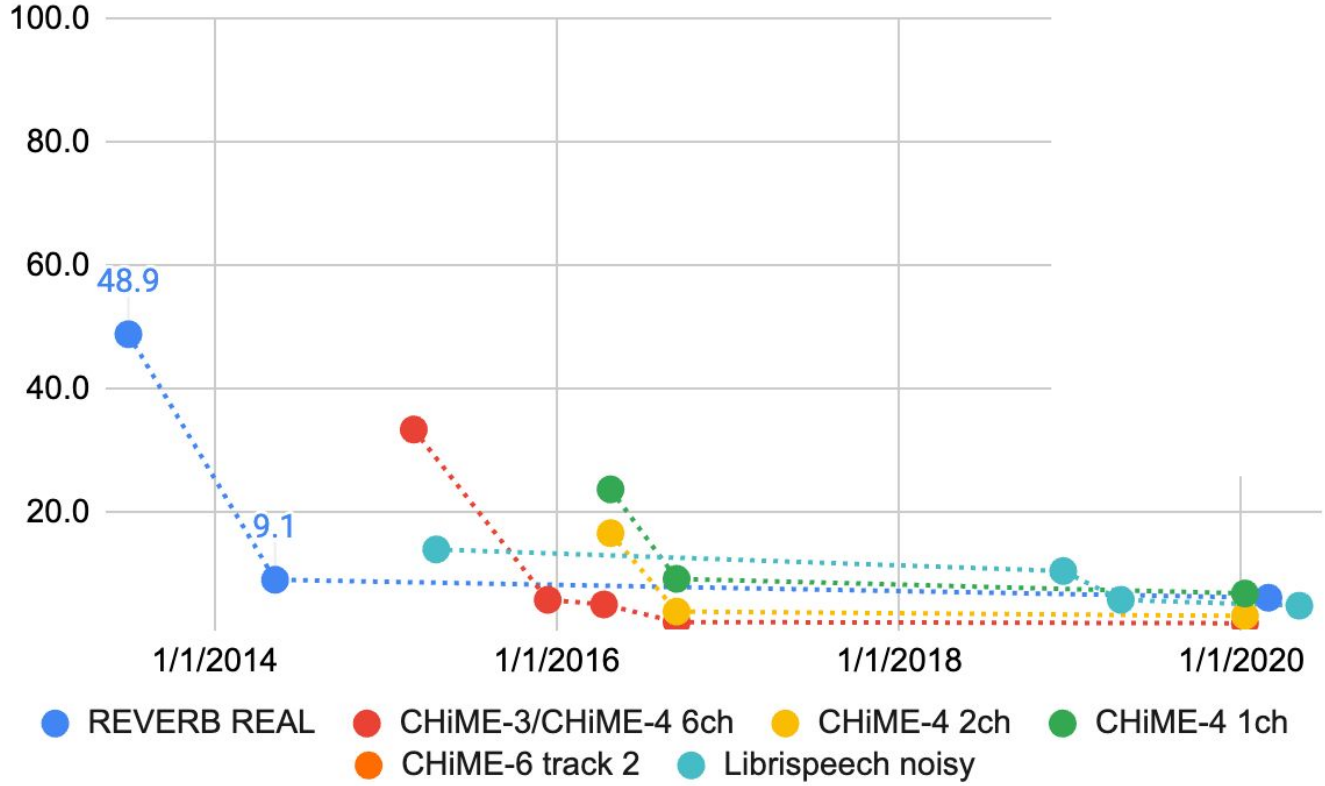
Not yet! (CHiME-6)



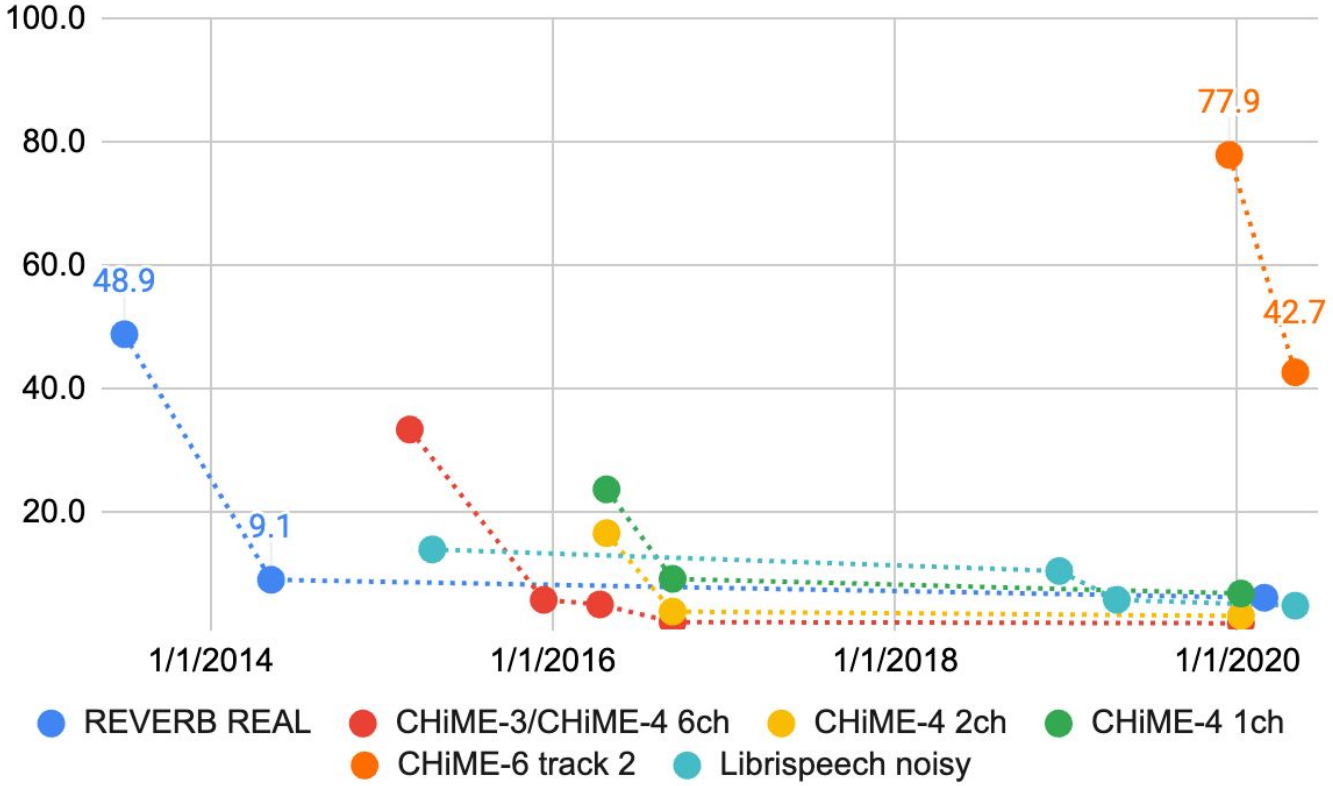
CHiME

We have to deal with various contexts

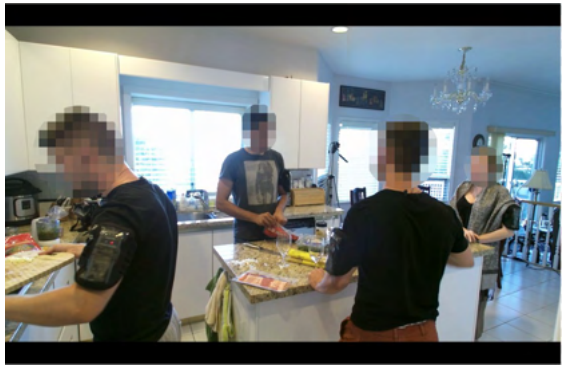
Is ASR solved?

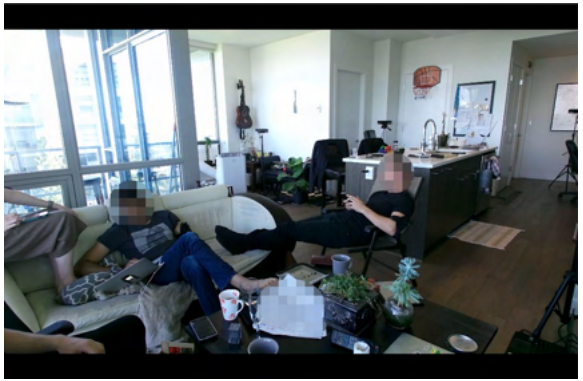


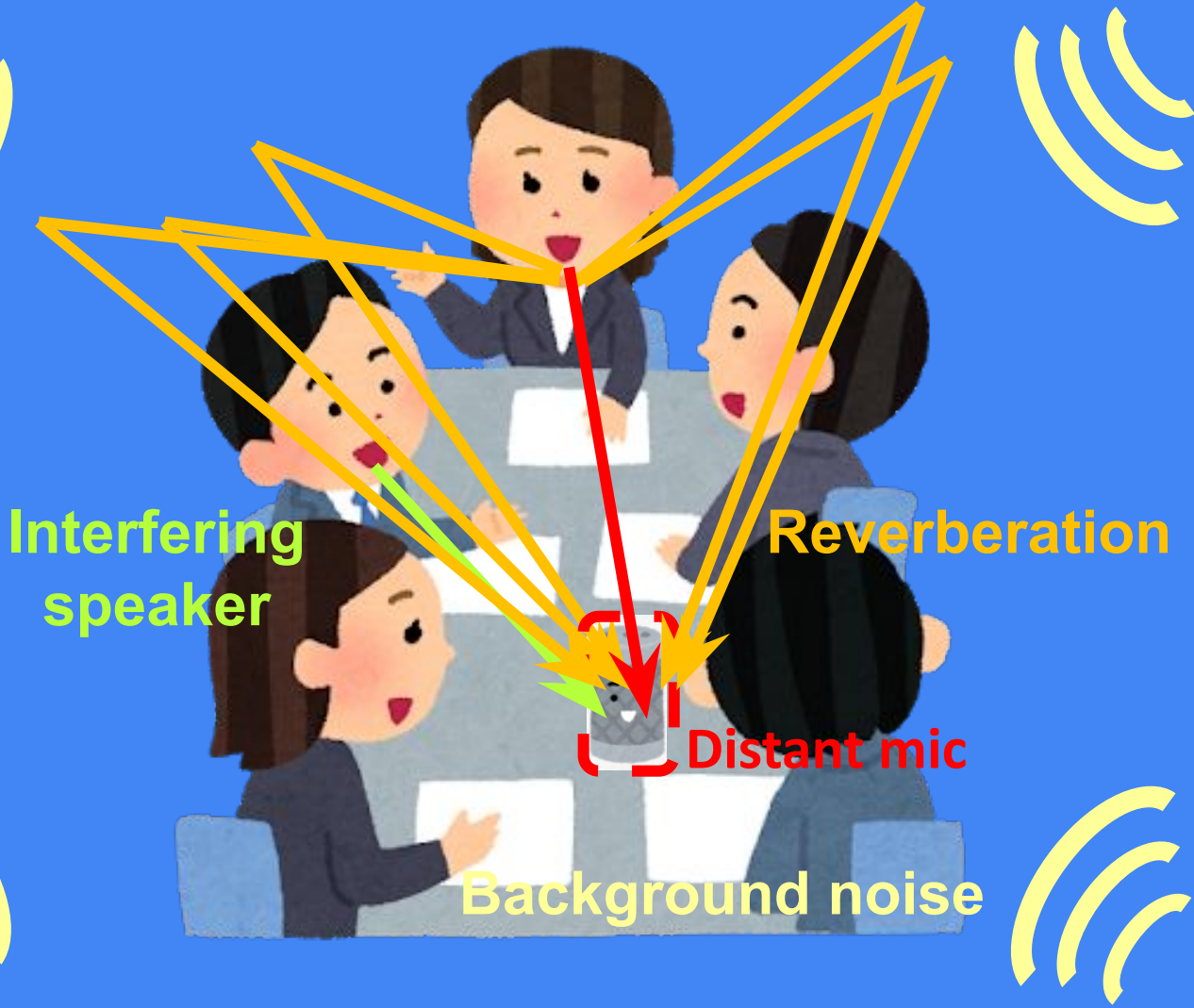
Not solved yet!

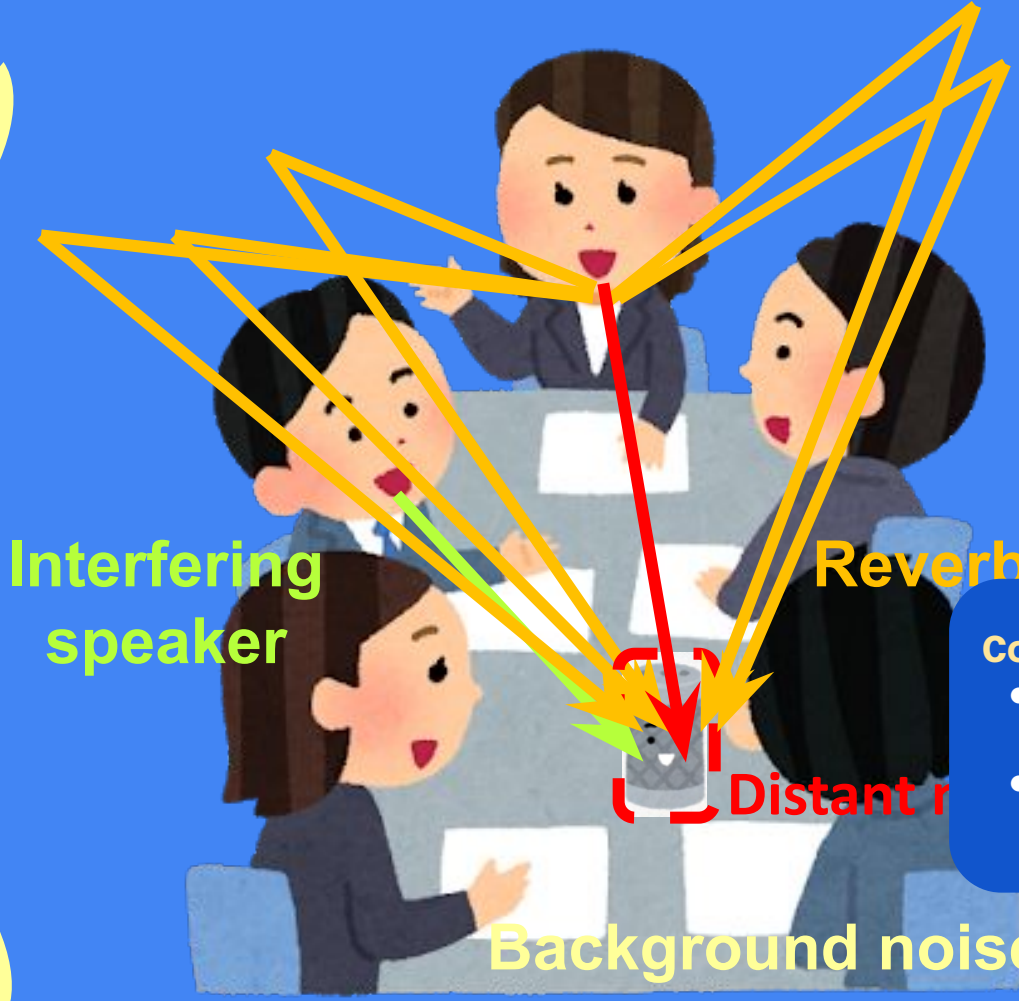


CHiME-6 https://chimechallenge.github.io/chime6/






Interfering speaker
Reverberation
Distant mic
Background noise

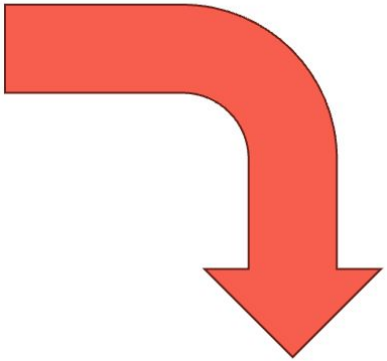


Context information

- Acoustic environments
- Speaker attribution

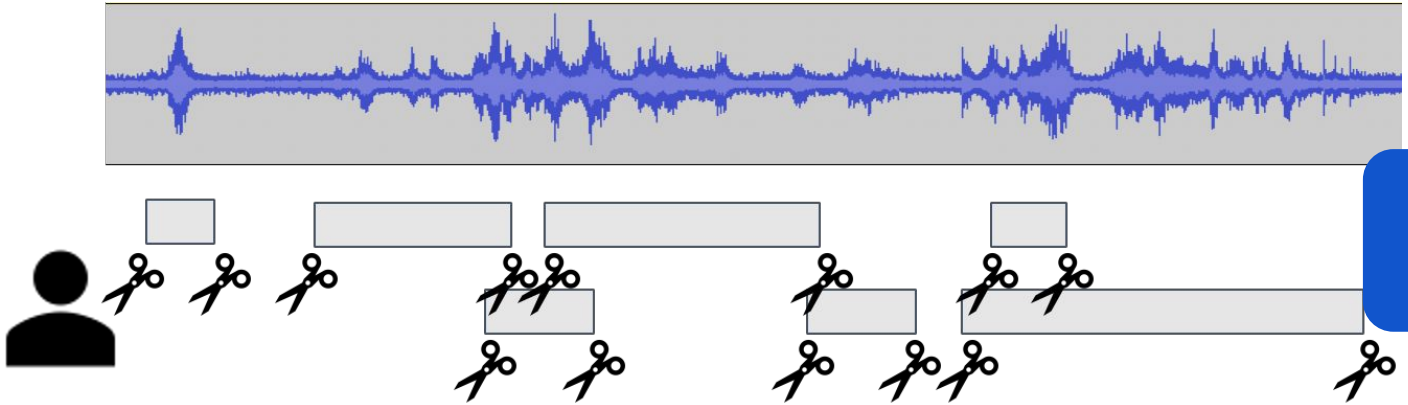
Rich transcriptions


```
{
  "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"
}
```

Segment ID τ	Speaker Cluster k	Onset (sec) $b(\tau)$	Offset (sec) $e(\tau)$	Transcription W_τ
S05_P03_001	P03	11.000	11.370	So ummm
S05_P01_002	P01	12.100	14.110	where is he?

Classical Speech-to-Text: Segmented speech

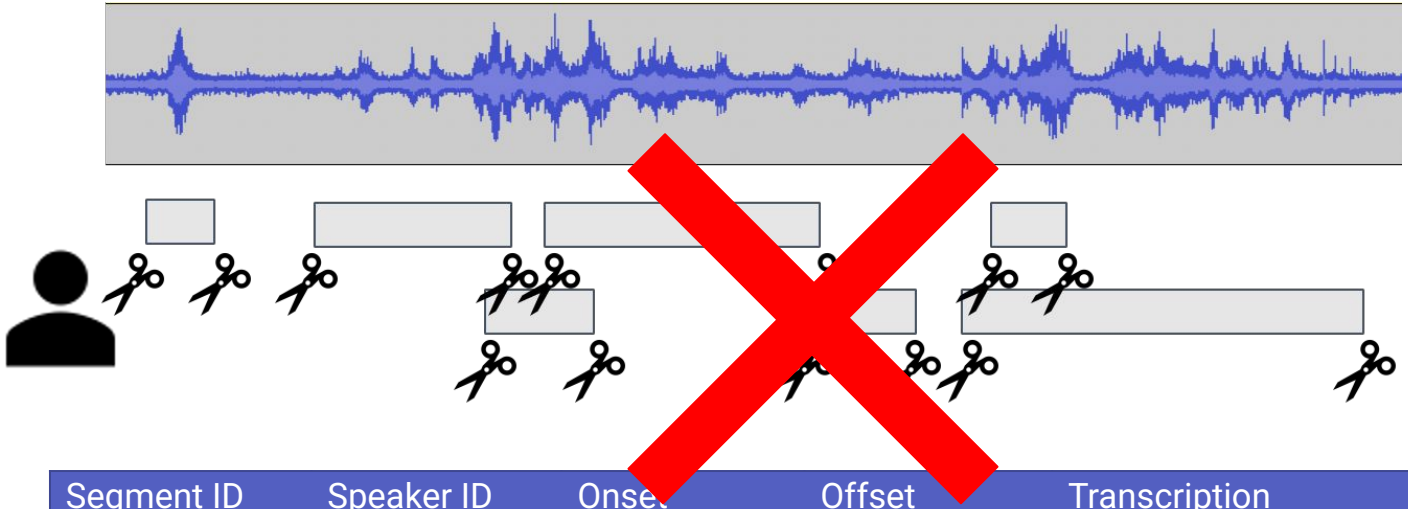


- Librispeech
- Switchboard
- etc.

Segment ID	Speaker ID	Onset	Offset	Transcription
A00011-001	A	0:01:26	0:05:04	Yeah, um, I'm kinda thinking about going to Pittsburgh
A00011-002	B	0:05:52	0:08:56	Oh really? Do you think you'll check out CMU?
A00011-003	A	0:09:22	0:13:04	Hmm, well, I've actually never been there.

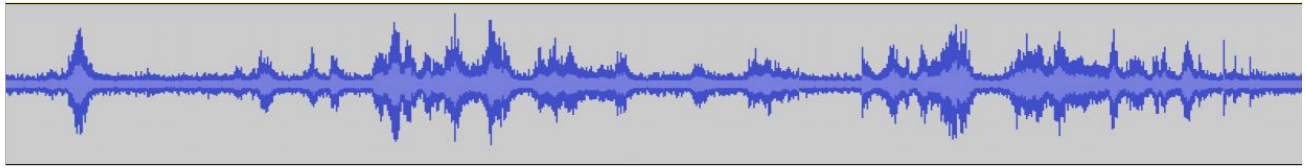
Do short-form ASR

Classical Speech-to-Text: Segmented speech



Segment ID	Speaker ID	Onset	Offset	Transcription
A00011-001	A	0:01:26	0:05:04	Yeah, um, I'm kinda thinking about going to Pittsburgh
A00011-002	B	0:05:52	0:08:56	Oh really? Do you think you'll check out CMU?
A00011-003	A	0:09:22	0:13:04	Hmm, well, I've actually never been there.

Long-form conversation recognition



ASR

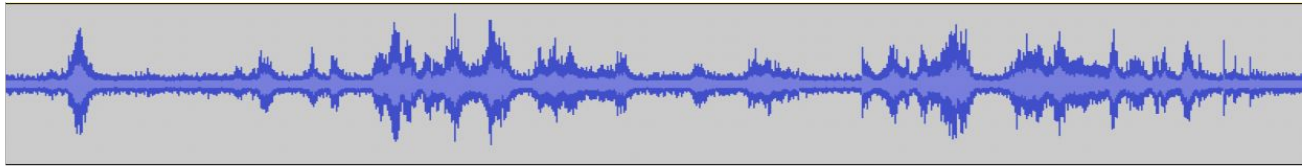
Separation

Diarization

LLMs

Segment ID	Speaker ID	Onset	Offset	Transcription
A00011-002	B	0:05:52	0:08:56	Oh really? Do you think you'll check out CMU?

Long-form conversation recognition



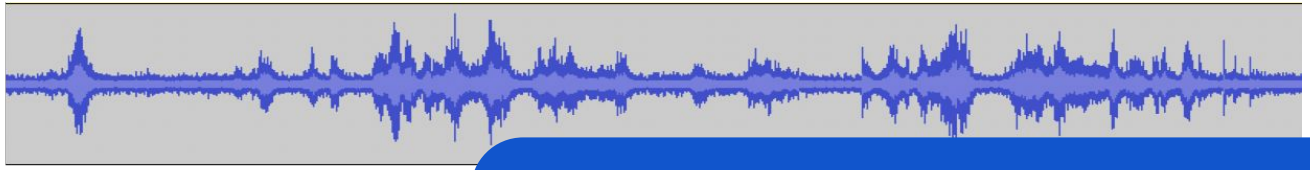
ASR | Separation

Diarization | LLMs

Segment ID	Speaker ID	Onset	Offset	Transcription
A00011-001	A	0:01:26	0:05:04	Yeah, um, I'm kinda thinking about going to Pittsburgh
A00011-002	B	0:05:52	0:08:56	Oh really? Do you think you'll check out CMU?
A00011-003	A	0:09:22	0:13:04	Hmm, well, I've actually never been there.

Predict everything by AI

Long-form conversation recognition



Context information

- Speaker attribution contents
- Timestamp
- Dialog

Segment ID	Speaker ID	Onset	Offset	Transcription
A00011-001	A	0:01:26	0:05:04	Yeah, um, I'm kinda thinking about going to Pittsburgh
A00011-002	B	0:05:52	0:08:56	Oh really? Do you think you'll check out CMU?
A00011-003	A	0:09:22	0:13:04	Hmm, well, I've actually never been there.

Predict everything by AI

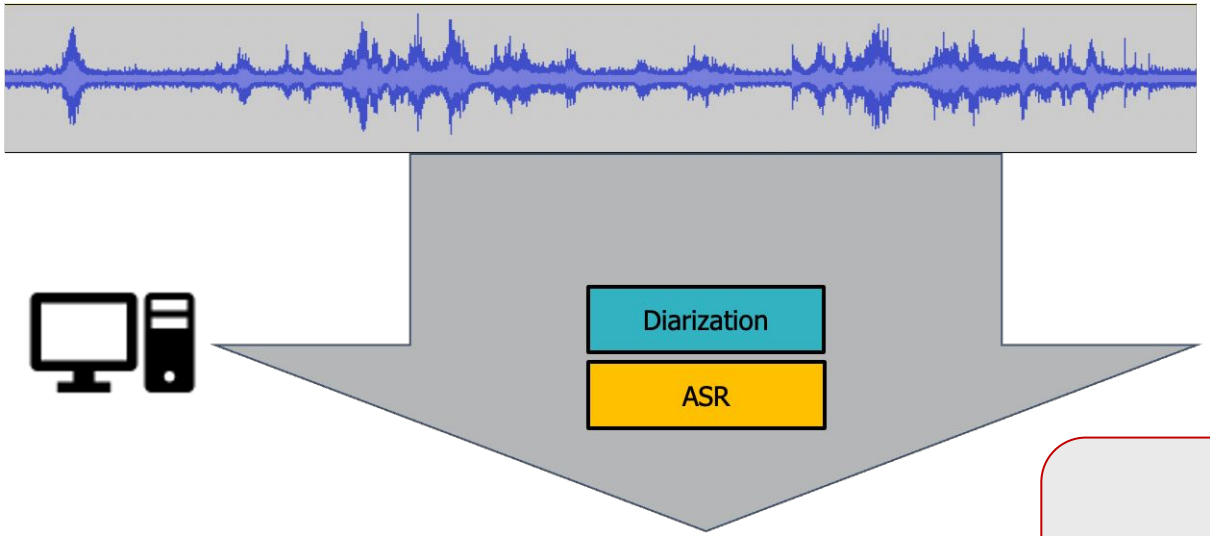
Semantic information

Segment ID	Speaker ID	Onset	Offset	Transcription
A00011-001	A	0:01:26	0:05:04	Yeah, um, I'm kinda thinking about going to Pittsburgh
A00011-002	B	0:05:52	0:08:56	Oh really? Do you think you'll check out CMU?
A00011-003	A	0:09:22	0:13:04	Hmm, well, I've actually never been there.

Context information

- Semantic information
  - Irrelevant filler information
  - keywords (Pittsburgh, CMU)
  - Location context → CMU is Carnegie Mellon University, not Central Michigan University

Long-form conversation recognition



Without any given information

Segment ID τ	Speaker Cluster k	Onset (sec) $b(\tau)$	Offset (sec) $e(\tau)$	Transc
A-0007400-0007588	A	74.0	75.88	yes
A-0008380-0009435	A	83.8	94.35	hey did you put…
B-0031242-0031700	B	312.42	317.0	hmm

Time Marked (RTTM)

- **Contextも言う**

Conversational speech (I will change it to better examples)

Shinji: Heyyy, Kyu! Man, it's been a while. How've you been?

Kyu: Oh, you know… same ol', same ol'. Still trying to win the war against my inbox. *laughs* How about you?

Shinji: Hmm… let's just say my "to-do later" pile is now an actual tower. Anyway—hey, did you check out the Interspeech 2025 lineup yet?

Kyu: Uh, not really. I've been meaning to, but… yeah, life. Anything good?

Shinji: Oh, yeah. Taejin's doing this tutorial—

Semantic information
- What is this conversation talking about
- Keywords: Interspeech, Taejin's tutorial, etc.

How to evaluate it?

- Speaker Attributed Contents
- Timestamp
- Dialog
- Semantics

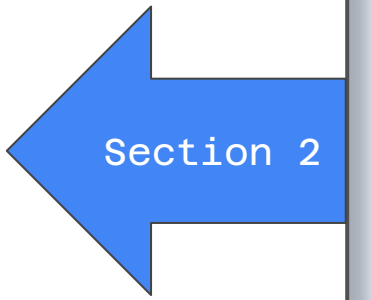
Evaluation Metrics

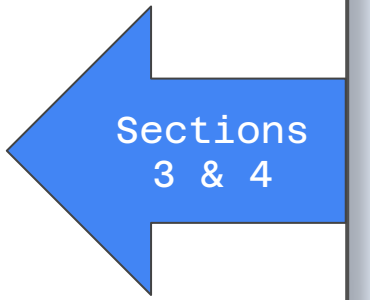
Evaluation Metrics

- **Speaker Attributed Contents**
  - concatenated minimum permutation WER (cpWER)
    - Speaker consistency across the segments
- **Timestamp**
  - Diarization error rate (DER)
  - Time-constrained cpWER (tcpWER)
- **Dialog**
  - Turn taking accuracy
- **Semantics**
  - Biased WER
  - n-gram matching approaches: BLEU, METEOR, …
  - Embedding based: BERTScore
  - LLM-based: Perplexity, LLM-as-a-judge

Evaluation Metrics

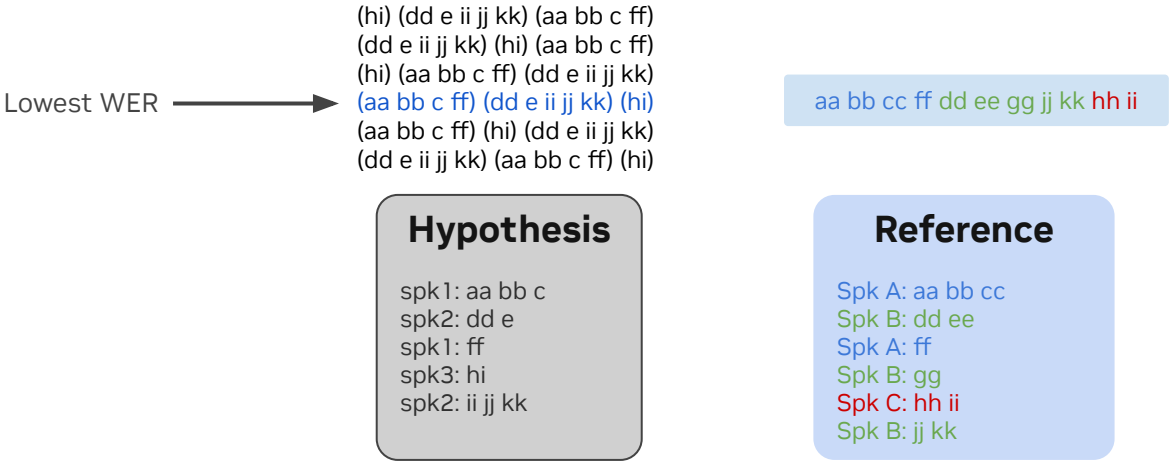
- **Speaker Attributed Contents**
  - concatenated minimum permutation WER (cpWER)
    - Speaker consistency across the segments
- **Timestamp**
  - Diarization error rate (DER)
  - Time-constrained cpWER (tcpWER)
- **Dialog**
  - Turn taking accuracy
- **Semantics**
  - Biased WER
  - n-gram matching approaches: BLEU, METEOR, …
  - Embedding based: BERTScore
  - LLM-based: Perplexity, LLM-as-a-judge





cpWER, tcpWER

Concatenated minimum-Permutation Word Error Rate



cpWER calculation:

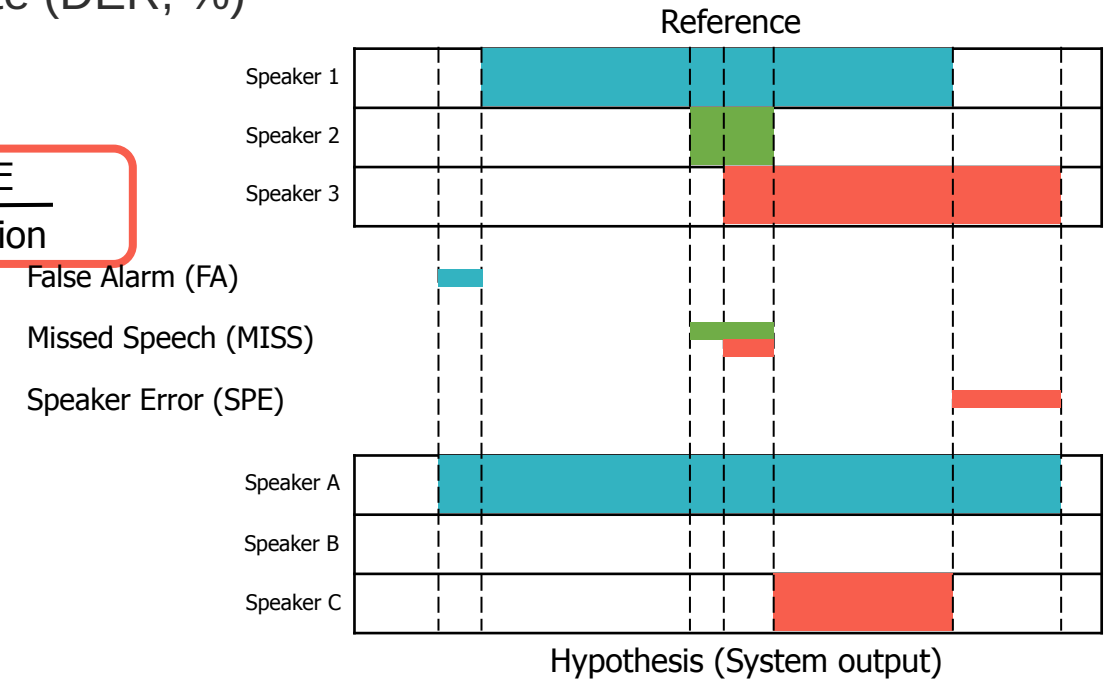
- Concatenate all utterances of each speaker for both reference and hypothesis files.
- Compute the WER between the reference and all possible speaker permutations of the hypothesis.
- Pick the lowest WER among them

• **tcp WER:** time-constrained cpWER

Speaker diarization metrics

- Diarization error rate (DER, %)

$$\text{DER} = \frac{\text{FA} + \text{MISS} + \text{SPE}}{\text{Total Speech Duration}} \quad \text{(\% of time)}$$



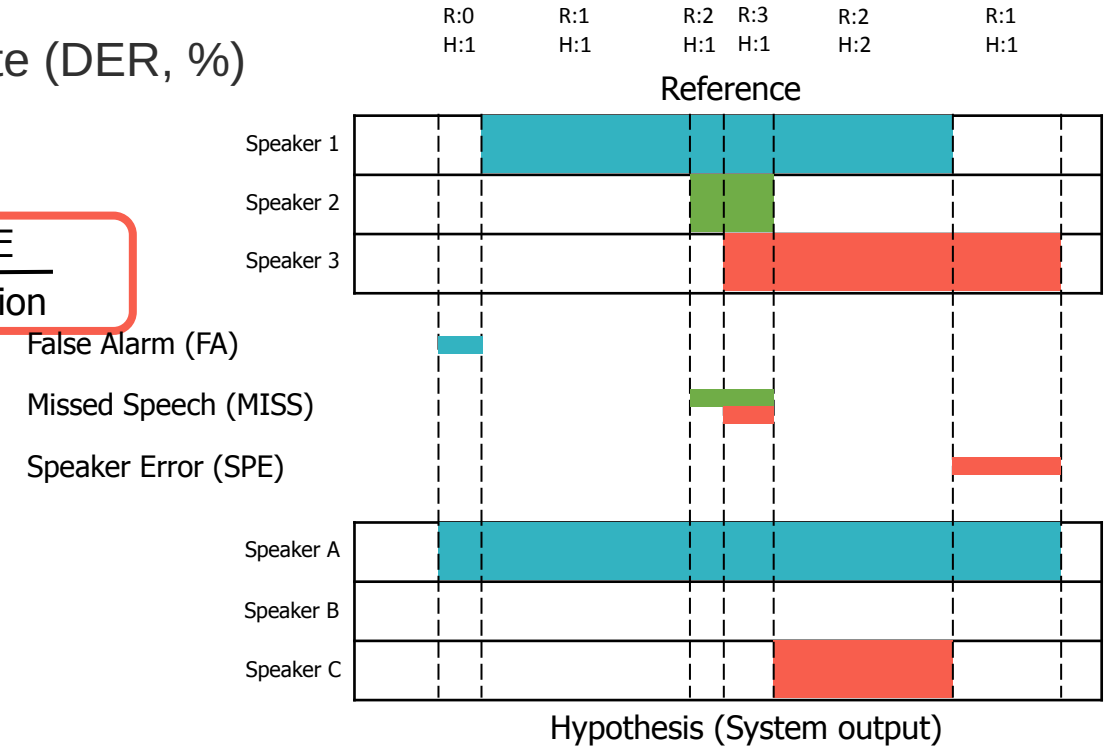
A → 1
B → 2
C → 3

Speaker diarization metrics

- Diarization error rate (DER, %)

$$\text{DER} = \frac{\text{FA} + \text{MISS} + \text{SPE}}{\text{Total Speech Duration}} \quad \text{(\% of time)}$$

We first compute #speakers for reference **R** and hypothesis **H**, respectively for each segment

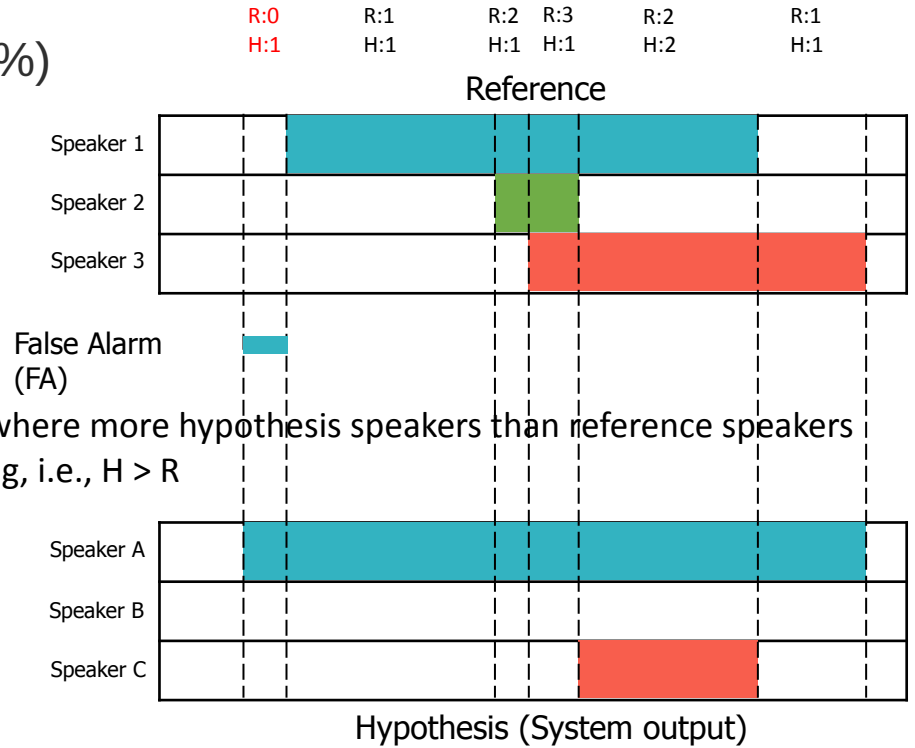


A → 1
B → 2
C → 3

Speaker diarization metrics

- Diarization error rate (DER, %)

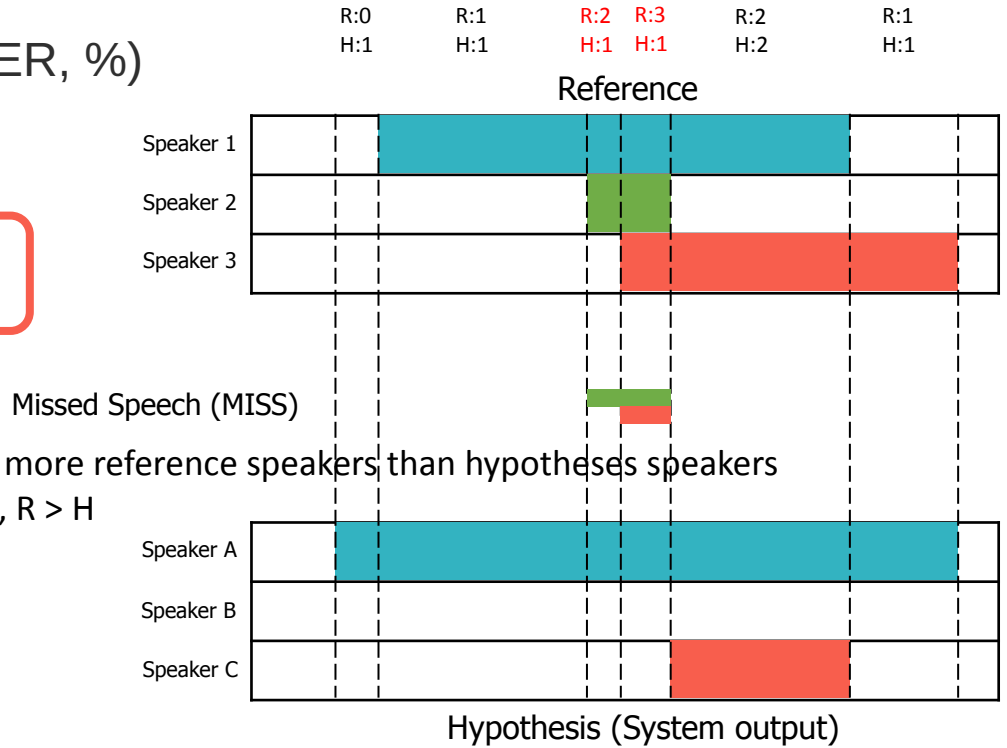
$$\text{DER} = \frac{\text{FA} + \text{MISS} + \text{SPE}}{\text{Total Speech Duration}} \quad \text{(\% of time)}$$



Speaker diarization metrics

- Diarization error rate (DER, %)

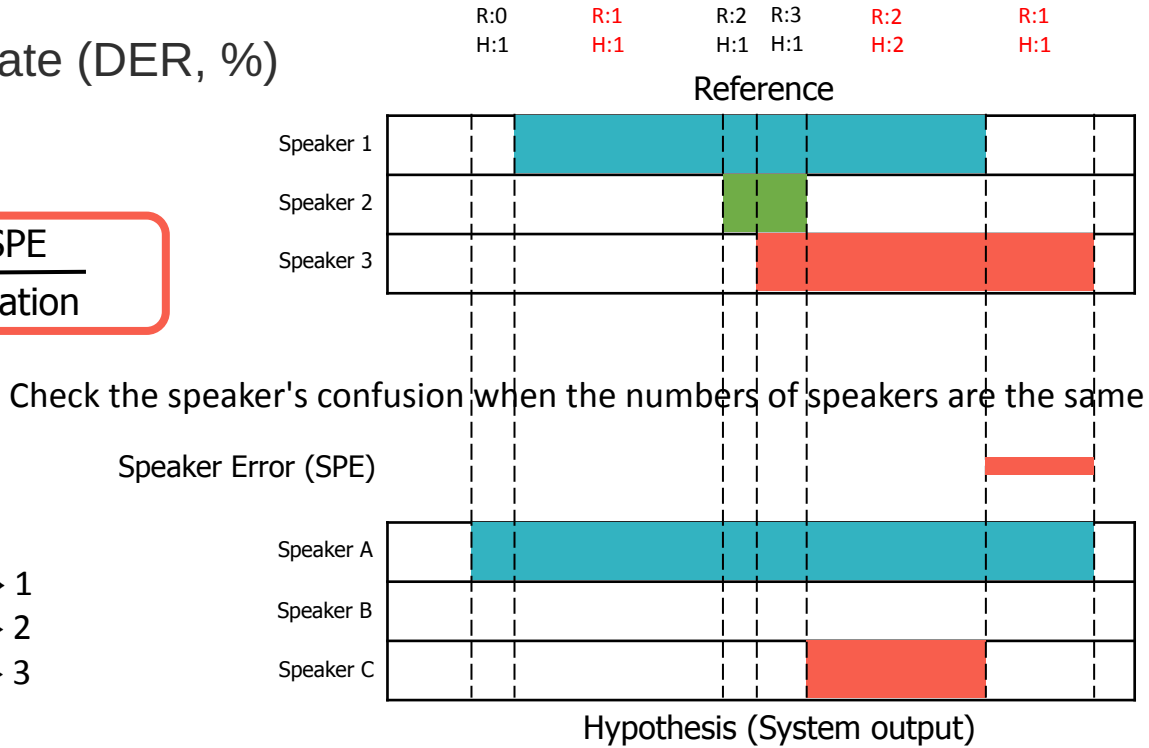
$$\text{DER} = \frac{\text{FA} + \text{MISS} + \text{SPE}}{\text{Total Speech Duration}} \quad \text{(\% of time)}$$



Speaker diarization metrics

- Diarization error rate (DER, %)

$$\text{DER} = \frac{\text{FA} + \text{MISS} + \text{SPE}}{\text{Total Speech Duration}} \quad \text{(\% of time)}$$



A → 1
B → 2
C → 3

Speaker diarization metrics

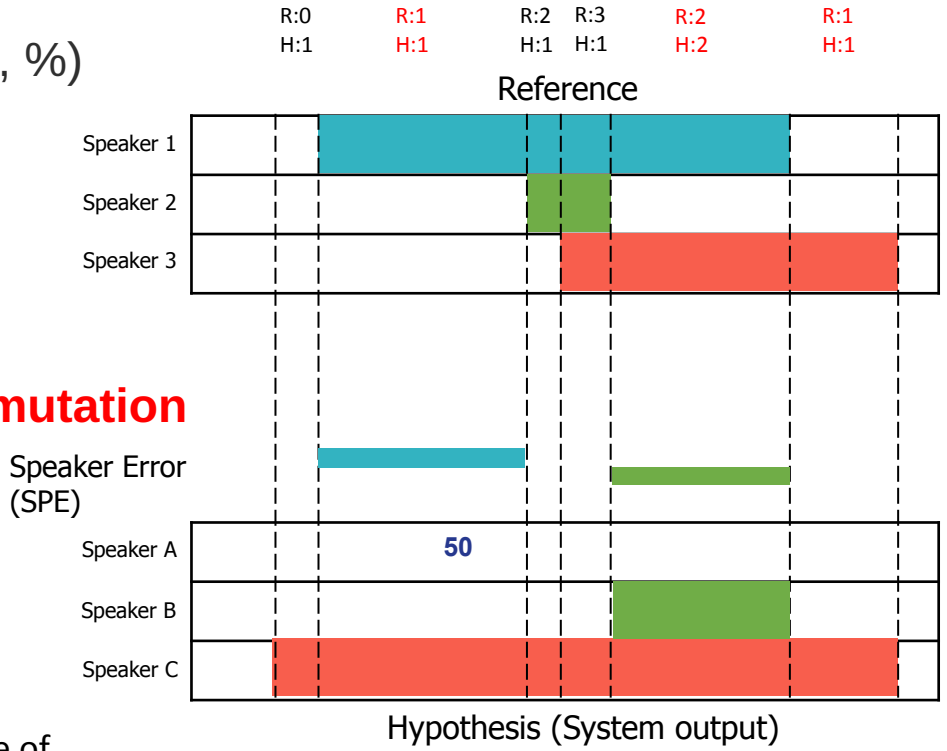
- Diarization error rate (DER, %)

$$\text{DER} = \frac{\text{FA} + \text{MISS} + \text{SPE}}{\text{Total Speech Duration}} \quad \text{(\% of time)}$$

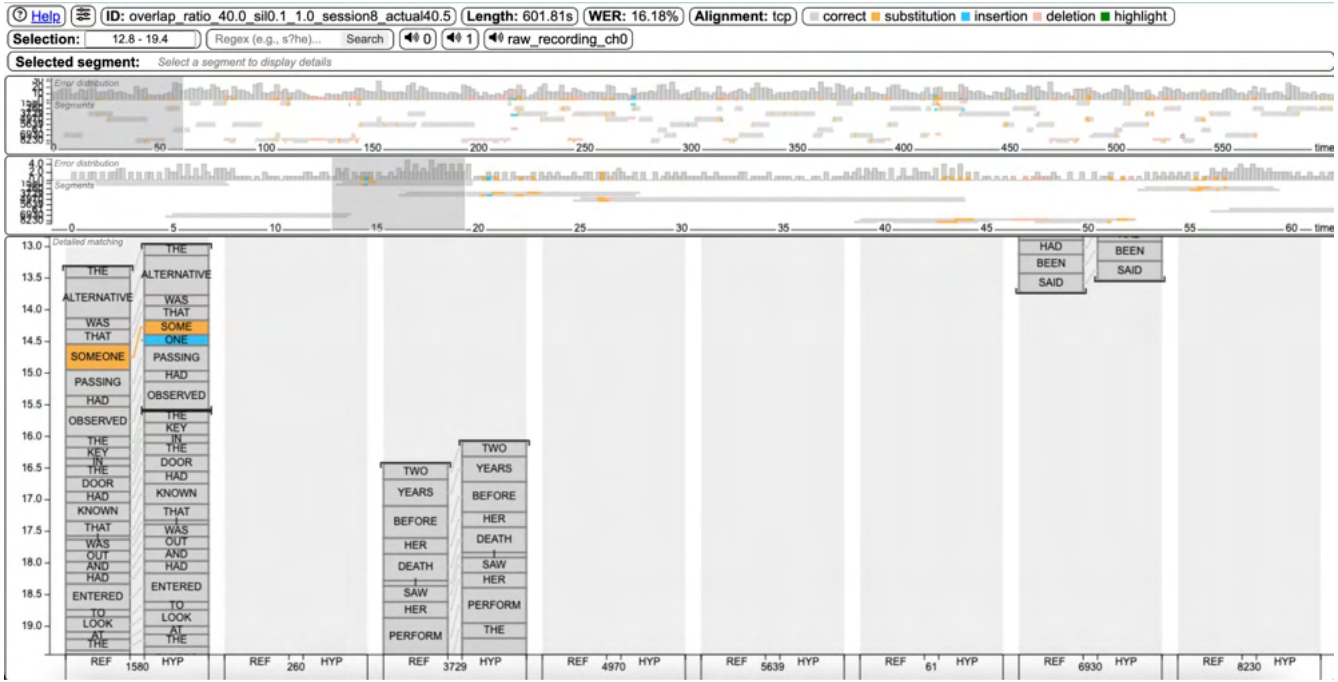
- Consider the **speaker permutation**

A → 2
B → 3
C → 1

We just pick up the minimum score of all possible permutations



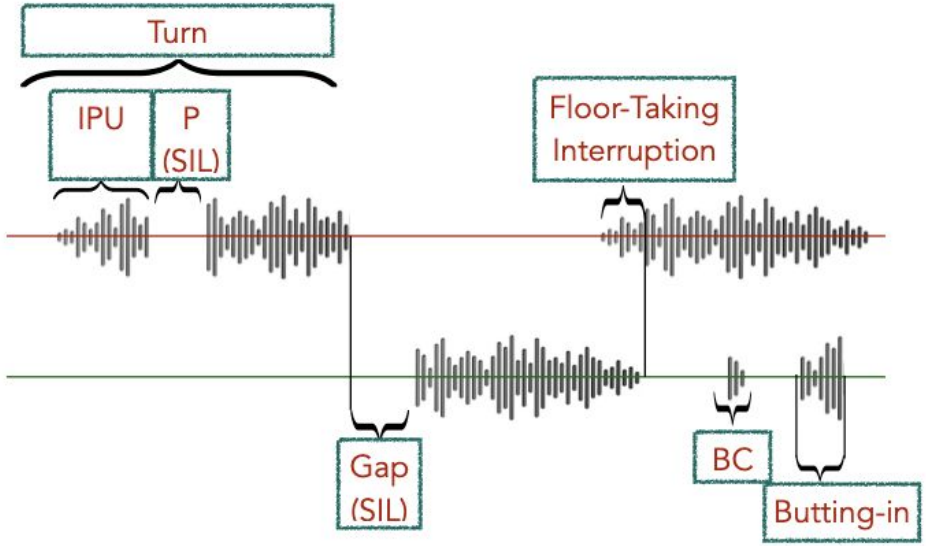
Time-constrained cpWER (tcpWER) https://github.com/fgnt/meeteval



Examples from
https://groups.uni-paderborn.de/nt/meeteval/icassp2024-demo/libricss_diarization/overlap_ratio_40.0_sil0.1_1.0_session8_actual40.5_System_tcp.html?selection=12.2-18.8

Turn taking

- Turn taking attributes
  - IPU (Inter-Pausal Unit)
  - Pause
  - Gap
  - BC (Backchannel)
  - Butting-in
  - Etc.
- Evaluate turn taking prediction by human annotations or proxy estimator trained by SWBD



Arora, Siddhant, et al. **"Talking Turns: Benchmarking Audio Foundation Models on Turn-Taking Dynamics."** *ICLR* (2025)
Inoue, Koji, et al. **"Yeah, Un, Oh: Continuous and Real-time Backchannel Prediction with Fine-tuning of Voice Activity Projection."** *NAACL* (2025).

Work by Inoue-kun

Semantic metrics

Can we use word error rate?

Candidate 1: Shinji Watanabe is an Associate Professor at the University of Pittsburgh, next to Carnegie Mellon, PA.

Candidate 2: Cindy Watanabe is an Associate Professor at Carnegie Mellon University, Pittsburgh, PA.

Candidate 3: At Carnegie Mellon University in Pittsburgh, PA, Shinji Watanabe is an Associate Professor.

Ref1: Shinji Watanabe is an Associate Professor at Carnegie Mellon University, Pittsburgh, PA.

- Which one is the lowest WER?
- Which one is semantically most correct?

Can we use word error rate?

Candidate 1: Shinji Watanabe is an Associate Professor at the University of Pittsburgh, next to Carnegie Mellon, PA. **WER = 58%**

Candidate 2: Cindy Watanabe is an Associate Professor at Carnegie Mellon University, Pittsburgh, PA. **WER = 17%**

Candidate 3: At Carnegie Mellon University in Pittsburgh, PA, Shinji Watanabe is an Associate Professor. **WER = 108%**

Ref1: Shinji Watanabe is an Associate Professor at Carnegie Mellon University, Pittsburgh, PA.

Can we use word error rate?

Candidate 1: Shinji Watanabe is an Associate Professor at the University of Pittsburgh, next to Carnegie Mellon, PA. **WER = 58%**

Candidate 2: Cindy Watanabe is an Associate Professor at Carnegie Mellon University, Pittsburgh, PA. **WER = 17%**

Candidate 3: At Carnegie Mellon University in Pittsburgh, PA, Shinji Watanabe is an Associate Professor. **WER = 108%**

Ref1: Shinji Watanabe is an Associate Professor at Carnegie Mellon University, Pittsburgh, PA.

- We should penalize the error of the important keywords → **Biased WER**
- We should avoid to use the order-dependent metrics → **BLEU** etc.

BLEU Score (Papineni+, 2002)

- Averaged N-gram precision (order agnostic)
- Extensions to deal with
  - Multiple references
  - Length penalty

Brevity Penalty

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}.$$

N=4 (typically)

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right).$$

Brevity Penalty

Modified Precision

n-gram weight (usually equal weight)

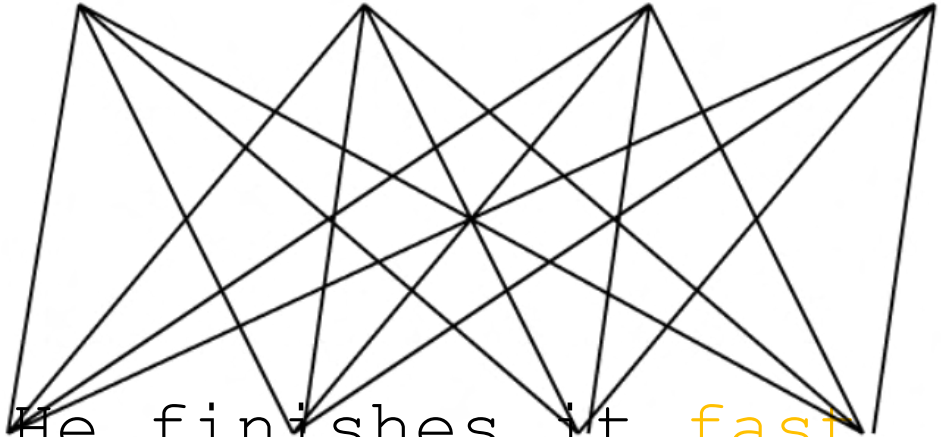
Modified Precision

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}{\sum_{C' \in \{Candidates\}} \sum_{n\text{-}gram' \in C'} Count(n\text{-}gram')}.$$

BERT Score

- Compute the pairwise similarity in the embedding space (again, order-agnostic)

Candidate: He finishes it quickly

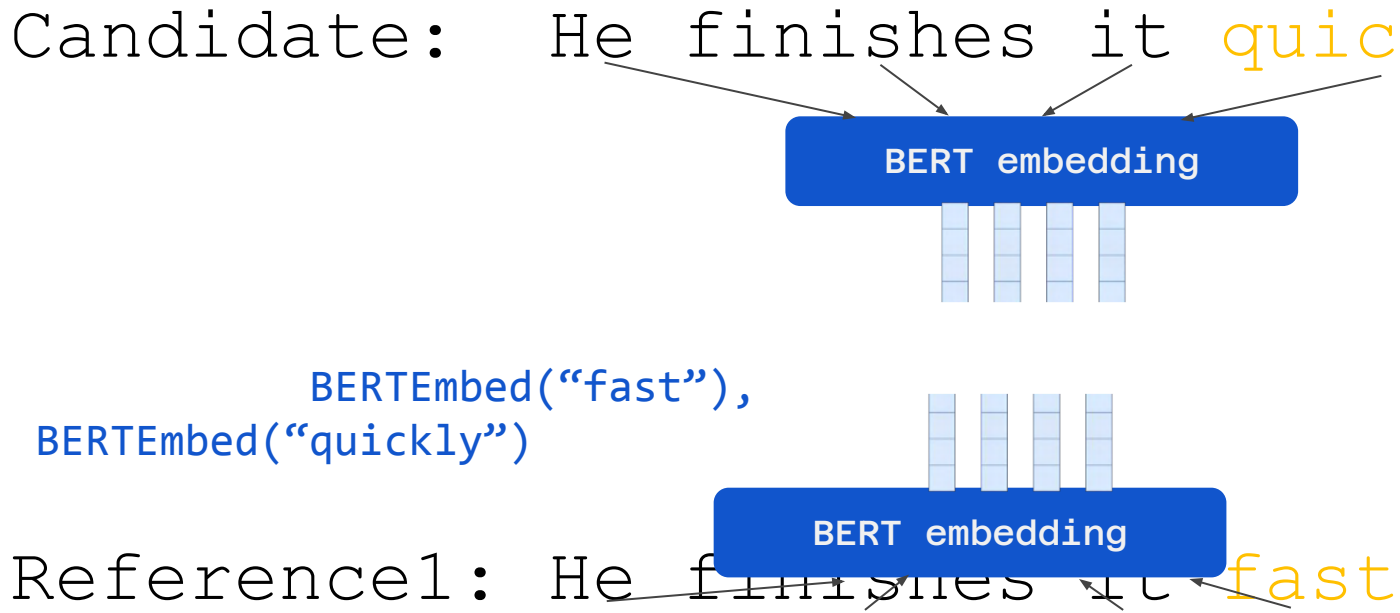


Similarity(“fast,” “quickly”)=0

Reference1: He finishes it fast

BERT Score

- Compute the pairwise similarity in the embedding space

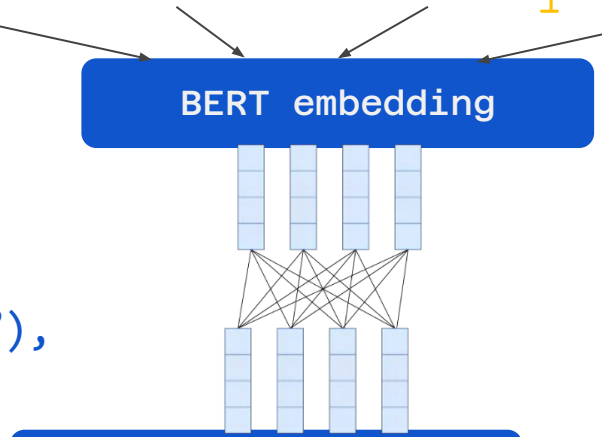


Zhang, Tianyi, et al. **"BERTScore: Evaluating Text Generation with BERT."** *ICLR* (2019)

BERT Score

- Compute the pairwise similarity in the embedding space

Candidate: He finishes it quickly



Similarity(BERTEmbed("fast"), BERTEmbed("quickly")

Reference1: He finishes it fast

Zhang, Tianyi, et al. **"BERTScore: Evaluating Text Generation with BERT."** *ICLR* (2019)

LLM-based metrics

- **Perplexity:**
  - Compute the average likelihood of generated sentences based on an LLM Θ
  - Given J-length sentence $W^{\text{test}} = (w_i \in \mathcal{V} | i = 1, \cdots, J)$ the perplexity is defined as

$$\text{PPL}(W^{\text{test}}) = \left(p(W^{\text{test}}|\Theta)\right)^{-\frac{1}{J}}$$

  - Lower is better
- **LLM as a judge:**
  - Ask an LLM to judge which model output is better

W_1 and W_2: generated sentences given prompt x

LLM judge model $J_\Theta(W_1, W_2, x)$:

- 1 If the LLM prefers W_1 over W_2
- 0 If the LLM prefers W_2 over W_1, and
- 0.5 If they are even

LLM as a judge

System prompt (to LLM Judge):
You are an impartial evaluator. Given a question, a reference answer, and two candidate answers, decide which candidate is more accurate, complete, and relevant. Provide a score from 1 to 5 for each candidate and a brief justification.

Question:
Who wrote *Pride and Prejudice*?

Reference Answer:
Jane Austen.

Model A Answer:
Pride and Prejudice was written by Jane Austen in 1813.

Model B Answer:
It was written by Charlotte Brontë.

Judge LLM's Output:

- **Model A Score:** 5/5 — Correct author, includes additional accurate publication year.
- **Model B Score:** 1/5 — Incorrect author.
- **Decision:** Model A is clearly better.

Databases

Databases

https://arxiv.org/abs/2507.18161

Corpus	Authors	Year	Real	Long	Multi Speaker	Far Field	Multi Domain
Santa Barbara	Du Bois et al.	2000	✓	✓	✓	✓	✓
Aurora2-6	Hirsch and Pearce	2000-2006	✗	✗	✗	✓	✓
RT evaluations	Garofolo et al.	2002-2009	✓	✓	✓	✓	✓
ICSI	Janin et al.	2003	✓	✓	✓	✓	✗
Fisher	Cieri et al.	2004	✓	✓	✓	✗	✗
AMI	Carletta et al.	2005	✓	✓	✓	✓	✗
Pascal	Cooke et al.	2006	✗	✗	✓	✗	✗
CHIL	Mostefa et al.	2007	✓	✓	✓	✓	✗
CHiME Corpus	Christensen et al.	2010	✗	✗	✗	✓	✗
Mixer 6 Speech	Brandschain et al.	2010	✓	✓	✓	✓	✗
CHiME-1	Barker et al.	2011	✗	✗	✗	✓	✗
COSINE	Stupakov et al.	2012	✓	✓	✓	✓	✗
Sheffield Wargames	Fox et al.	2013	✓	✓	✓	✓	✗
REVERB	Kinoshita et al.	2013	✗	✗	✗	✓	✗
CHiME-2	Vincent et al.	2013	✗	✗	✓	✓	✗
DIRHA	Cristoforetti et al.	2014	✗	✗	✗	✓	✗
CHiME-3	Barker et al.	2015	✗	✗	✓	✗	✗
CHiME-4	Vincent et al.	2015	✗	✗	✓	✗	✗
ASpIRE	Harper	2015	✓	✓	✗	✓	✗
CHiME-5	Barker et al.	2018	✓	✗	✓	✓	✗
VOiCES	Richey et al.	2018	✗	✗	✗	✓	✗
DiPCo	Van Segbroeck et al.	2019	✓	✓	✓	✓	✗
CHiME-6	Watanabe et al.	2020	✓	✓	✓	✓	✗
Aishell-4	Fu et al.	2020	✓	✓	✓	✓	✗
AliMeeting	Yu et al.	2020	✓	✓	✓	✓	✗
LibriCSS	Chen et al.	2020	✗	✓	✓	✓	✗
EGO4D	Grauman et al.	2022	✓	✓	✓	✓	✓
MISP	Wang et al.	2022	✓	✓	✓	✓	✓
CHiME-7 DASR	Cornell et al.	2023	✓	✓	✓	✓	✓
CHiME-8 DASR	Cornell et al.	2024	✓	✓	✓	✓	✓
CHiME-8 NOTSOFAR-1	Vinnikov et al.	2024	✓	✓	✓	✓	✗
CHiME-8 MMCSG	Zmolikova et al.	2024	✓	✓	✓	✓	✗

Cornell, Samuele, et al. **"Recent Trends in Distant Conversational Speech Recognition: A Review of CHiME-7 and 8 DASR Challenges."** *arXiv preprint arXiv:2507.18161* (2025).

Databases

https://arxiv.org/abs/2507.18161

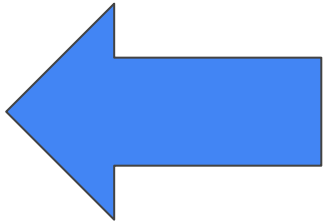
Corpus	Authors	Year	Real	Long	Multi Speaker	Far Field	Multi Domain
Sant							
Auro							
RT e							
ICSI							
Fisher	Cieri et al.	2004	✓	✓	✓	✗	✗
AMI	Carletta et al.	2005	✓	✓	✓	✓	✗
Pascal	Cooke et al.	2006	✗	✗	✓	✗	✗
CHIL	Mostefa et al.	2007	✓	✓	✓	✓	✗
CHiME Corpus	Christensen et al.	2010	✗	✗	✗	✓	✗
Mixer 6 Speech	Brandschain et al.	2010	✓	✓	✓	✓	✗
CHiME-1	Barker et al.	2011	✗	✗	✗	✓	✗
COSINE	Stupakov et al.	2012	✓	✓	✓	✓	✗
Sheffield Wargames	Fox et al.	2013	✓	✓	✓	✓	✗
REVERB	Kinoshita et al.	2013	✗	✗	✗	✓	✗
CHiME-2	Vincent et al.	2013	✗	✗	✓	✓	✗
DIRHA	Cristoforetti et al.	2014	✗	✗	✗	✓	✗
CHiME-3	Barker et al.	2015	✗	✗	✓	✗	✗
CHiME-4	Vincent et al.	2015	✗	✗	✓	✗	✗
ASpIRE	Harper	2015	✓	✓	✗	✓	✗
CHiME-5	Barker et al.	2018	✓	✗	✓	✓	✗
VOiCES	Richey et al.	2018	✗	✗	✗	✓	✗
DiPCo	Van Segbroeck et al.	2019	✓	✓	✓	✓	✗
CHiME-6	Watanabe et al.	2020	✓	✓	✓	✓	✗
Aishell-4	Fu et al.	2020	✓	✓	✓	✓	✗
AliMeeting	Yu et al.	2020	✓	✓	✓	✓	✗
LibriCSS	Chen et al.	2020	✗	✓	✓	✓	✗
EGO4D	Grauman et al.	2022	✓	✓	✓	✓	✓
MISP	Wang et al.	2022	✓	✓	✓	✓	✓
CHiME-7 DASR	Cornell et al.	2023	✓	✓	✓	✓	✓
CHiME-8 DASR	Cornell et al.	2024	✓	✓	✓	✓	✓
CHiME-8 NOTSOFAR-1	Vinnikov et al.	2024	✓	✓	✓	✓	✗
CHiME-8 MMCSG	Zmolikova et al.	2024	✓	✓	✓	✓	✗

Cornell, Samuele, et al. **"Recent Trends in Distant Conversational Speech Recognition: A Review of CHiME-7 and 8 DASR Challenges."** *arXiv preprint arXiv:2507.18161* (2025).

Databases

https://arxiv.org/abs/2507.18161

Corpus	Authors	Year	Real	Long	Multi Speaker	Far Field	Multi Domain
Santa Barbara	Du Bois et al.	2000	✓	✓	✓	✓	✓
Aurora2-6	Hirsch and Pearce	2000-2006	✗	✗	✗	✓	✓
RT evaluations	Garofolo et al.	2002-2009	✓	✓	✓	✓	✓
ICSI	Janin et al.	2003	✓	✓	✓	✓	✗
Fisher	Cieri et al.	2004	✓	✓	✓	✗	✗
AMI	Carletta et al.	2005	✓	✓	✓	✓	✗
Pascal	Cooke et al.	2006	✗	✗	✓	✗	✗
CHIL	Mostefa et al.	2007	✓	✓	✓	✓	✗
CHiME Corpus	Christensen et al.	2010	✗	✗	✗	✓	✗
Mixer 6 Speech	Brandschain et al.	2010	✓	✓	✓	✓	✗
CHiME-1	Barker et al.	2011	✗	✗	✗	✓	✗
COSINE	Stupakov et al.	2012	✓	✓	✓	✓	✗
Sheffield Wargames	Fox et al.	2013	✓	✓	✓	✓	✗
REVERB	Kinoshita et al.	2013	✗	✗	✗	✓	✗
CHiME-2	Vincent et al.	2013	✗	✗	✓	✓	✗
DIRHA	Cristoforetti et al.	2014	✗	✗	✗	✓	✗
CHiME-3	Barker et al.	2015	✗	✗	✓	✗	✗
CHiME-4	Vincent et al.	2015	✗	✗	✓	✗	✗
ASpIRE	Harper	2015	✓	✓	✗	✓	✗
CHiME-5	Barker et al.	2018	✓	✗	✓	✓	✗
VOiCES	Richey et al.	2018	✗	✗	✗	✓	✗
DiPCo	Van Segbroeck et al.	2019	✓	✓	✓	✓	✗
CHiME-6	Watanabe et al.	2020	✓	✓	✓	✓	✗
Aishell-4	Fu et al.	2020	✓	✓	✓	✓	✗
AliMeeting	Yu et al.	2020	✓	✓	✓	✓	✗
LibriCSS	Chen et al.	2020	✗	✓	✓	✓	✗
EGO4D	Grauman et al.	2022	✓	✓	✓	✓	✓
MISP	Wang et al.	2022	✓	✓	✓	✓	✓
CHiME-7 DASR	Cornell et al.	2023	✓	✓	✓	✓	✓
CHiME-8 DASR	Cornell et al.	2024	✓	✓	✓	✓	✓
CHiME-8 NOTSOFAR-1	Vinnikov et al.	2024	✓	✓	✓	✓	✗
CHiME-8 MMCSG	Zmolikova et al.	2024	✓	✓	✓	✓	✗



Cornell, Samuele, et al. **"Recent Trends in Distant Conversational Speech Recognition: A Review of CHiME-7 and 8 DASR Challenges."** *arXiv preprint arXiv:2507.18161* (2025).

Long-form Databases https://domklement.github.io/sbcsae/

	SBCSAE	AliMeeting	AMI	CHiME-6	DIHARD3	DiPCo
# Recordings	60	237	169	20	513	10
# Speakers	439	537	189	48	581	32
Average Duration (min.)	23:18 ± 04:25	31:59 ± 03:09	35:11 ± 13:32	150:31 ± 11:26	07:51 ± 02:52	32:00 ± 12:12
Overlap by Time (%)	10.3 ± 8.3	27.2 ± 18.4	12.9 ± 6.1	33.2 ± 10.5	9.4 ± 10.8	26.9 ± 7.5

- Much longer than the utterance-based processing (5–20 sec)
- Various overlap ratios

Maciejewski, Matthew, et al. **"Evaluating the Santa Barbara Corpus: Challenges of the Breadth of Conversational Spoken Language."** *Interspeech* (2024).

CHiME-5 → CHiME-6

- **Revisiting** the CHiME-5 corpus
- Removing the segmentation boundary information
- Multi-speaker scenario
- Longer duration (2.5 hours)
- Similar corpus: AMI, DiPCo, AliMeeting,NOTSOFAR
- This setup is often discussed in Section 2

	Segmentation	Metrics	Length
CHiME-5	Given	WER	5 – 20 sec.
CHiME-6	Not given	DER + cpWER	2.5 hours

Watanabe, Shinji, et al. **"CHiME-6 Challenge: Tackling Multispeaker Speech Recognition for Unsegmented Recordings."** Proc. CHiME *Workshop* (2020).

Librispeech → Libriheavy

- **Revisiting** the Librispeech corpus
- 50,000h English read speech from LibriVox
- Rich transcripts: punctuation, casing, **previous text context** ➡ Enables **context-aware ASR research**
- Open-source audio-text alignment pipeline
- Longer duration version (from 20 - 100 seconds)

	Amount	Transcripts	Length
Librispeech	~1000h	Normalized text	5 – 20 sec.
Libriheavy	~50,000h	Punctuation, casing, previous text context	20 – 100 sec. (long version)

Kang, Wei, et al. **"Libriheavy: A 50,000 hours ASR corpus with punctuation casing and context."** *ICASSP* (2024)

Selected Benchmarks

CHiME series https://www.chimechallenge.org/

- **CHiME-5 → CHiME-6**
  - One of the first benchmarks for long-form setup
- **CHiME-7 DASR**
  - Multiple devices in diverse scenarios
  - CHiME-6 + DiPCo + Mixer 6
  - cpWER → tcpWER
- **CHiME-8 DASR**
  - + NOTSOFAR-1
  - Allow the use of **a strong LLM** (related to Section 3)

Librispeech for contextual biasing

- Specifies a biasing list with N = {100, 500, 1000, 2000} distractors
- U-WER: **Unbiased WER** measured on words **NOT IN** the biasing List
- B-WER: **biased WER** measured on words **IN** the biasing list

Model	N = 100		N = 500		N = 1000		N = 2000	
	test-clean	*test-other*	*test-clean*	*test-other*	*test-clean*	*test-other*	*test-clean*	*test-other*
B1: RNNT Baseline	3.65 (2.4/14.1)	9.61 (7.2/30.6)	3.65 (2.4/14.1)	9.61 (7.2/30.6)	3.65 (2.4/14.1)	9.61 (7.2/30.6)	3.65 (2.4/14.1)	9.61 (7.2/30.6)
S1: B1 + DB-RNNT	3.11 (2.3/9.8)	8.79 (7.1/23.4)	3.24 (2.3/10.7)	9.03 (7.2/25.1)	3.30 (2.4/11.0)	9.12 (7.2/26.1)	3.34 (2.3/11.4)	9.28 (7.3/27.0)
S2: B1 + WFST	3.06 (2.3/9.4)	8.60 (7.1/22.2)	3.10 (2.3/9.6)	8.72 (7.1/22.5)	3.11 (2.3/9.7)	8.78 (7.2/22.8)	3.09 (2.3/9.6)	8.83 (7.2/22.9)
S3: S2 + DB-RNNT	**2.81 (2.2/7.4)**	**8.10 (7.0/17.7)**	**2.91 (2.3/8.1)**	**8.30 (7.1/19.1)**	**3.00 (2.3/8.5)**	**8.45 (7.1/20.5)**	**3.04 (2.3/8.9)**	**8.75 (7.3/21.8)**
B2: B1 + NNLM	2.79 (1.7/11.6)	7.35 (5.2/26.3)	2.79 (1.7/11.6)	7.35 (5.2/26.3)	2.79 (1.7/11.6)	7.35 (5.2/26.3)	2.79 (1.7/11.6)	7.35 (5.2/26.3)
S4: S3 + NNLM	2.28 (1.6/7.9)	6.50 (5.1/18.7)	2.35 (1.6/8.2)	6.64 (5.2/19.6)	2.40 (1.7/8.4)	6.72 (5.2/20.2)	2.41 (1.7/8.6)	6.81 (5.2/20.9)
S5: S3 + DB-NNLM	**1.98 (1.5/5.7)**	**5.86 (4.9/14.1)**	**2.09 (1.6/6.2)**	**6.09 (5.1/15.1)**	**2.14 (1.6/6.7)**	**6.35 (5.1/17.2)**	**2.27 (1.6/7.3)**	**6.58 (5.2/18.9)**

Table 1: *LibriSpeech results with different biasing list size N. Reported metrics are in the following format: WER (U-WER/B-WER).*

Le, Duc, et al. **"Contextualized Streaming End-to-End Speech Recognition with Trie-Based Deep Biasing and Shallow Fusion."** *Interspeech*. 2021.

Librispeech for contextual biasing

- Specifies a biasing list with N = {100, 500, 1000, 2000} distractors
- U-WER: **Unbiased WER** measured on words **NOT IN** the biasing List
- B-WER: **biased WER** measured on words **IN** the biasing list
- This will be further discussed in Section 4

Model	N = 100		N = 500		N = 1000		N = 2000	
	test-clean	*test-other*	*test-clean*	*test-other*	*test-clean*	*test-other*	*test-clean*	*test-other*
B1: RNNT Baseline	3.65 (2.4/14.1)	9.61 (7.2/30.6)	3.65 (2.4/14.1)	9.61 (7.2/30.6)	3.65 (2.4/14.1)	9.61 (7.2/30.6)	3.65 (2.4/14.1)	9.61 (7.2/30.6)
S1: B1 + DB-RNNT	3.11 (2.3/9.8)	8.79 (7.1/23.4)	3.24 (2.3/10.7)	9.03 (7.2/25.1)	3.30 (2.4/11.0)	9.12 (7.2/26.1)	3.34 (2.3/11.4)	9.28 (7.3/27.0)
S2: B1 + WFST	3.06 (2.3/9.4)	8.60 (7.1/22.2)	3.10 (2.3/9.6)	8.72 (7.1/22.5)	3.11 (2.3/9.7)	8.78 (7.2/22.8)	3.09 (2.3/	8.83
S3: S2 + DB-RNNT	**2.81 (2.2/7.4)**	**8.10 (7.0/17.7)**	**2.91 (2.3/8.1)**	**8.30 (7.1/19.1)**	**3.00 (2.3/8.5)**	**8.45 (7.1/20.5)**	**3.0 (2.3/**	
B2: B1 + NNLM	2.79 (1.7/11.6)	7.35 (5.2/26.3)	2.79 (1.7/11.6)	7.35 (5.2/26.3)	2.79 (1.7/11.6)	7.35 (5.2/26.3)	2.7 (1.7/11.6)	(5.2/26.3)
S4: S3 + NNLM	2.28 (1.6/7.9)	6.50 (5.1/18.7)	2.35 (1.6/8.2)	6.64 (5.2/19.6)	2.40 (1.7/8.4)	6.72 (5.2/20.2)	2.41 (1.7/8.6)	6.81 (5.2/20.9)
S5: S3 + DB-NNLM	**1.98 (1.5/5.7)**	**5.86 (4.9/14.1)**	**2.09 (1.6/6.2)**	**6.09 (5.1/15.1)**	**2.14 (1.6/6.7)**	**6.35 (5.1/17.2)**	**2.27 (1.6/7.3)**	**6.58 (5.2/18.9)**

Table 1: *LibriSpeech results with different biasing list size N. Reported metrics are in the following format: WER (U-WER/B-WER).*

Normal WER (U-WER/B-WER)

8.75 (7.3/21.8)

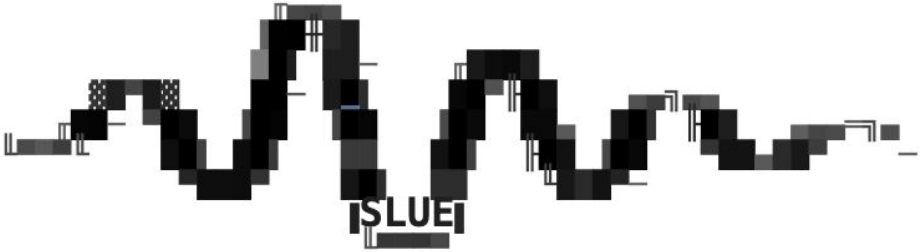
Le, Duc, et al. **"Contextualized Streaming End-to-End Speech Recognition with Trie-Based Deep Biasing and Shallow Fusion."** *Interspeech*. 2021.

SLUE (Spoken Language Understanding Evaluation)

- A benchmark suite for spoken language understanding on **natural** speech (not read or synthesized)
  - ASR, Named Entity Recognition (NER), and Sentiment Analysis (SA)
- SLUE-Voxpopli/SLUE-Voxcereb:
  - Built using freely available VoxCeleb and VoxPopuli datasets, with new transcriptions and annotations

Later extended to deal with

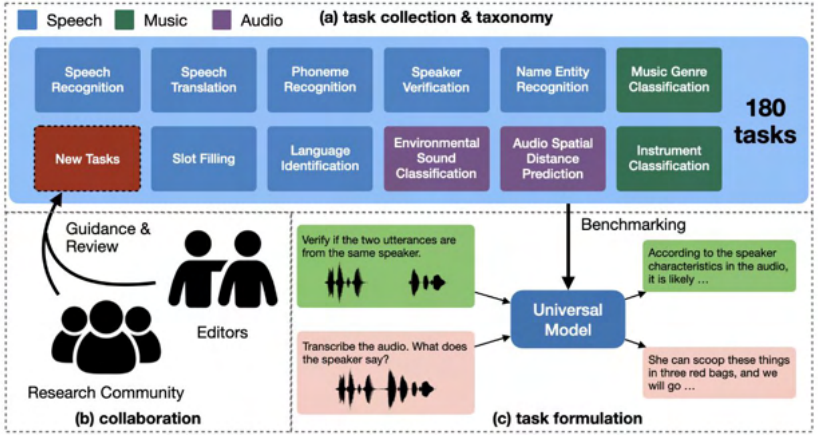
- Dialog act classification (DAC)
- Question answering (QA)
- Summarization (SUMM), and
- Named entity localization (NEL)



Shon, Suwon, et al. "**SLUE: New benchmark tasks for spoken language understanding evaluation on natural speech.**" *ICASSP* (2022)
Shon, Suwon, et al. **"SLUE Phase-2: A Benchmark Suite of Diverse Spoken Language Understanding Tasks."** *ACL* (2023).

Dynamic SUPERB (Speech processing Universal PERformance Benchmark)

- Focuses on **dynamic** updates with community collaboration to keep benchmarks relevant.
- Covers a **wide range of speech/audio tasks** including ASR, speaker recognition, and more.
- Provides an open framework for researchers to add new tasks and datasets
- **Over 70 contributors and 180 tasks**



Huang, Chien-yu, et al. **"Dynamic-superb: Towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech."** *ICASSP* (2024)

Huang, Chien-yu, et al. **"Dynamic-superb phase-2: A collaboratively expanding benchmark for measuring the capabilities of spoken language models with 180 tasks."** *ICLR* (2024).

Computer Science > Computation and Language

[Submitted on 8 Nov 2024 (v1), last revised 9 Jun 2025 (this version, v2)]

Dynamic-SUPERB Phase-2: A Collaboratively Expanding Benchmark for Measuring the Capabilities Spoken Language Models with 180 Tasks

Chien-yu Huang, Wei-Chih Chen, Shu-wen Yang, Andy T. Liu, Chen-An Li, Yu-Xiang Lin, Wei-Cheng Tseng, Anuj Diwan, Yi-Jen Shih, Jiatong Shi, William Chen, Chih-Kai Yang, Wenze Ren, Xuanjun Chen, Chi-Yuan Hsiao, Puyuan Peng, Shih-Heng Wang, Chun-Yi Kuan, Ke-Han Lu, Kai-Wei Chang, Fabian Ritte Gutierrez, Kuan-Po Huang, Siddhant Arora, You-Kuan Lin, Ming To Chuang, Eunjung Yeo, Kalvin Chang, Chung-Ming Chien, Kwanghee Choi, Jun-You Wa Cheng-Hsiu Hsieh, Yi-Cheng Lin, Chee-En Yu, I-Hsiang Chiu, Heitor R. Guimarães, Jionghao Han, Tzu-Quan Lin, Tzu-Yuan Lin, Homu Chang, Ting-Wu Ch Chun Wei Chen, Shou-Jen Chen, Yu-Hua Chen, Hsi-Chun Cheng, Kunal Dhawan, Jia-Lin Fang, Shi-Xin Fang, Kuan-Yu Fang Chiang, Chi An Fu, Hsien-Fu H Ching Yu Hsu, Shao-Syuan Huang, Lee Chen Wei, Hsi-Che Lin, Hsuan-Hao Lin, Hsuan-Ting Lin, Jian-Ren Lin, Ting-Chun Liu, Li-Chun Lu, Tsung-Min Pai, Ankita Pasad, Shih-Yun Shan Kuan, Suwon Shon, Yuxun Tang, Yun-Shao Tsai, Jui-Chiang Wei, Tzu-Chieh Wei, Chengxi Wu, Dien-Ruei Wu, Chao-Han Huc Yang, Chieh-Chi Yang, Jia Qi Yip, Shao-Xiang Yuan, Vahid Noroozi, Zhehuai Chen, Haibin Wu, Karen Livescu, David Harwath, Shinji Watanabe, Hung-yi Le

Multimodal foundation models, such as Gemini and ChatGPT, have revolutionized human-machine interactions by seamlessly integrating various forms of data. Developing a universal spoken language model that comprehends a wide range of natural language instructions is critical for bridging communication gaps and facilitating more intuitive interactions. However, the absence of a comprehensive evaluation benchmark poses a significant challenge. We present Dynamic-SUPERB Phase-2, an open and evolving benchmark for the comprehensive evaluation of instruction-based universal speech models. Building upon the first generation, this second version incorporates 125 new tasks contributed collaboratively by the global research community, expanding the benchmark to a total of 180 tasks, making it the largest benchmark for speech and audio evaluation. While the first generation of Dynamic-SUPERB was limited to classification tasks, Dynamic-SUPERB Phase-2 broadens its evaluation capabilities by introducing a wide array of novel and diverse tasks, including regression and sequence generation, across speech, music, and environmental audio. Evaluation results show that no model performed well universally. SALMONN-13B excelled in English ASR and Qwen2-Audio-7B-Instruct showed high accuracy in emotion recognition, but current models still require further innovations to handle a broader range of tasks. We open-source all task data and the evaluation pipeline at this https URL.

Dynamic SUPERB (Speech processing Universal PERformance Benchmark)

Figure 4: Task taxonomy of speech.

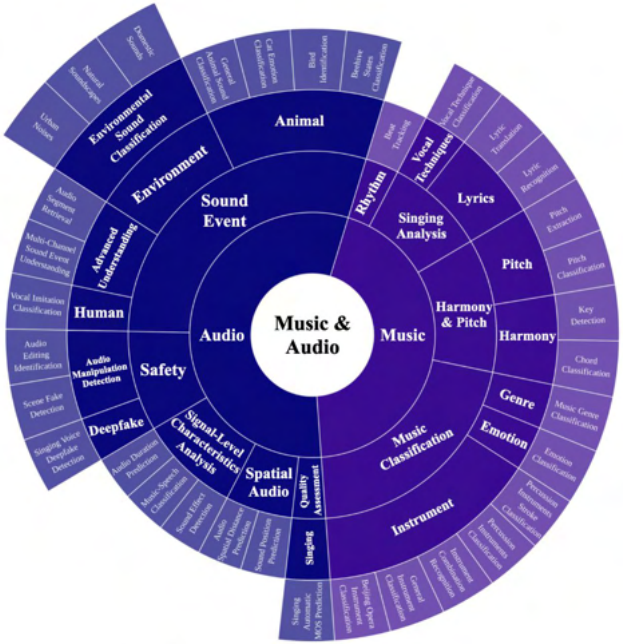


Figure 5: Task taxonomy of audio and music.

Huang, Chien-yu, et al. **"Dynamic-superb: Towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech."** *ICASSP* (2024)

Huang, Chien-yu, et al. **"Dynamic-superb phase-2: A collaboratively expanding benchmark for measuring the capabilities of spoken language models with 180 tasks."** *ICLR* (2024).

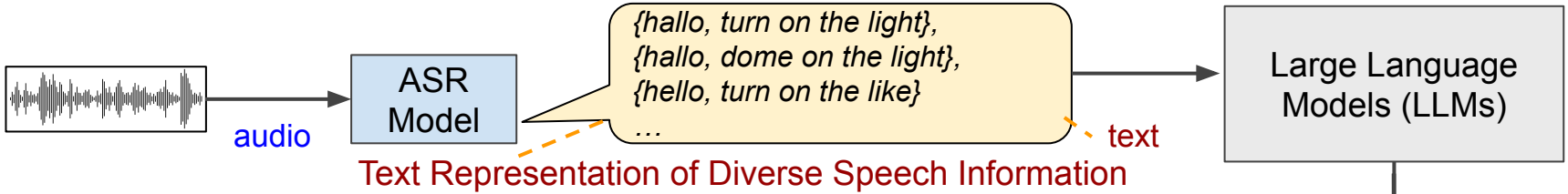
Speech QA Benchmarks

- Spoken SQuAD
  - **Synthesized** speech version of the SQuAD reading comprehension dataset.

- SLUE phase 2
  - Crowdsourcing platform for collecting spoken questions **read** by human speakers
  - 5 QA text database, including SQuAD
  - Frame F-1 score

SLT 24 Post-ASR Multi-Task LM Benchmark (1/2)

- **Motivation**: Could ASR-Decoding **Text secretly contains Acoustic information**?
  - We design three new post-ASR LM tasks given decoding text as inputs



Task 1 Post-ASR LM Words Correction (Acoustics / Semantics)

task-instruction

Task 2 Post-ASR LM for Speaker Tagging (Acoustics / Semantics)

task-instruction

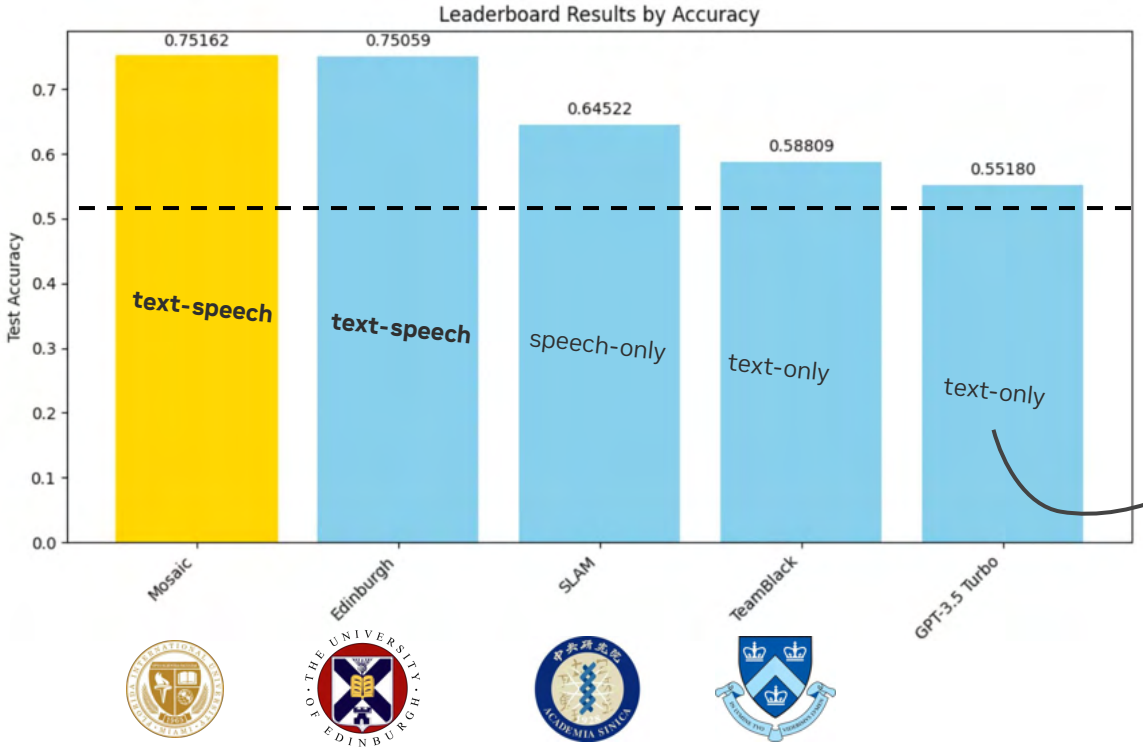
Task 3 Post-ASR LM for Emotion Recognition (Acoustics / Semantics)

task-instruction

IEEE SLT 24 “LLM Based Generative Error Correction: A Challenge and Baselines for Speech Recognition, Speaker Tagging, and Emotion Recognition”

SLT 24 Post-ASR Task 3 Emotion Understanding (2/2)

- **Key finding**: text-only LM can identify emotion from ASR systems' hypotheses
  - Text model with speech information injection archived the best results



ICASSP 23 HuBERT 300M baseline

text-only GPT-3.5 baseline

→ Leaderboard
huggingface.co/spaces/GenSEC-LLM/Post-ASR-N-Best-Text-Speech-Emotion-Recognition

Takeaways

- Evolution in speech-to-text systems
  - From short-form ASR to contextual conversation understanding
- Evaluation metrics
  - Extended from WER
  - Semantic metrics
- Benchmarks
  - Handles long-form speech
  - Some are redesigned for long-form speech

To accelerate this research direction

- **We need more realistic, conversational, and preferably large-scale public datasets!**

Table of Contents

Download Slides

Introduction (10 mins) 15:30-15:40	
Shinji (30 min) 15:40-16:10	Speech-to-Text Benchmark
Taejin (40 min) 16:10-16:50	**Leveraging Long Acoustic Context**
Recess (10 min) 16:50-17:00	
Huck (40 min) 17:00-17:40	Semantic Context and Speech-Language Modeling
Kyu (30 min) 17:40-18:10	Contextual Biasing and Methods Leveraging Longer Semantic Context for Speech Systems
Closing Remark (10 min) 18:10-18:20	
Q&A Session (10 min) 18:20-18:30	

Leveraging Long Acoustic Context

Taejin Park

Taejin Park

Leveraging Long Acoustic Context

1. **The Dynamics of Acoustic Context in End-to-End ASR**
   a. Speech Variability as a Source of Acoustic Context
   b. Acoustic Context Experiments on ASR Systems

2. **Speaker-attributed ASR and Acoustic Context**
   a. Acoustic Context Problem in Multispeaker ASR
   b. Multi-talker ASR
   c. Target speaker ASR

3. **Streaming with Long Acoustic Context**
   a. Streaming Speaker Attributed ASR
   b. Offline models to streaming
      i. Wait-K and others
      ii. ASR + MT: AST and Streaming ST

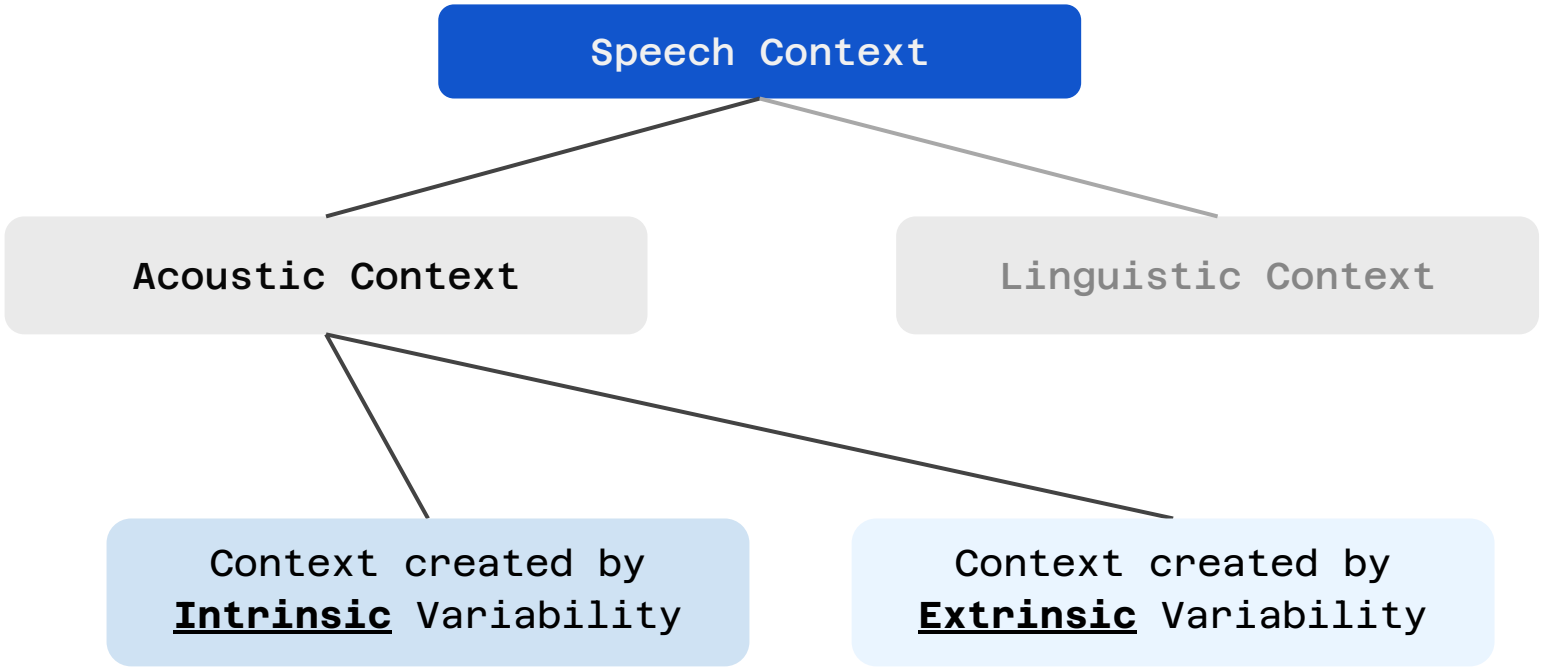
4. **Alternate Architectures for Long Context ASR**
   a. Variants of Transformer and Transducer
   b. State space model for speech recognition

5. **Voice-agent, Speech-LLM and Long context**
   a. Modular (EOU-ASR + LLM + TTS) Systems and Long Context
   b. Jointly trained duplex Speech LMs

The Dynamics of Acoustic Context in End-to-End ASR

The Dynamics of Acoustic Context in End-to-End ASR

- Speech Variability as a Source of Acoustic Context

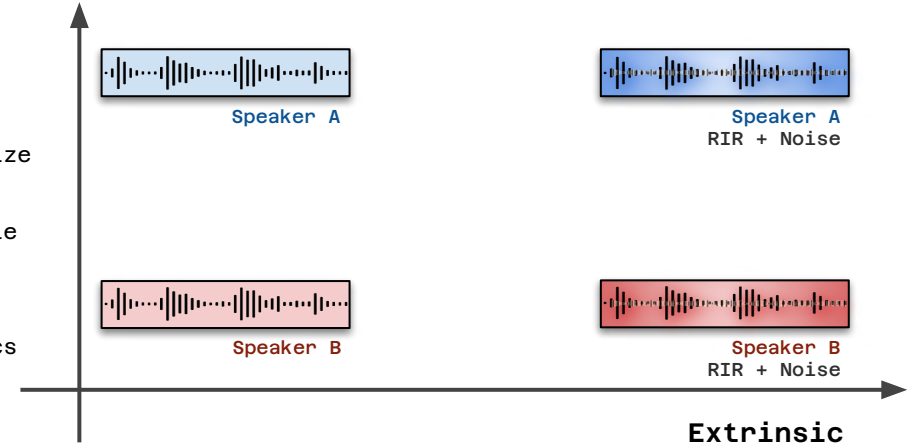


The Dynamics of Acoustic Context in End-to-End ASR

- Speech Variability as a Source of Acoustic Context

Intrinsic Variability

- **Physiological**: Vocal tract size
- Speech rate
- Emotional state
- Emphasis
- Speaker Characteristics



Extrinsic Variability

- **Environmental**: Background noise, reverberation, channel distortion
- Microphone quality
- Codec

The Dynamics of Acoustic Context in End-to-End ASR

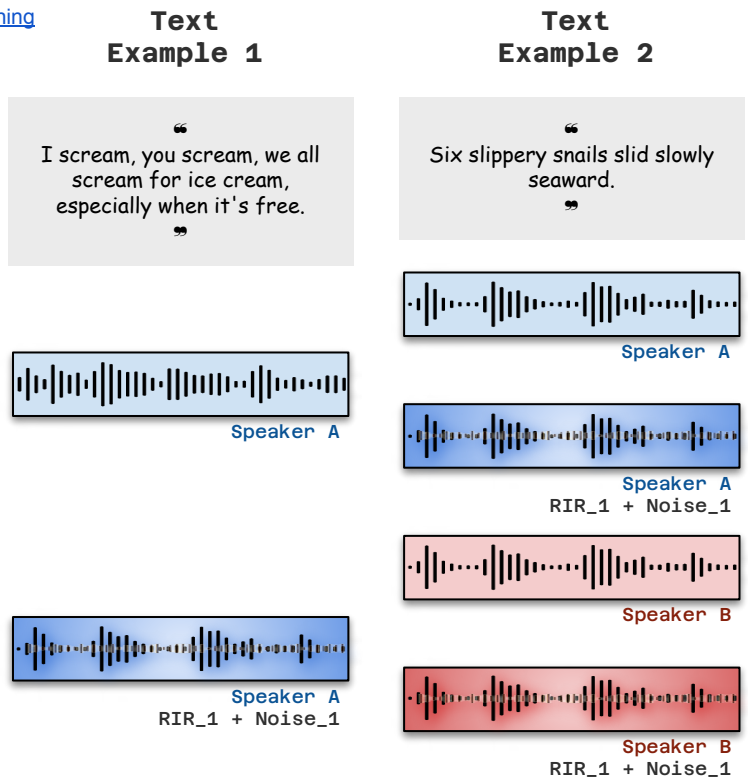
- **Acoustic Context Experiments on ASR Systems: Experiment Design**

Hugging Face Dataset Download: huggingface.co/datasets/taejinp/acoustic_context_switching

1. **Use TTS to generate speech samples**
   a. TTS enables Variable controlled experiments
   b. Created 384 utterances per each config
   c. Simulates intrinsic variability with multiple gender-balanced Speakers

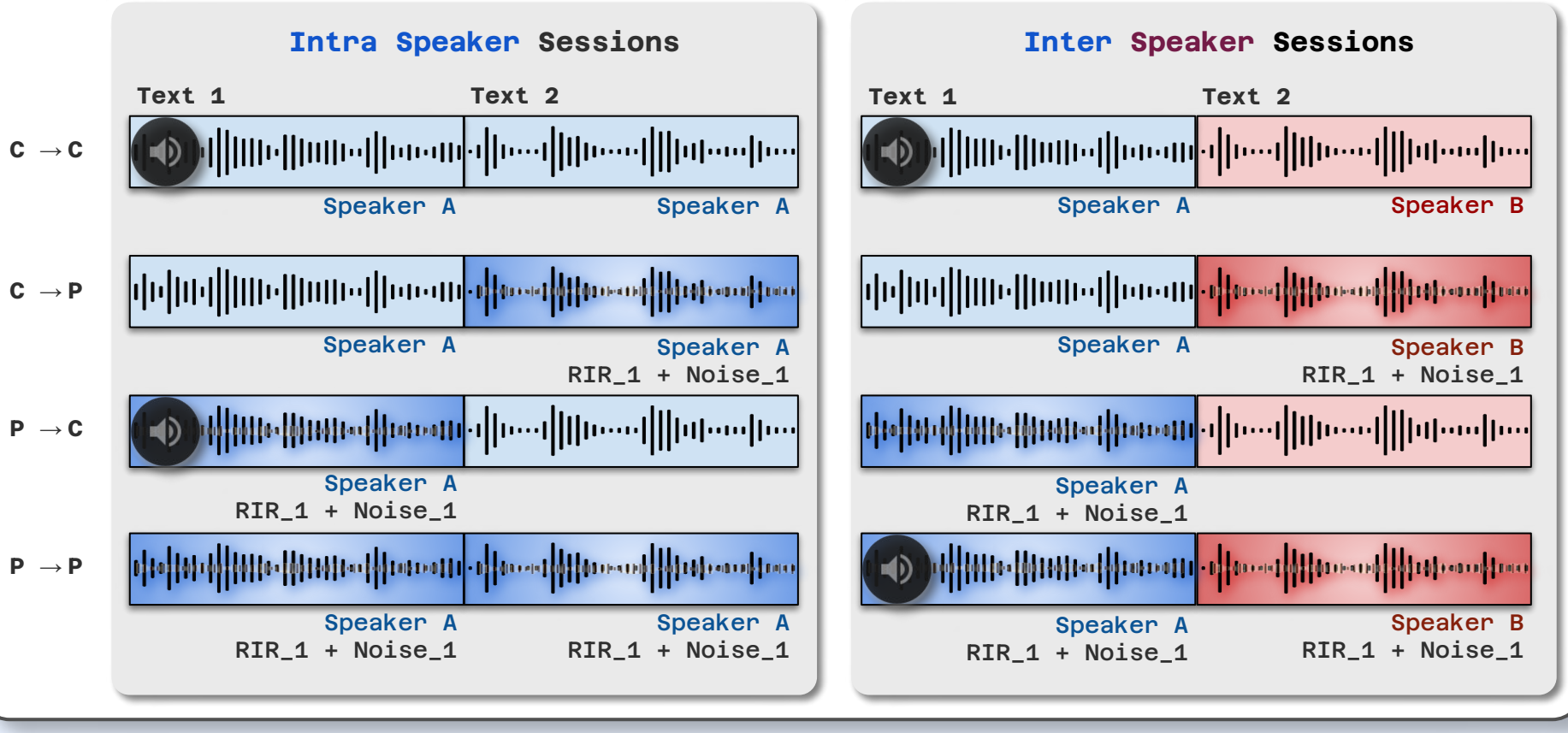
2. **Generate high-perplexity sentences to test Acoustic Modality**
   a. To minimize the effect of the internal language model in an End-to-end ASR model
   b. Tongue twisters and sophisticated words

3. **Room impulse response (RIR) and additive noise**
   a. To simulate extrinsic variabilities
   b. Harsh enough perturbations that drop WER



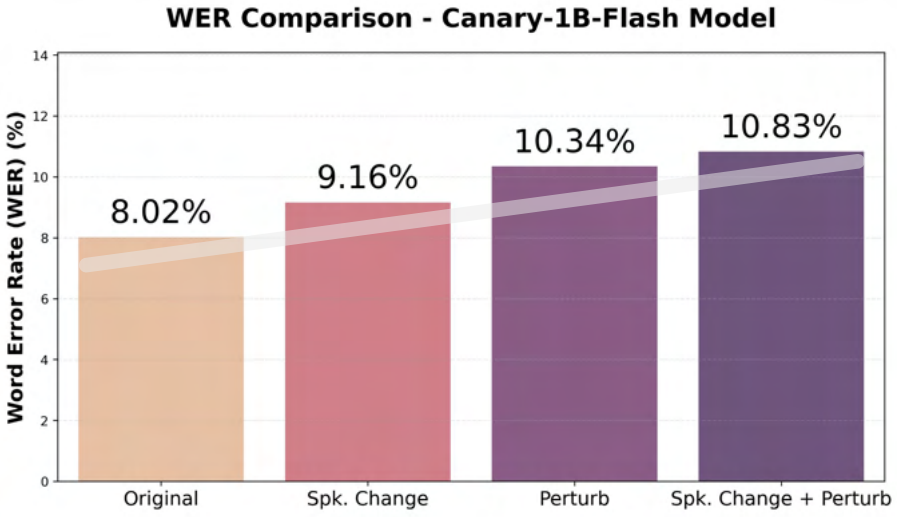
The Dynamics of Acoustic Context in End-to-End ASR

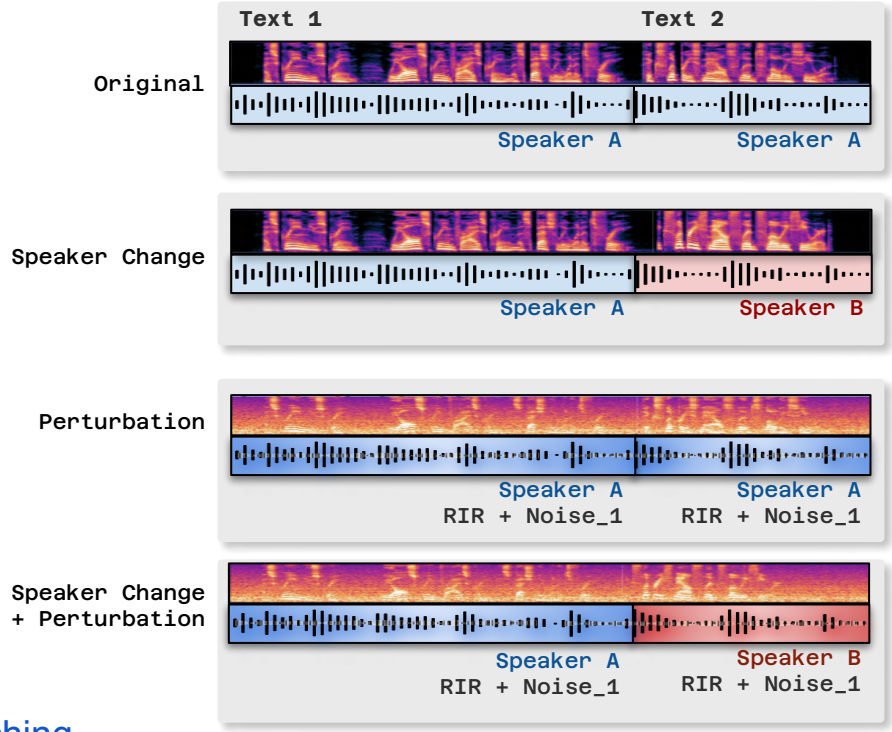
- Acoustic Context Experiments on ASR Systems: Example data points



The Dynamics of Acoustic Context in End-to-End ASR

- The Mechanism of Acoustic Contextual Information in ASR





Hugging Face Dataset Download:

huggingface.co/datasets/taejinp/acoustic_context_switching

The Dynamics of Acoustic Context in End-to-End ASR

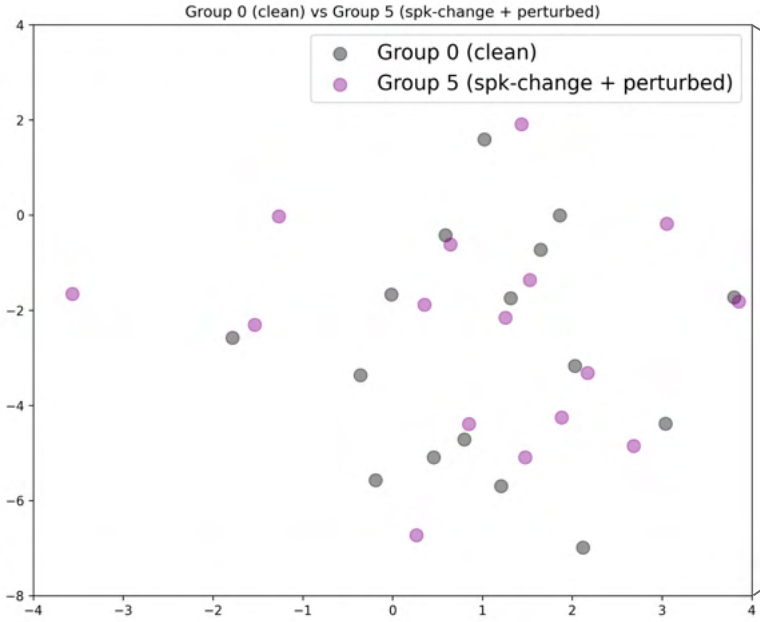
- Acoustic Context Experiments on ASR Systems

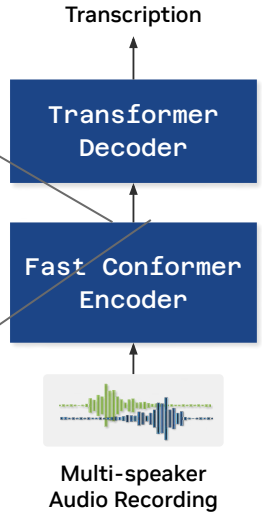
Clean

Clean vs Perturbation

Clean vs Speaker Change

Clean vs Speaker Change + Perturbation

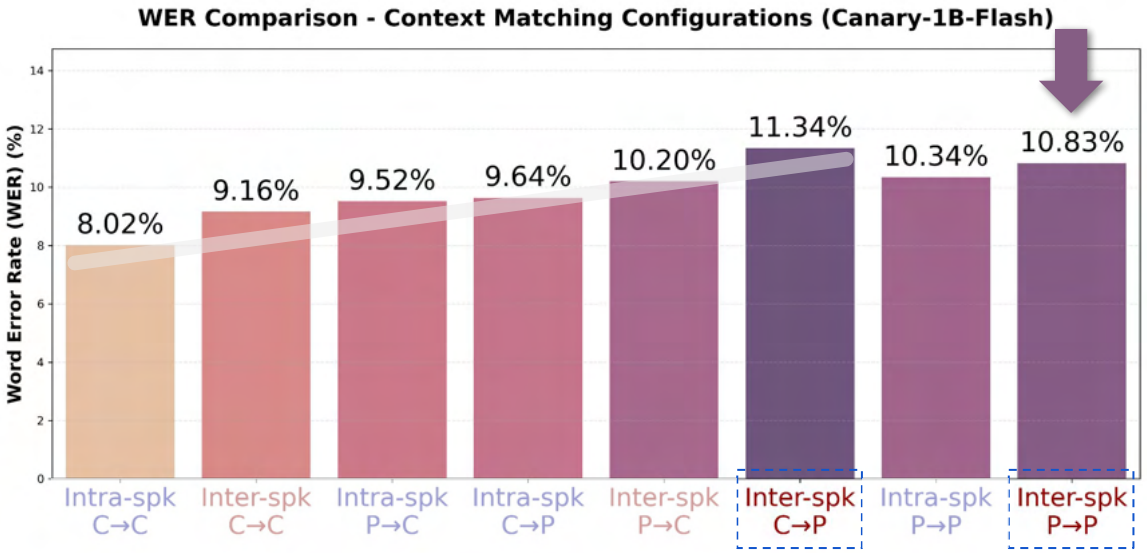


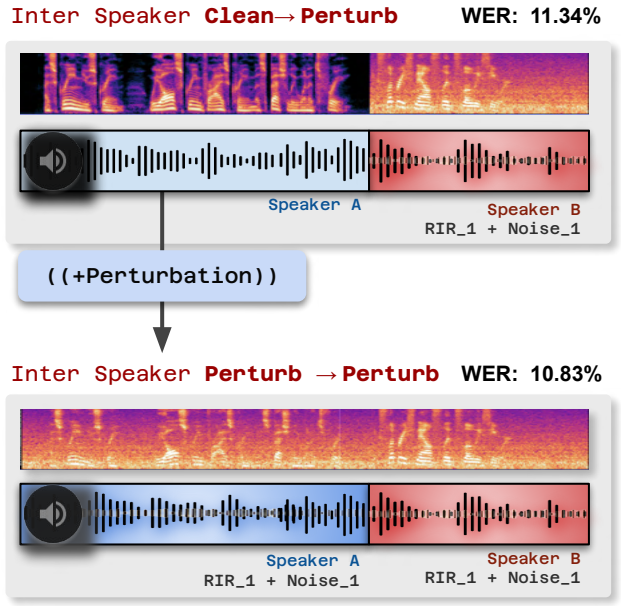


- **ASR encoder state shifts** and thus lead to WER drop.
- The shifted encoder state **will affect the future inputs** during decoding.

The Dynamics of Acoustic Context in End-to-End ASR

- Acoustic Context Experiments on ASR Systems





Inter Speaker Clean→ Perturb **WER** 11.34% **>** **Inter Speaker Perturb → Perturb** **WER** 10.83%

Add more perturbation then get a lower WER ?

Takeaways

1. **The sudden shift of acoustic context change leads to word error rate degradation.**

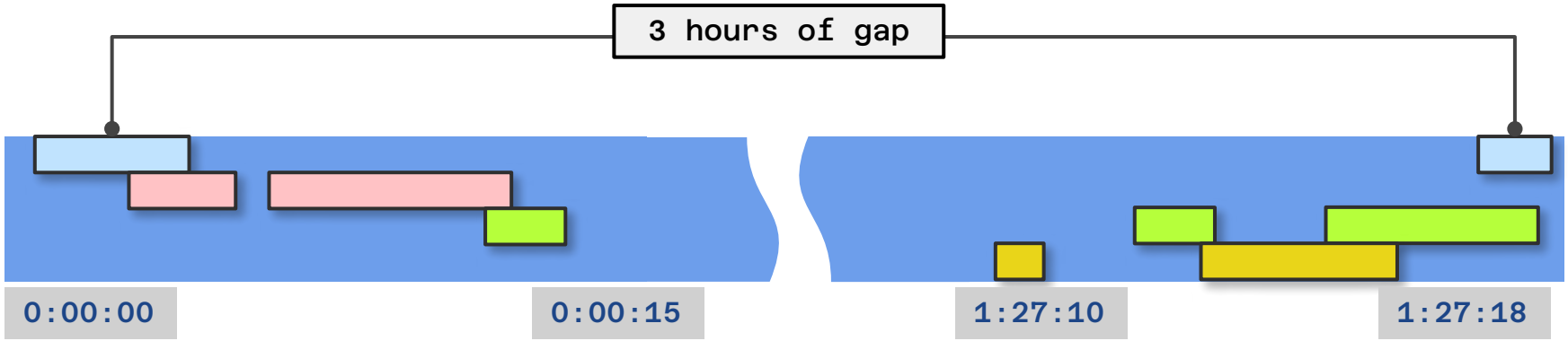
2. **Preconditioning with proper contextual information can be beneficial for ASR performance.**

3. **How can we leverage such behavior of end-to-end ASR model ?**
    a. For speaker-attributed ASR models (Multi-talker, target-speaker ASR models)
    b. For streaming ASR models

Speaker-attributed ASR and Acoustic Context

Speaker-attributed ASR and Long-Context

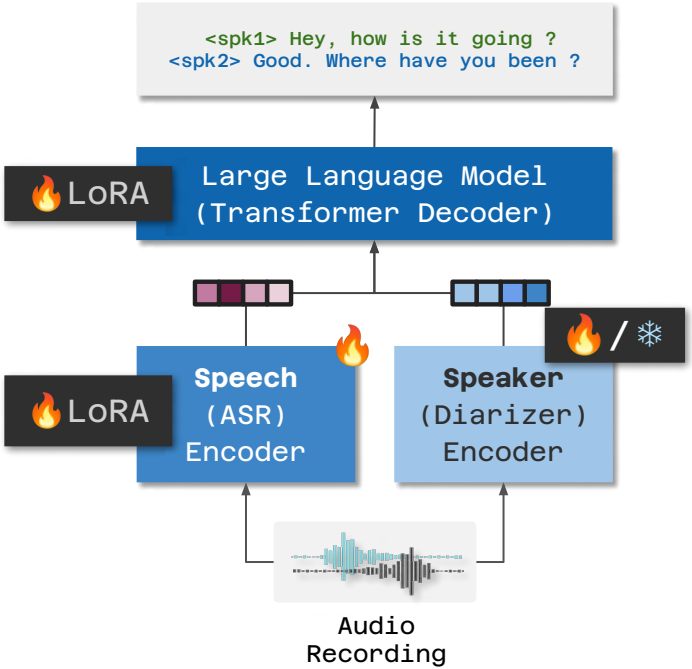
- Acoustic Context Problem in Multispeaker ASR



- Multispeaker ASR or Speaker Diarization need an **exceptionally long acoustic context**
- Imbalanced speaking time poses lack of context for certain speakers.

Speaker-attributed ASR and Long-Context

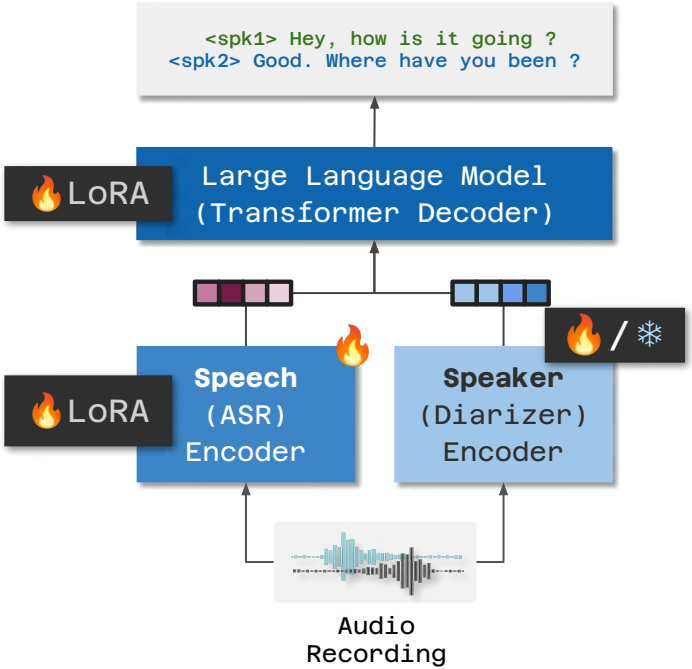
- Acoustic Context Problem in Multispeaker ASR



- The most recent multispeaker ASR systems share similarities:
  - Speech (ASR) Encoder part
  - Speaker (Diarization) encoder part
  - Adapter (LoRA) on Transformer Decoder
- Why separately...?

Speaker-attributed ASR and Long-Context

- Acoustic Context Problem in Multispeaker ASR

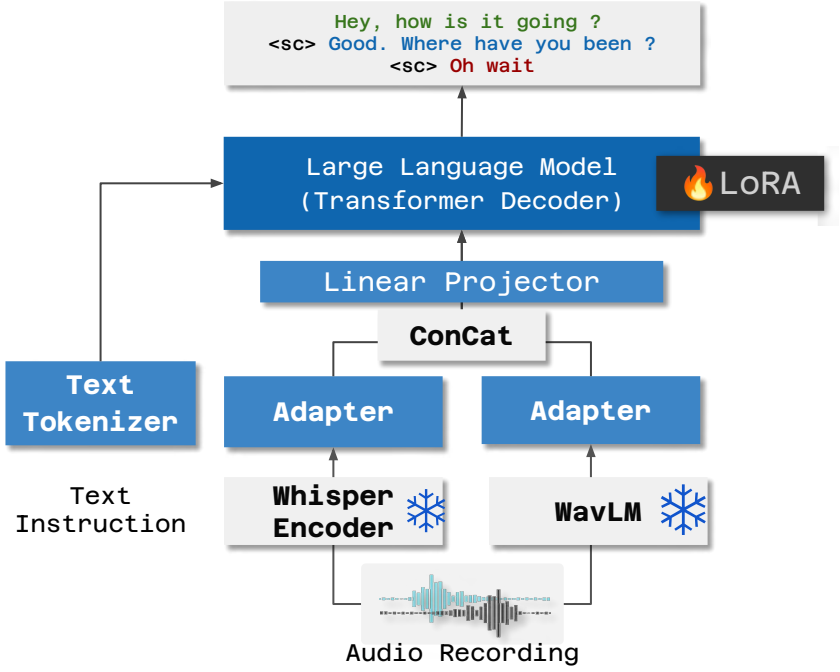


- Speech (ASR) encoder state minimizes:
  Variabilities other than **phonetic information**.
- Speaker encoder state minimizes:
  Variabilities other than **speaker information**.

- Speaker encoder and Speech encoder can be initialized from the same base speech encoder
- BUT, **universal embeddings fail to achieve SoTA.**
- In addition, for exceptionally long acoustic context, we need to save **speaker-cache separately.**

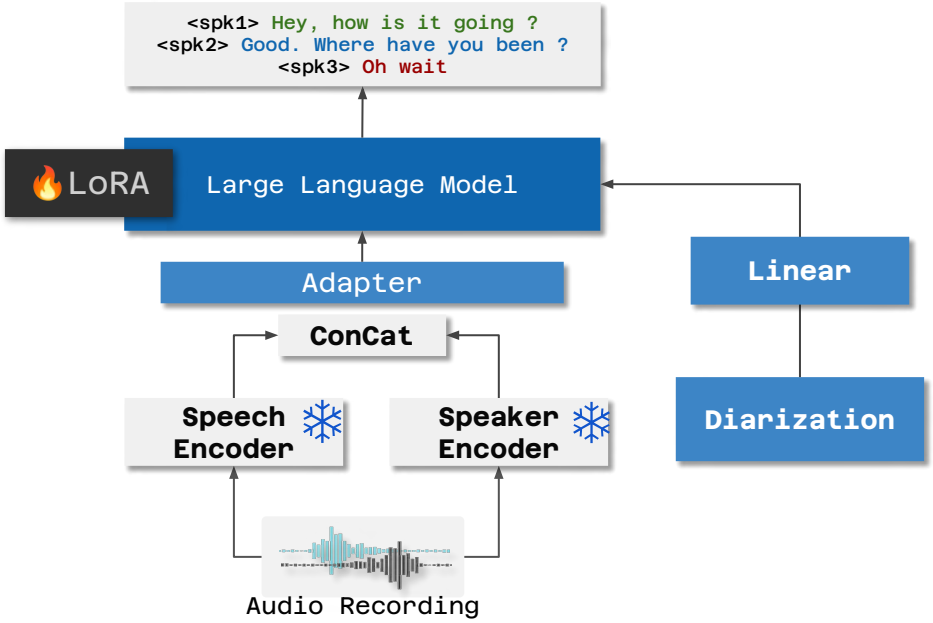
Speaker-attributed ASR and Long-Context

- Multi-speaker ASR with Speaker Context



MT-LLM: Multi-Talker LLLM

Meng, Lingwei, et al. **"Large language model can transcribe speech in multi-talker scenarios with versatile instructions."** *ICASSP 2025*

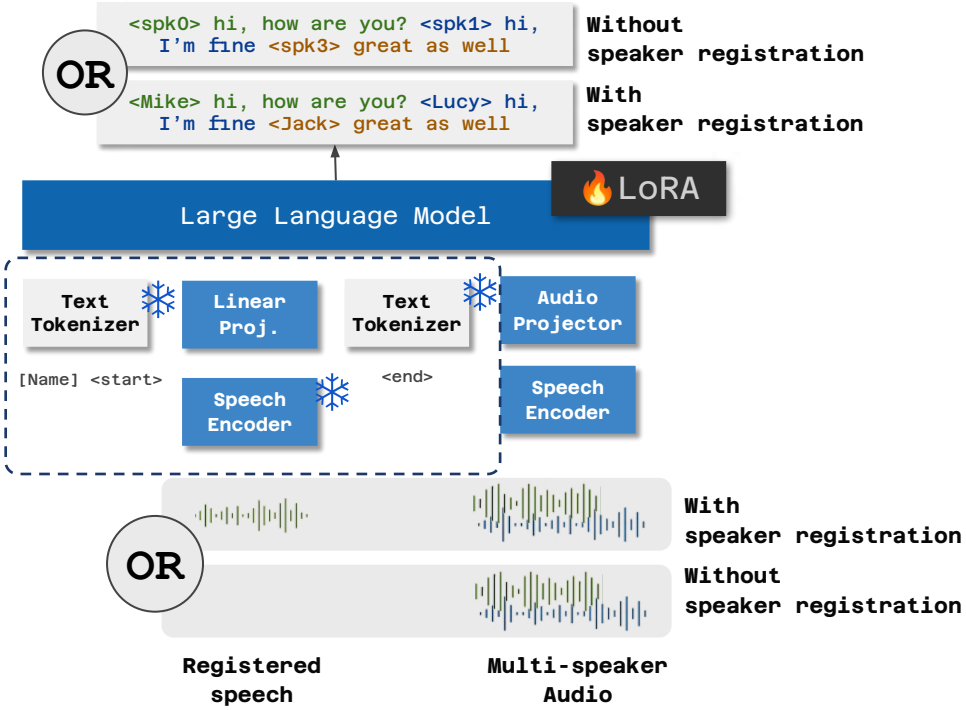


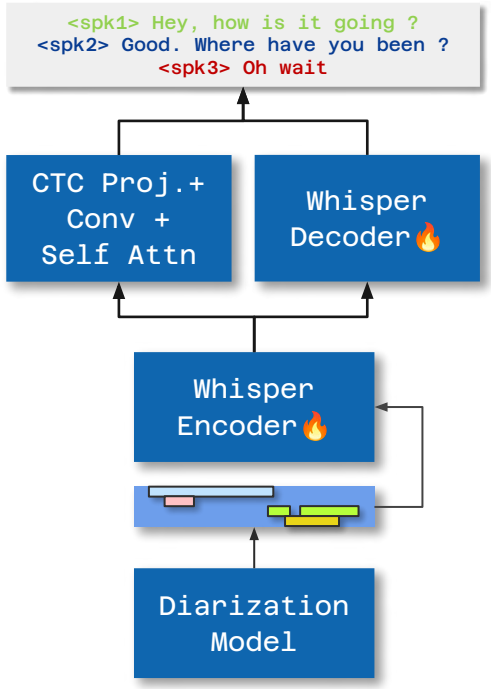
Diar-aware MS-ASR via LLM

Lin, Yuke, et al. **"Diarization-Aware Multi-Speaker Automatic Speech Recognition via Large Language Models."** *arXiv:2506.05796* (2025).

Speaker-attributed ASR and Long-Context

- Multi-speaker ASR with Speaker Context



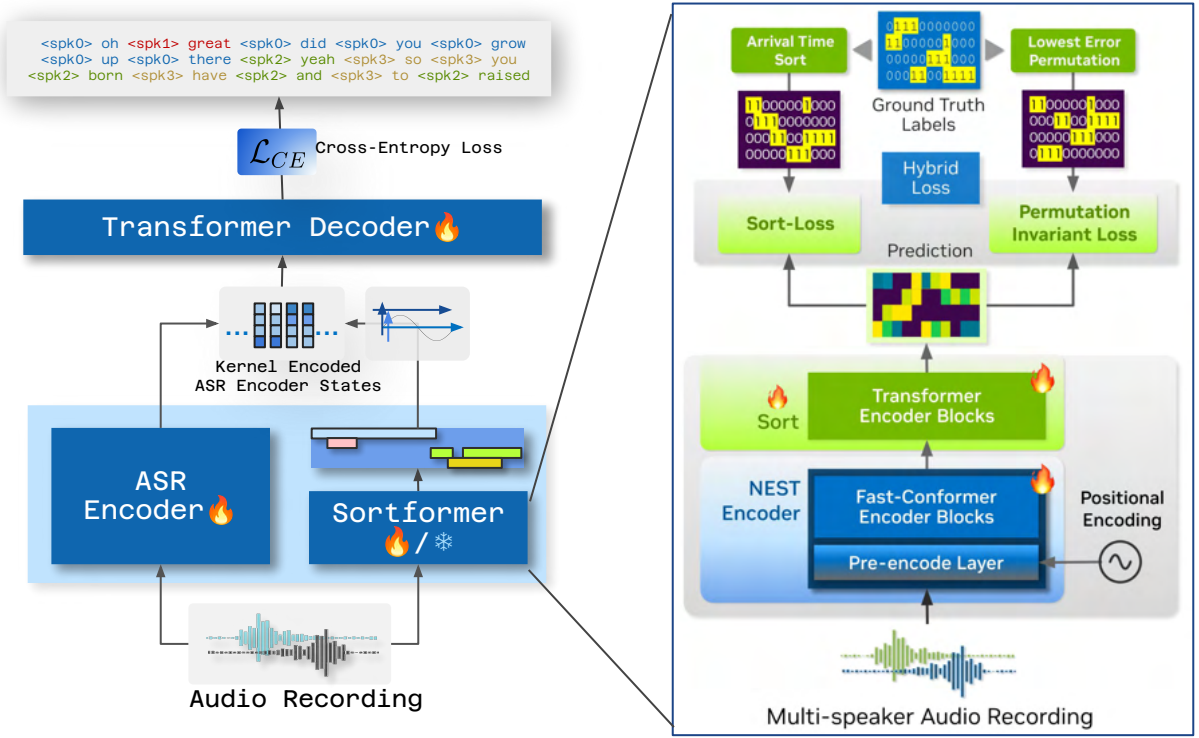


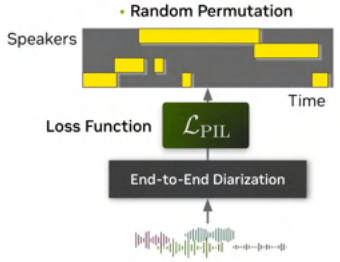
Yin, Han, et al. SpeakerLM: "**End-to-End Versatile Speaker Diarization and Recognition with Multimodal Large Language Models.**" 2025. arXiv:2508.0637 (2025).

Polok, Alexander, et al. **"DiCow: Diarization-conditioned whisper for target speaker automatic speech recognition."** *Computer Speech & Language* 95 (2026): 101841.

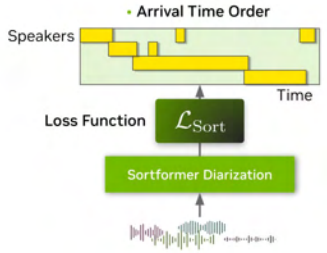
Speaker-attributed ASR and Long-Context

- Multi-speaker ASR with Speaker Context





Conventional E2E diarizers (e.g., EEND)

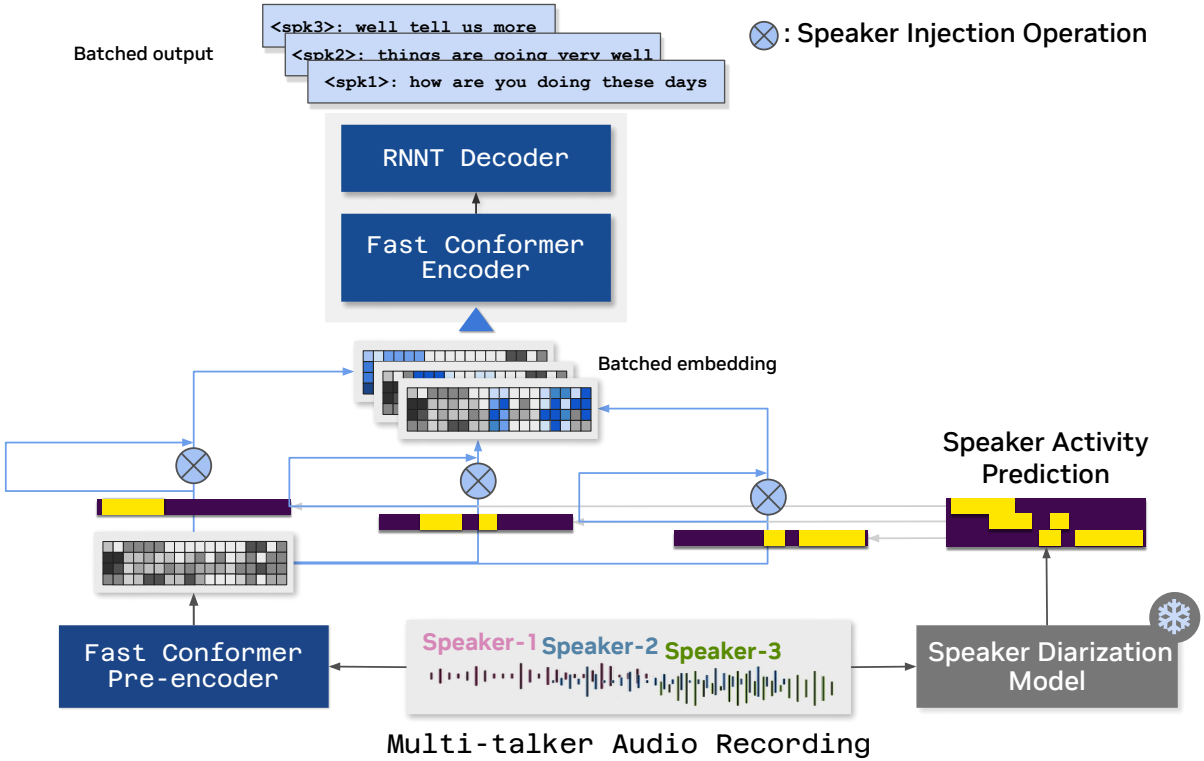


Sortformer Diarizer

Park, Taejin, Ivan Medennikov, Kunal Dhawan, Weiqing Wang, He Huang, et. al., **"Sortformer: A Novel Approach for Permutation-Resolved Speaker Supervision in Speech-to-Text Systems."** *Forty-second International Conference on Machine Learning (ICML 2025)*.

Speaker-attributed ASR and Long-Context

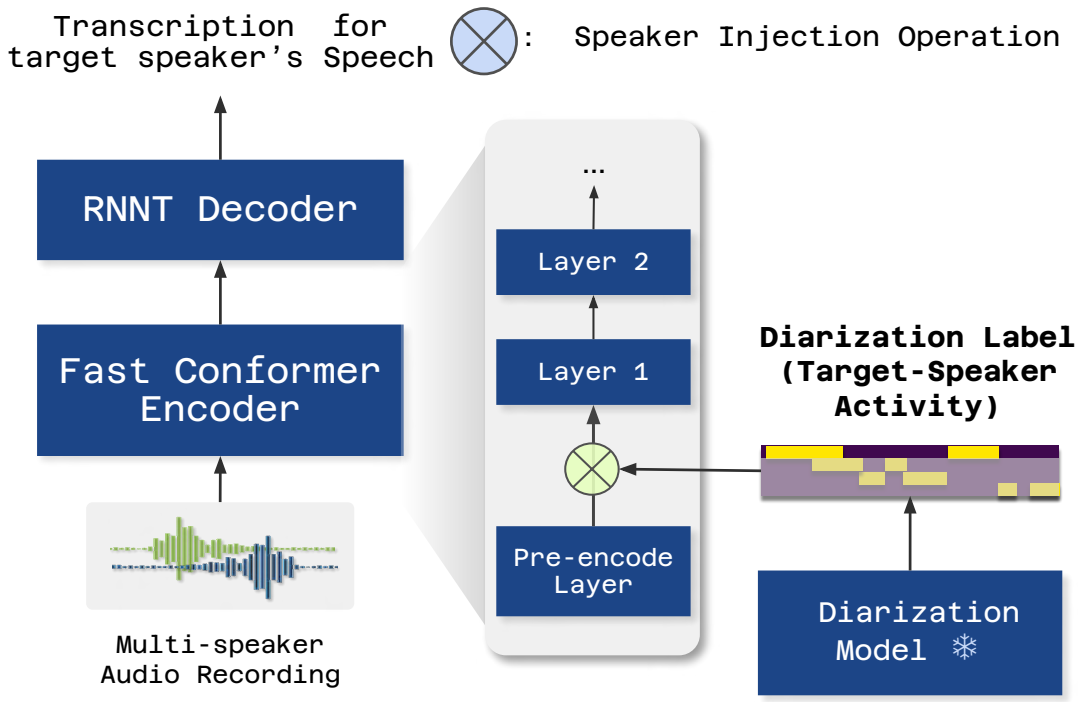
- Multi-speaker ASR with Speaker Context: Use Speaker (Acoustic) Context!



Wang, Weiqing, et al. **"Speaker Targeting via Self-Speaker Adaptation for Multi-talker ASR."** Interspeech 2025

Speaker-attributed ASR and Long-Context

- Multi-speaker ASR with Speaker Context: Batch the multiple speakers!



Wang, Weiqing, et al. **"Speaker Targeting via Self-Speaker Adaptation for Multi-talker ASR."** Interspeech 2025

Speaker-attributed ASR and Long-Context

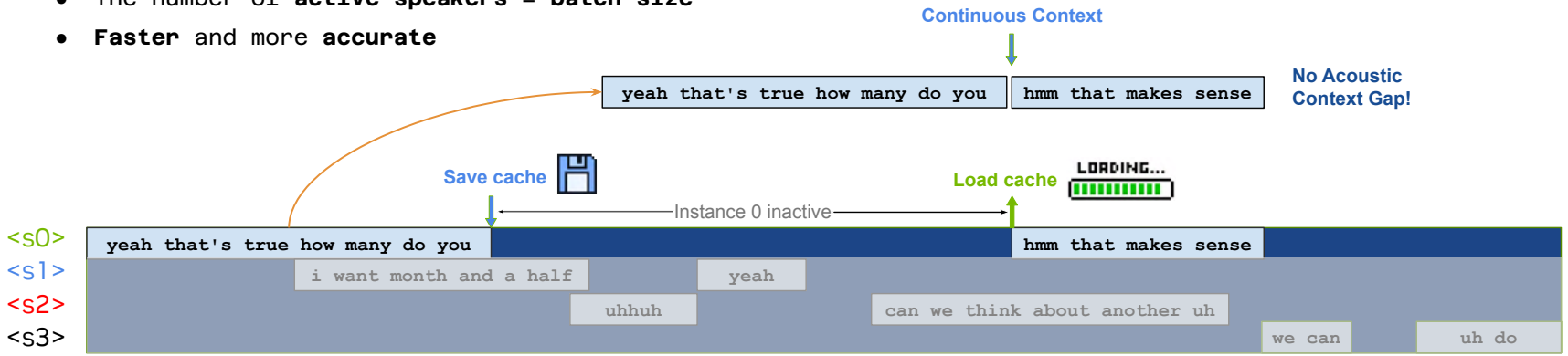
- Multi-speaker ASR with Speaker Context

Diarizatin Context Gating

- **Inefficient** when keeping 4 instance simultaneously
- Computations are **4x more than single speaker ASR**

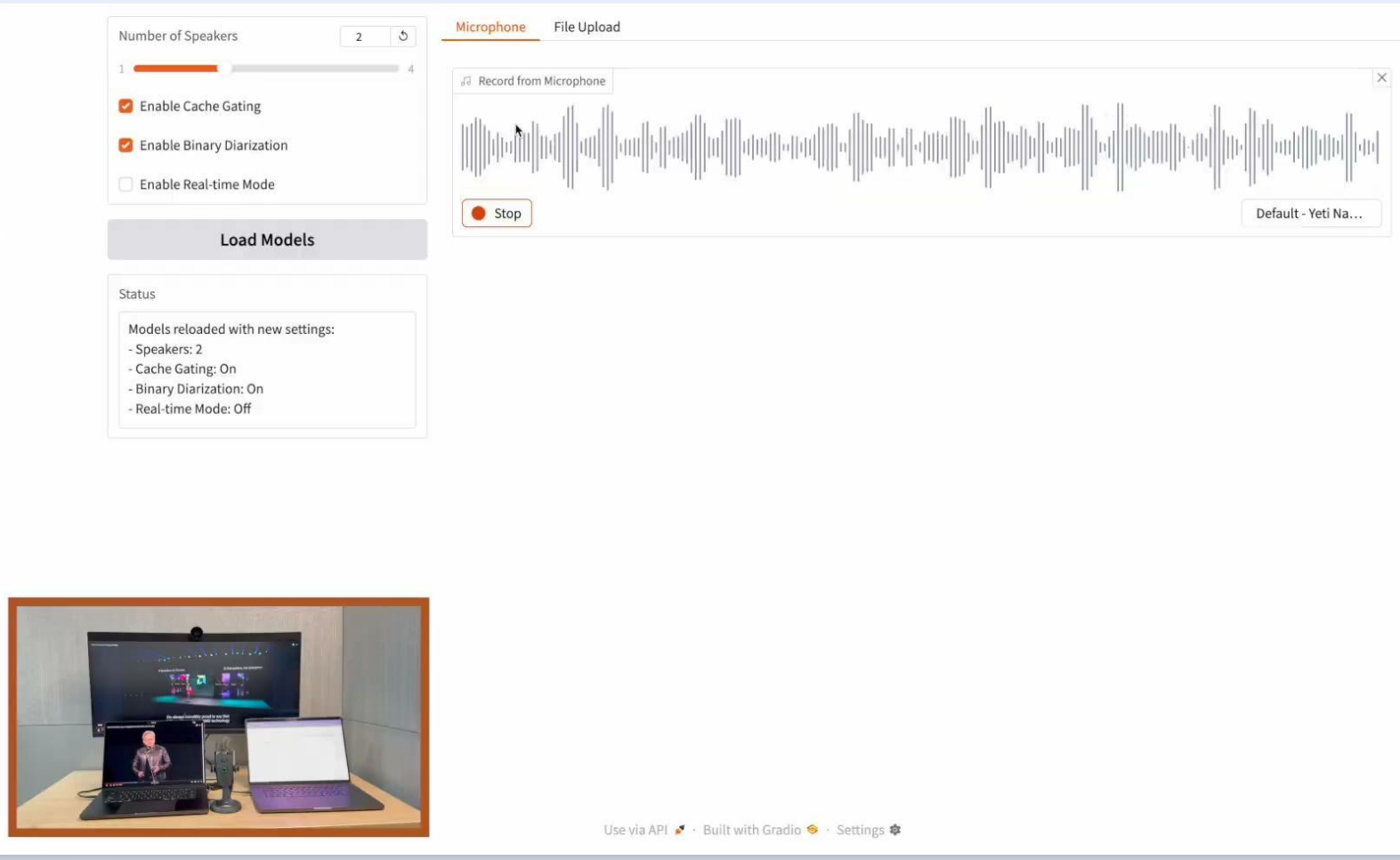
Solution ?

- **Save** speaker cache and **load** at the next speaker-active region
- The number of **active speakers** = **batch size**
- **Faster** and more **accurate**



Wang, Weiqing, et al. **"Speaker Targeting via Self-Speaker Adaptation for Multi-talker ASR."** Interspeech 2025

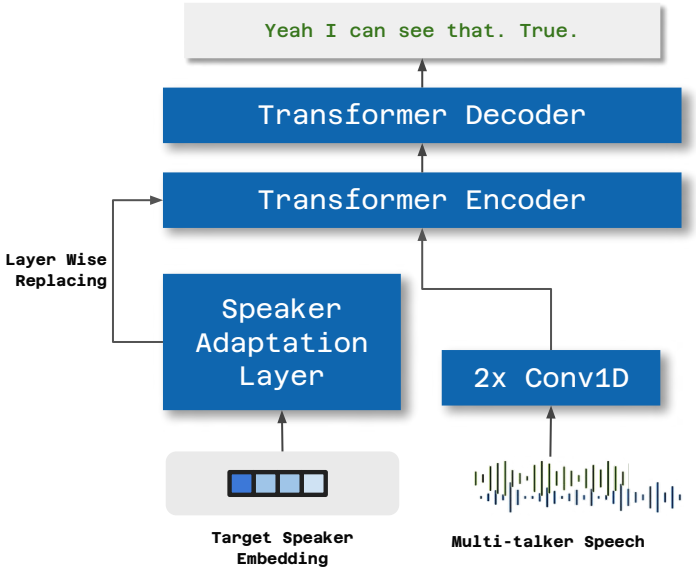
Speaker-attributed ASR and Long-Context



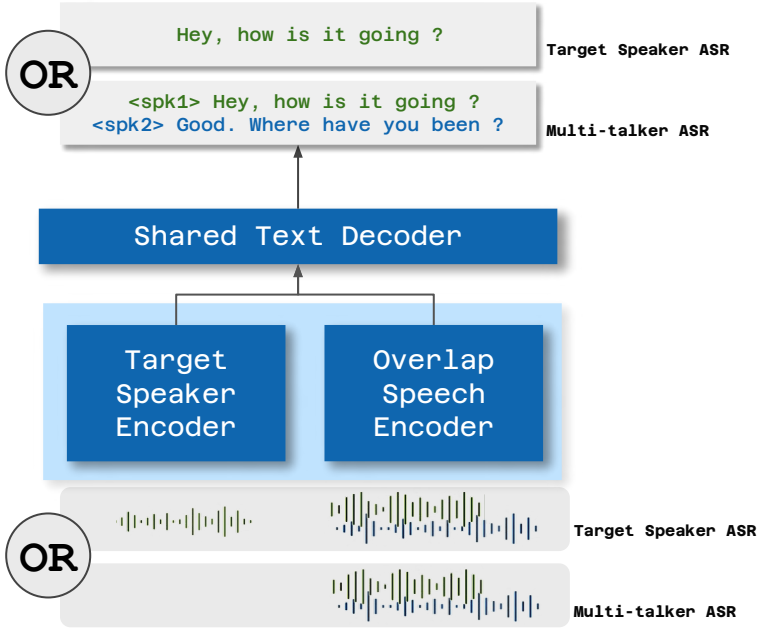
Speaker-attributed ASR and Long-Context

- Target speaker ASR with Speaker Context

Whisper with prompt:Target-speaker ASR [1]



Unified Multi-Talker/Target-speaker ASR [2]

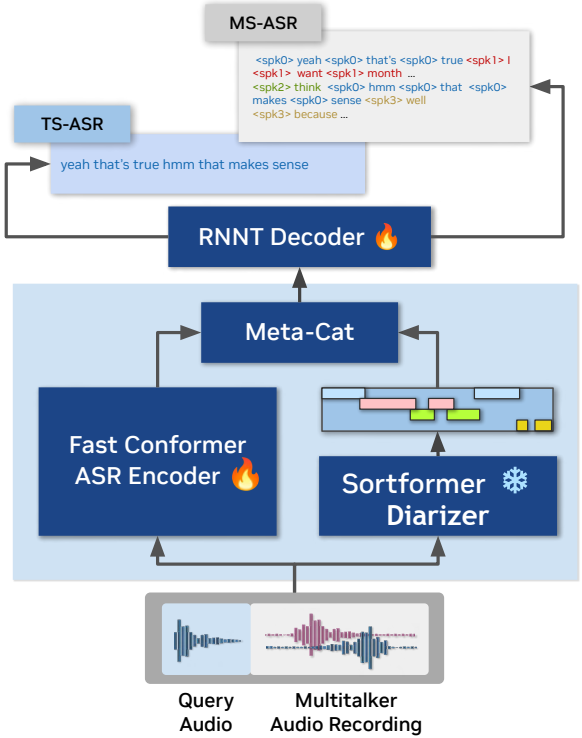


[1] Ma, Hao, et al. **"Extending Whisper with prompt tuning to target-speaker ASR."** *ICASSP 2024*

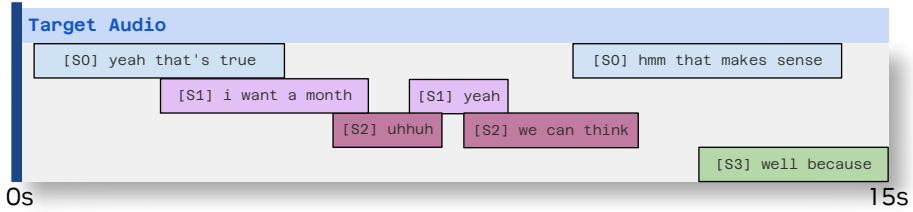
[2] Masumura, Ryo, et al. **"Unified multi-talker ASR with and without target-speaker enrollment."** *Interspeech*. Vol. 2024. 2024.

Speaker-attributed ASR and Long-Context

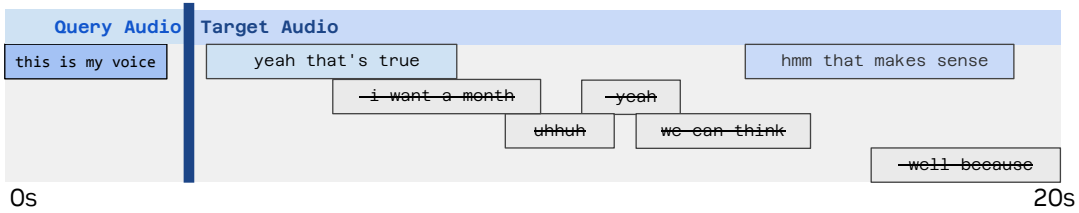
- Target speaker ASR with Speaker Context



(a) Multispeaker ASR Scenario



(b) Targetspeaker ASR Scenario



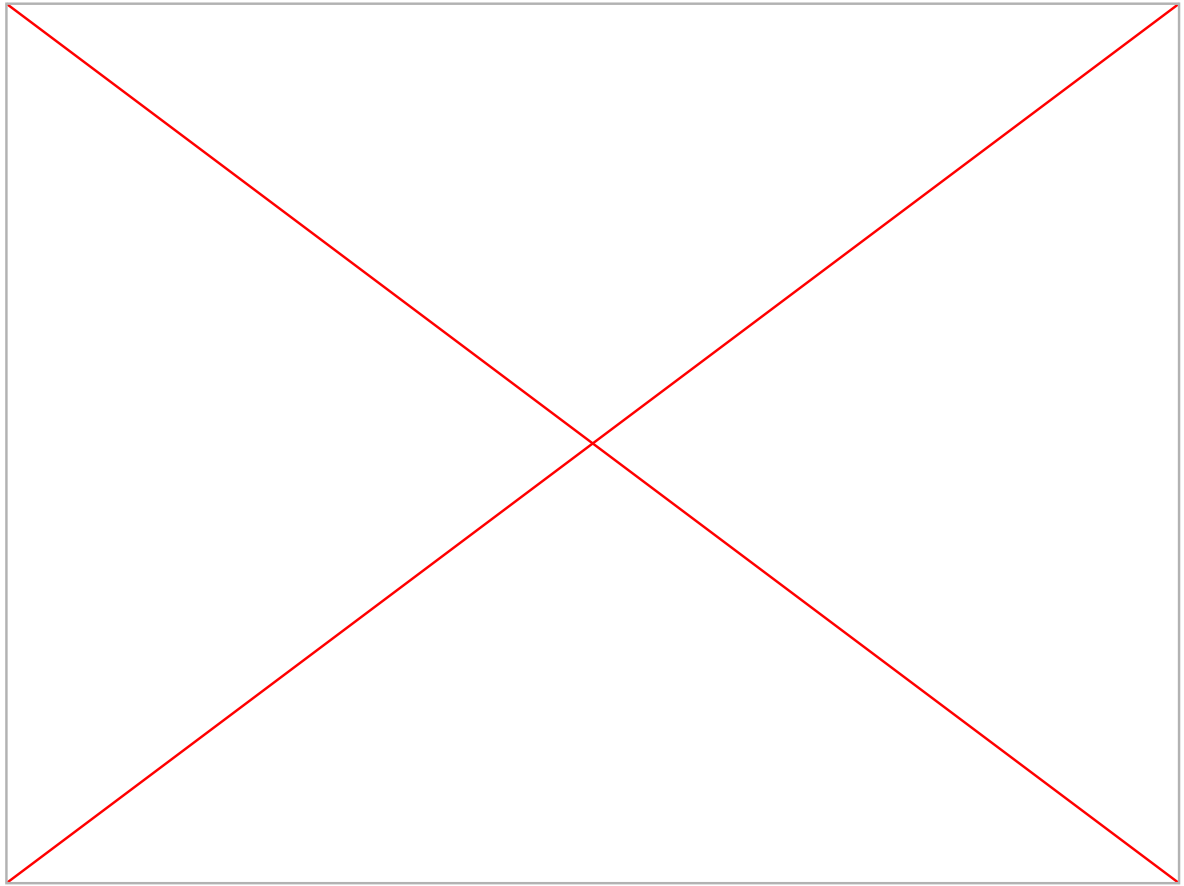
- Use Sortformer Diarizer: First speaker = query speaker.
- No architecture changes for performing target speaker ASR and multi-speaker ASR.

Wang, Jinhan, et al. **"META-CAT: Speaker-Informed Speech Embeddings via Meta Information Concatenation for Multi-talker ASR."** ICASSP 2025.

Speaker-attributed ASR and Long-Context

- Target speaker ASR

  Demo Video



Speaker-attributed ASR and Long-Context

- Takeaways

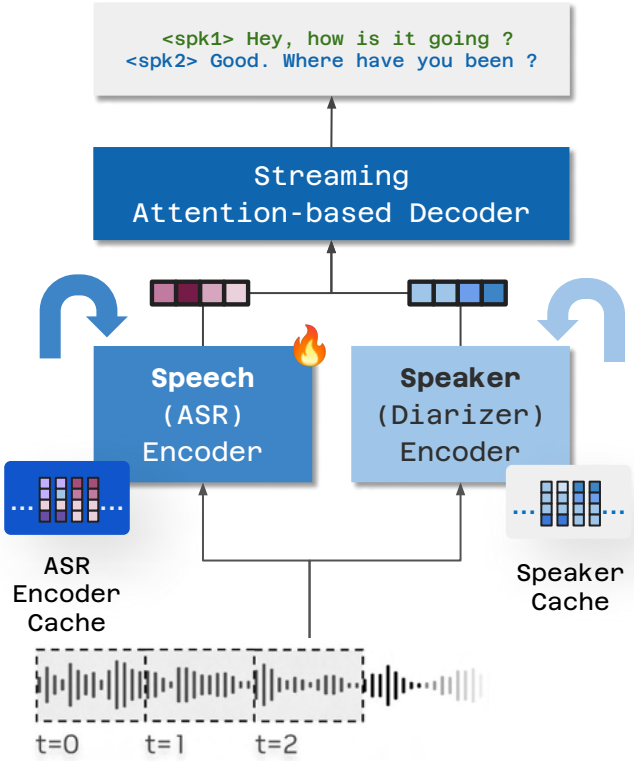
1. We can leverage the behavior of acoustic context biasing for multispeaker ASR and target speaker ASR.

2. Speaker cache should be handled separately from ASR cache for streaming and long form audio handling.

Streaming with Long Acoustic Context

Speaker-attributed ASR and Long-Context

- Streaming multispeaker ASR with Speaker Cache



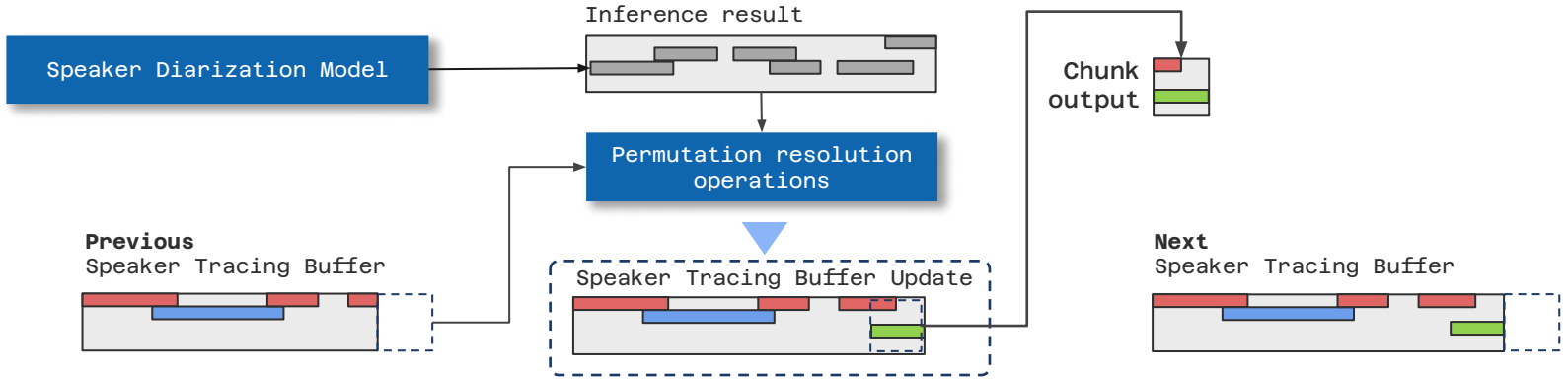
- Since we have separate encoders for both ASR and speaker diarization, we need to maintain **ASR cache and speaker cache separately**.

- The arrival time ordering of Sortformer makes speaker cache operating.

Streaming with Long Acoustic Context

- Streaming Speaker Diarization for SA-ASR: Speaker Tracing Buffer

  - How can we maintain speaker cache for diarization ?



  - "Permutation resolution operations" requires **cubic** time complexity
    n=(# of speakers)
  - Needs to maintain the whole history

Xue, Yawen, et al. **"Online end-to-end neural diarization with speaker-tracing buffer."** 2021 IEEE SLT Workshop.

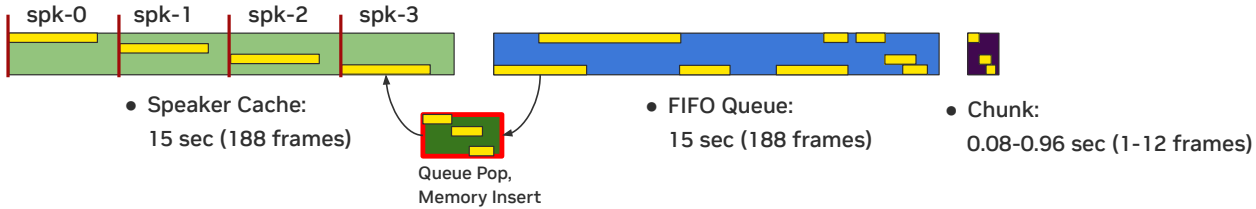
Horiguchi, Shota, et al. **"Online neural diarization of unlimited numbers of speakers using global and local attractors."** IEEE/ACM TASLP (2022): 706-720.

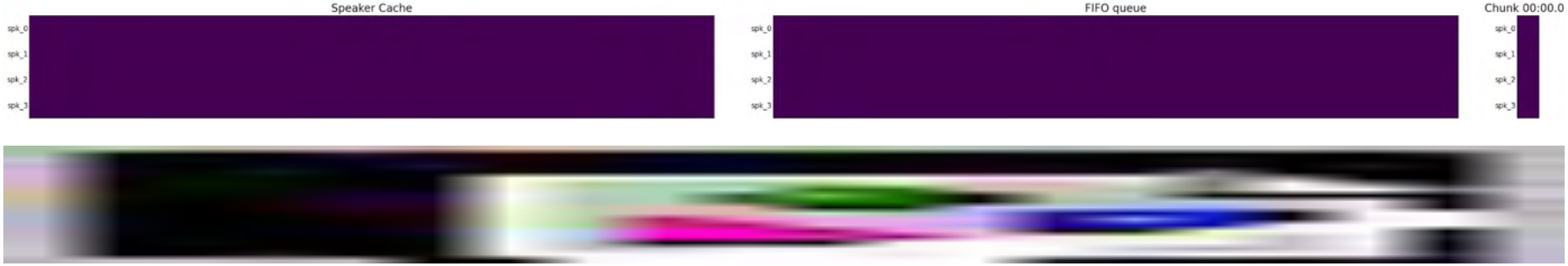
Streaming with Long Acoustic Context

- **Streaming Speaker Diarization for SA-ASR**

How it works: **Streaming Sortformer**

- Chunk-wise processing
- Arrival-Ordered Speaker Cache (AOSC) resolves between-chunk permutations
- FIFO queue adds recent context

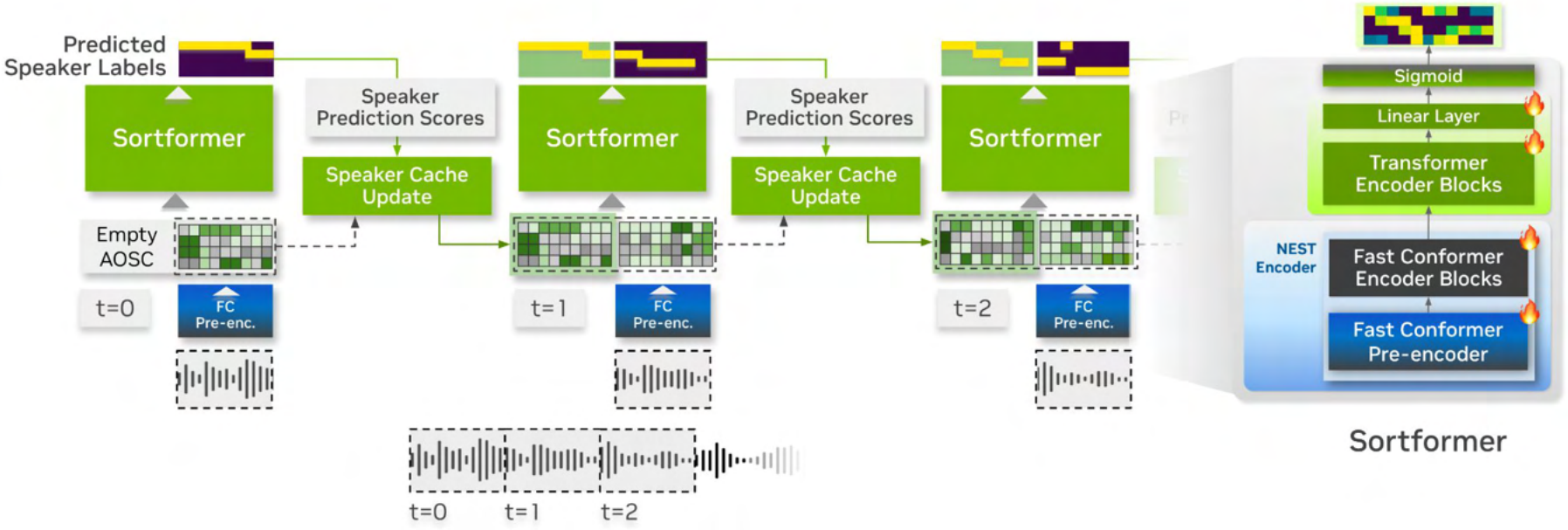




Medennikov, Ivan, et al. **"Streaming Sortformer: Speaker Cache-Based Online Speaker Diarization with Arrival-Time Ordering."** Interspeech 2025

Streaming with Long Acoustic Context

- Streaming Speaker Diarization for SA-ASR

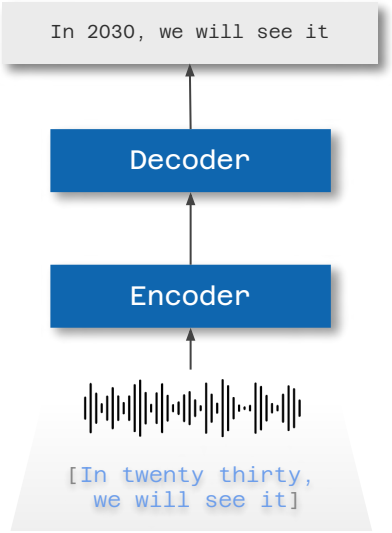


Medennikov, Ivan, et al. **"Streaming Sortformer: Speaker Cache-Based Online Speaker Diarization with Arrival-Time Ordering."** Interspeech 2025

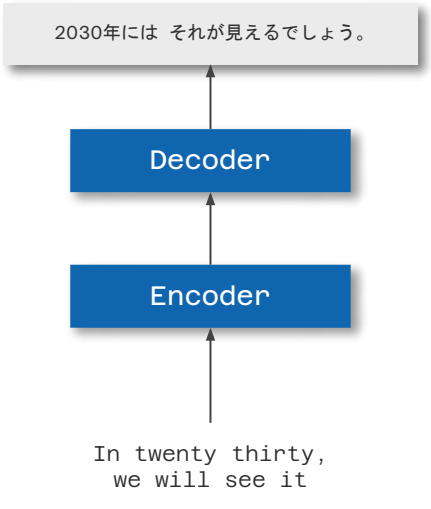
Streaming with Long Acoustic Context

- Offline models to streaming: ASR with Translation

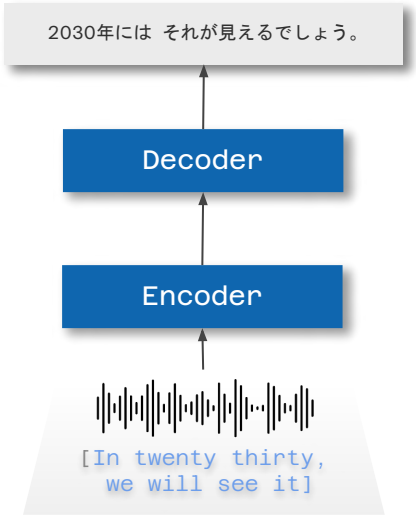
Beyond ASR: ASR vs MT vs AST



Automatic Speech Recognition (ASR)



Machine Translation (MT)

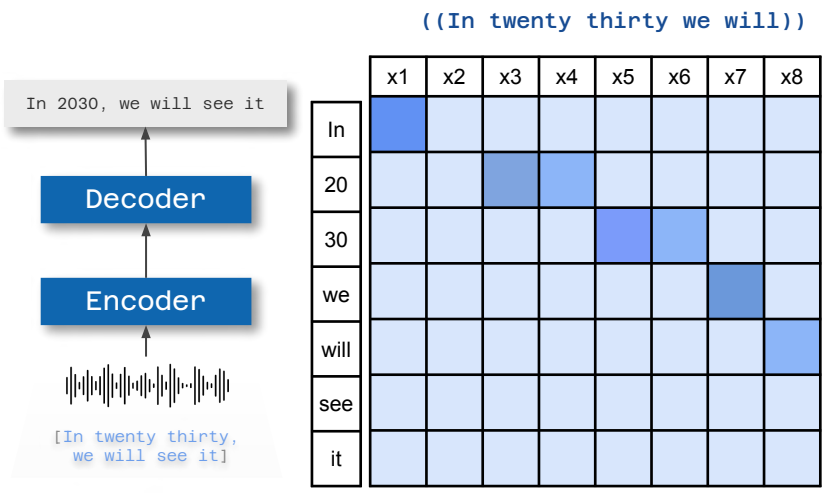


Automatic Speech Translation (AST)

Streaming with Long Acoustic Context

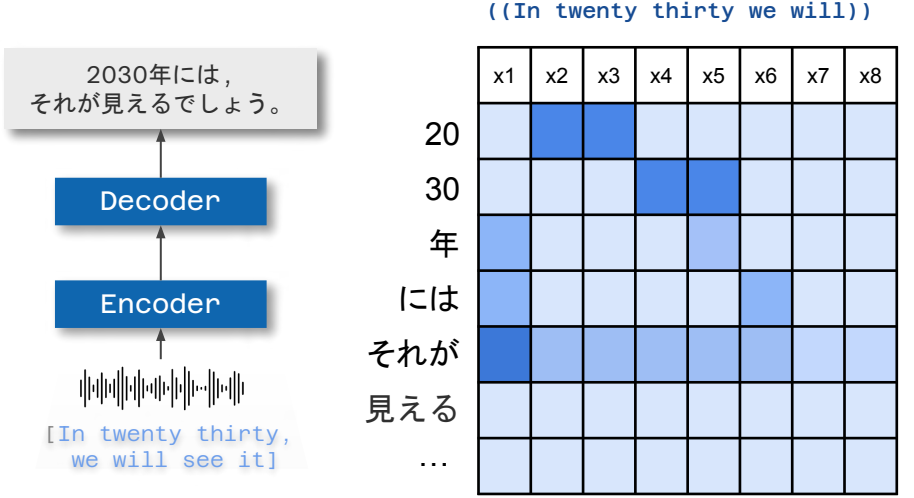
- **Offline models to streaming: Streaming ASR with Translation**

Streaming ASR with Translation = Simultaneous Speech Translation



Automatic Speech Recognition (ASR)

Monotonic Alignment



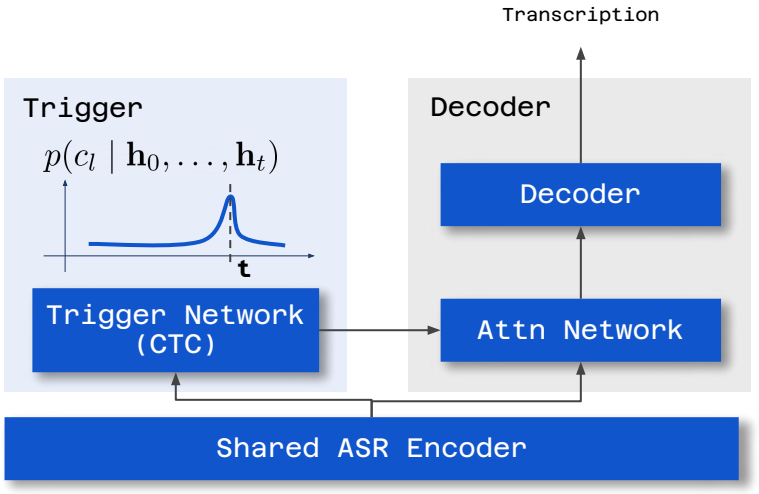
Automatic Speech Translation (AST)

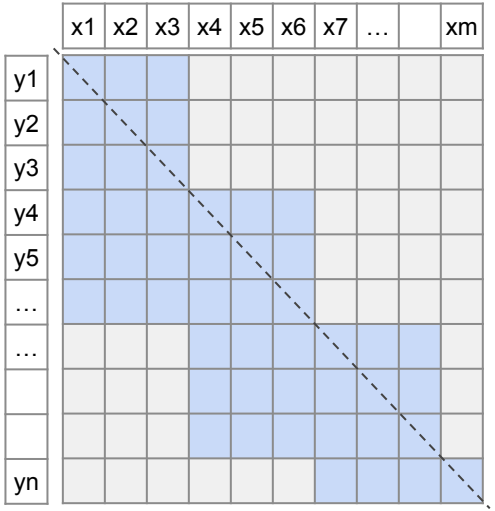
Non-monotonic Alignment

Liu, Xiaoqian, et al. **"Recent advances in end-to-end simultaneous speech translation."** *arXiv preprint arXiv:2406.00497* (2024).

Streaming with Long Acoustic Context

- Long-context Transformers: Attention Mechanisms for Streaming





- CTC-trained trigger network estimates when to activate the attention

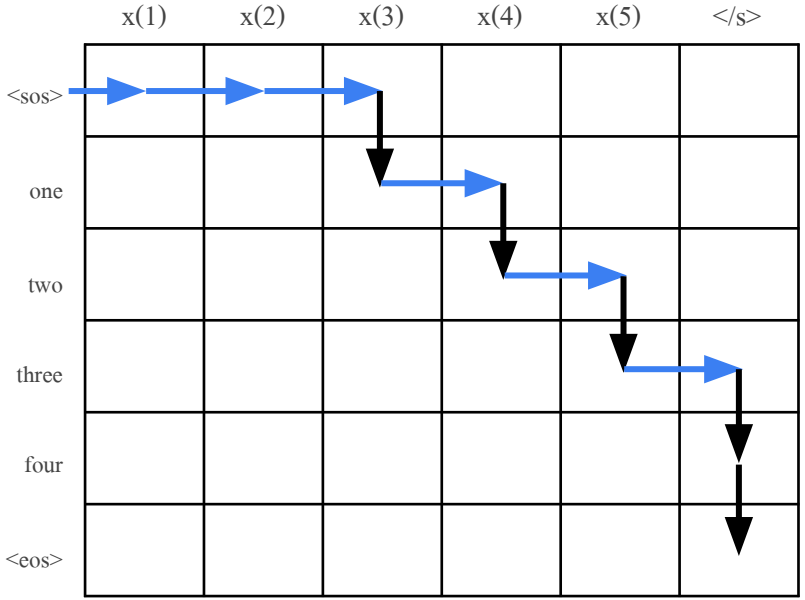
Moritz, Niko, Takaaki Hori, and Jonathan Le Roux. **"Triggered attention for end-to-end speech recognition."** ICASSP 2019
Hori, Takaaki, et al. **"Transformer-Based Long-Context End-to-End Speech Recognition."** *Interspeech 2020*
Hori, Takaaki, et al. **"Advanced Long-Context End-to-End Speech Recognition Using Context-Expanded Transformers."** *Interspeech 2021*

- Masked Attention for Streaming: Train with mask and limit the left context cache

Chen, Xie, et al. **"Developing real-time streaming transformer transducer for speech recognition on large-scale dataset."** ICASSP 2021
Noroozi, Vahid, et al. **"Stateful conformer with cache-based inference for streaming automatic speech recognition."** ICASSP 2024

Streaming with Long Acoustic Context

- Offline models to streaming **Without Training** : Wait-K and other methods



Wait-3 Example:

Input:	x(1)	x(2)	x(3)	x(4)	x(5)	</s>
Output:	one	two	three	four	<eos>	

👍 **Pros:**

- Stable latency
- Simple to implement

👎 **Cons:**

- Fixed delay could be suboptimal
- Poor with long reordering

Elbayad, Maha, Laurent Besacier, and Jakob Verbeek. **"Efficient Wait-k Models for Simultaneous Machine Translation."** *Proc. Interspeech 2020*. 2020.

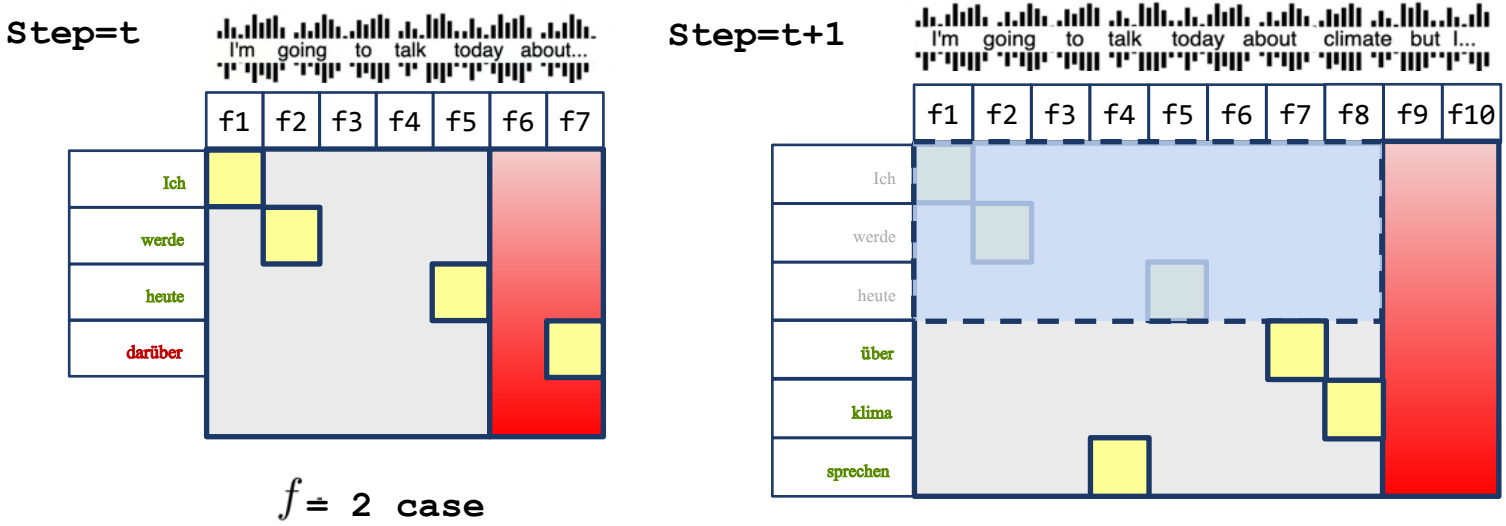
Elbayad, M. Ma, L. Huang, H. Xiong et al., **"STACL: Simultaneous trans-ation with implicit anticipation and controllable latency using prefix-to-prefix framework,"** in Proc. ACL 2019, Florence, Italy, Jul. 2019.

Streaming with Long Acoustic Context

- Offline models to streaming: Wait-K and other methods

AlignATT: Simple decision policy that aligns attention

Decision Policy:
Iterate over the prediction and continue the emission until: $Align_i \notin \{n - f + 1, ..., n\}$



$f = 2$ case

Papi, Sara, Marco Turchi, and Matteo Negri. **"ALIGNATT: Using Attention-based Audio-Translation Alignments as a Guide for Simultaneous Speech Translation.",** Interspeech 2023.

Speaker-attributed ASR and Long-Context

- Takeaways

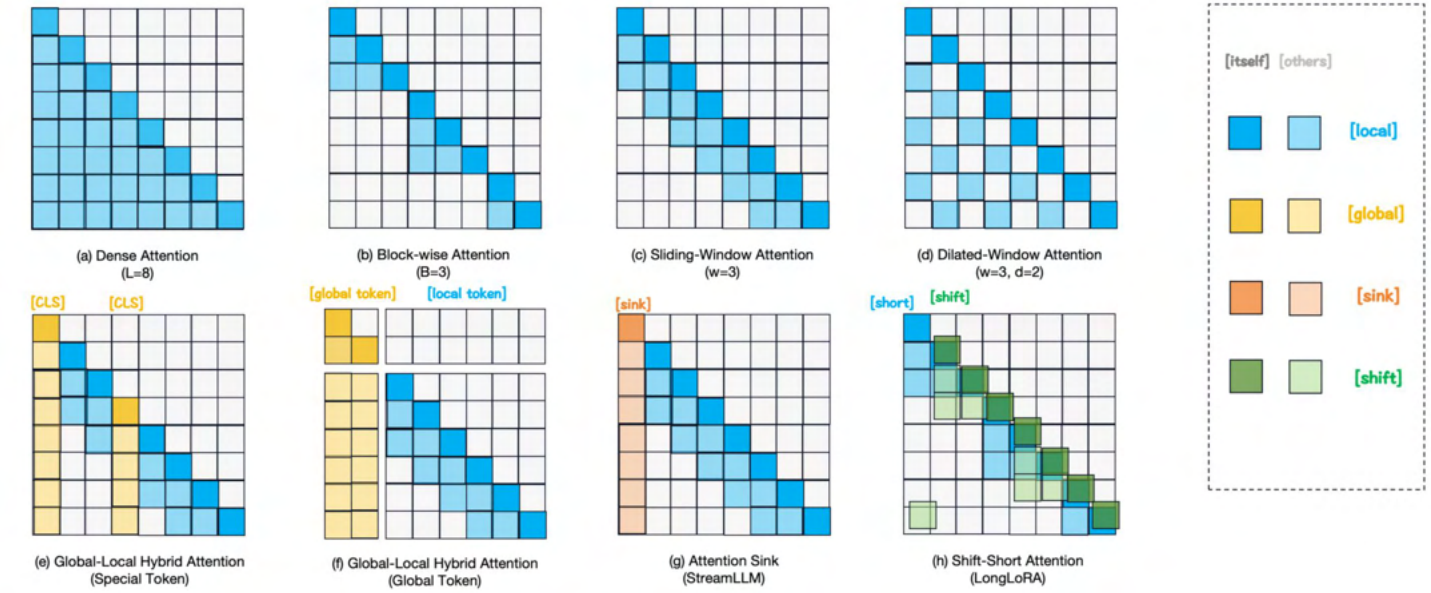
1. Speaker information needs to be separately cached or attended to have a competitive multispeaker ASR performance

2. Even without retraining, Attention Encoder-Decoder models for streaming ASR or simultaneous speech translation models can perform to a certain extent.

Alternative Architectures

Alternative Architectures for Long Context ASR

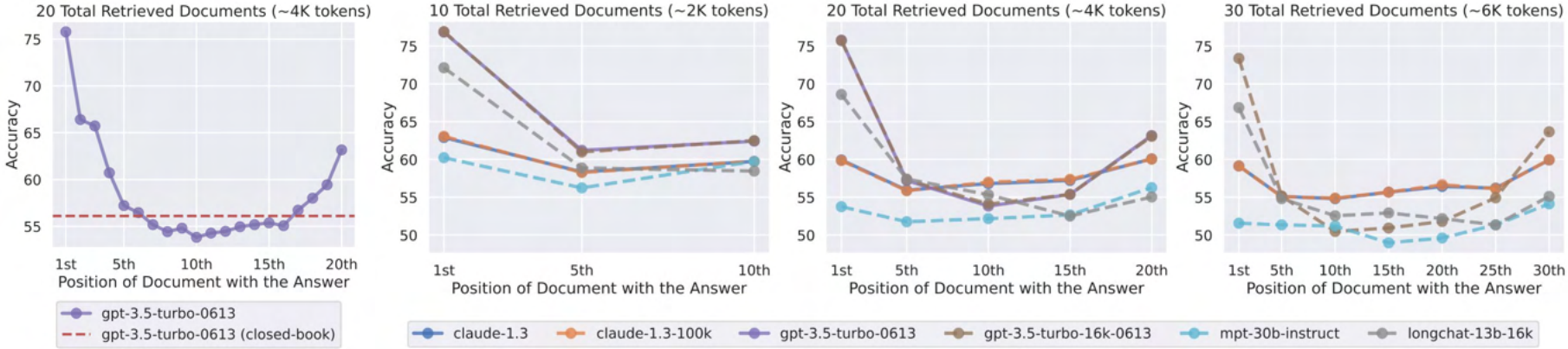
- Long-context Transformers: Attention Mechanisms for Streaming



Huang, Y., et al. **"Advancing transformer architecture in long-context large language models: A comprehensive survey. arXiv 2023."** arXiv preprint arXiv:2311.12351.

Alternative Architectures for Long Context ASR

- Long-context Transformers: Text vs Speech



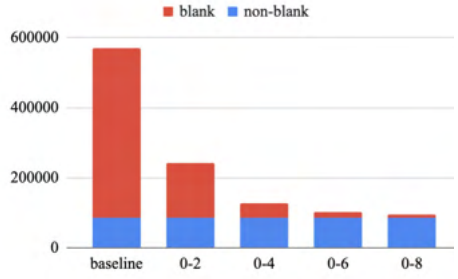
- Only the beginning and end of the long context survives. 😩

Liu, Nelson F., et al. **"Lost in the Middle: How Language Models Use Long Contexts."** Transactions of the Association for Computational Linguistics (2024).

Alternative Architectures for Long Context ASR

- **Token Duration Transducer (TDT)**

Predict Token AND Duration!



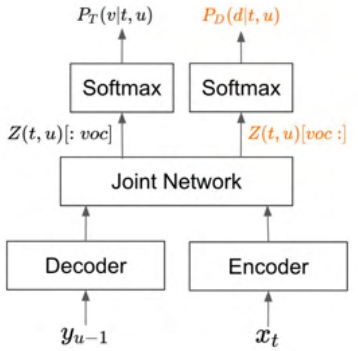
- Same as RNN-T with additional duration predictions
- Outputs token and its duration at each time frame
- TDT loss (2D lattice loss)
- **Faster** inference and **More accurate** than RNN-T

TDT config	WER(%)	time(s)	rel. speed-up
RNNT	2.14	256	-
0-2	2.35	175	1.46X
0-4	2.17	129	1.98X
0-6	2.14	119	2.15X
0-8	2.11	117	2.19X

TDT config	WER(%)	time(s)	rel. speed-up
RNNT	5.11	244	-
0-2	5.50	171	1.43X
0-4	5.06	128	1.91X
0-6	5.05	118	2.07X
0-8	5.16	115	2.12X

- **TDT decoding:** Skip frames based on duration predictions.

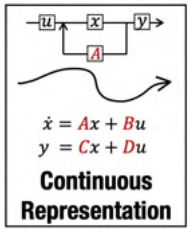
Frame:	**0**	**1**	**2**	**3**	**4**	**5**	**6**		**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Token:	This	->	_is	**blank**	->	->	_output		_from	**blank**	->				…			
Duration:	2			1	3				0	1	**2**							



Xu, Hainan, et al. **"Efficient Sequence Transduction by Jointly Predicting Tokens and Durations."** Proceedings of the 40th International Conference on Machine Learning (ICML), 2023.

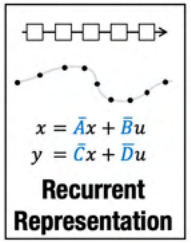
Alternative Architectures for Long Context ASR

- State-Space Models: SSM in a nutshell



$$\dot{\mathbf{x}}(t) = A\mathbf{x}(t) + B\mathbf{u}(t)$$
$$\mathbf{y}(t) = C\mathbf{x}(t) + D\mathbf{u}(t)$$

Discretize



$$x_k = \overline{\boldsymbol{A}} x_{k-1} + \overline{\boldsymbol{B}} u_k$$
$$y_k = \overline{\boldsymbol{C}} x_k$$

- SSMs are broadly used in many scientific disciplines.
- SSMs are related to **latent state models** such as Hidden Markov Models (HMM).
- A **State Space Model (SSM)** describes a system using hidden "state" variables that evolve over time and produce "observable outputs" in the form of differential equations.

[1] https://en.wikipedia.org/wiki/State-space_representation
[2] Gu, Albert, Karan Goel, and Christopher Ré. "**Efficiently modeling long sequences with structured state spaces.**" arXiv preprint arXiv:2111.00396 (2021)

Alternative Architectures for Long Context ASR

- State-Space Models: Applications of SSMs

	MFCC	Raw	0.5×
Transformer	90.75	✗	✗
Performer	80.85	30.77	30.68
ODE-RNN	65.9	✗	✗
NRDE	89.8	16.49	15.12
ExpRNN	82.13	11.6	10.8
LipschitzRNN	88.38	✗	✗
CKConv	**95.3**	71.66	65.96
WaveGAN-D	✗	96.25	✗
LSSL	93.58	✗	✗
S4	93.96	**98.32**	**96.30**

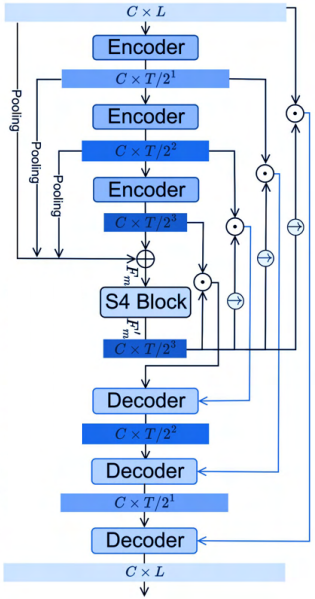
- **S4**[1] **model's** Raw timeseries speech classification
- **S4** model outperforms RNN and attention-variants

[1] Gu, Albert, **"Efficiently Modeling Long Sequences with Structured State Spaces"**, *ICLR 2022*

⊕ Sum operator

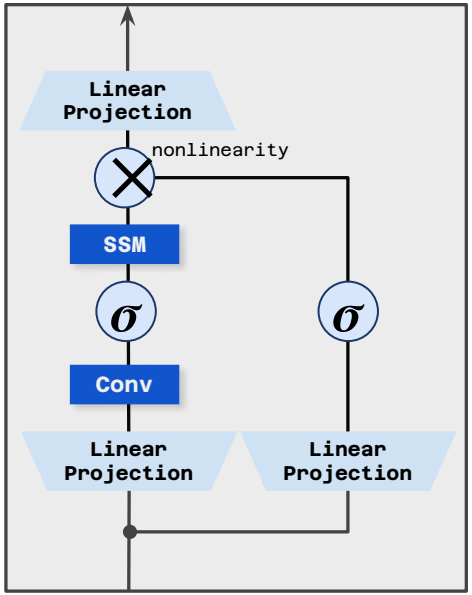
⊙ Element-wise product

↑ Nearest Neighbour interpolation



- S4M: Neural Speech Separation

Chen, Chen, Chao-Han Huck Yang, Kai Li, Yuchen Hu, Pin-Jui Ku, and Eng Siong Chng. **"A Neural State-Space Model Approach to Efficient Speech Separation."**, *Interspeech 2025*

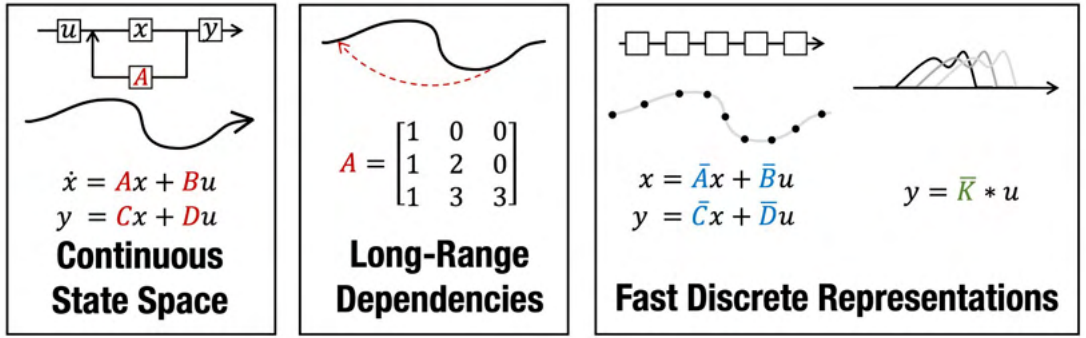


- SSMs were later popularized with Mamba[2], a specialized architecture built on the SSM framework.

[2] Gu, Albert, and Tri Dao. **"Mamba: Linear-Time Sequence Modeling with Selective State Spaces."** *COLM 2024.*

Alternative Architectures for Long Context ASR

- State-Space Models: SSM in a nutshell



- Sequence modeling struggles with **long-range dependencies (LRD).**
- SSMs have **linear memory scaling** while having a good **LRD.**
- Natural inductive bias for sequential and temporal data. Thus, widely used **for modeling time-series.**

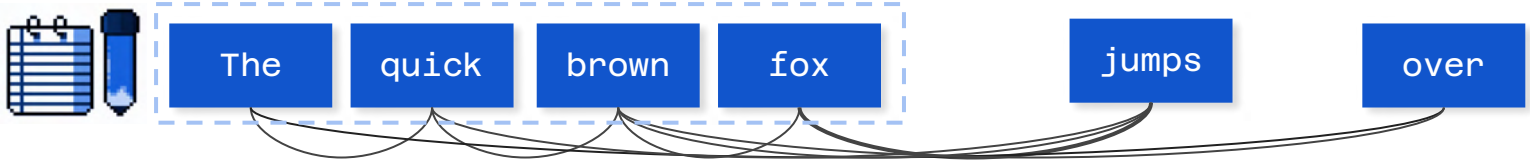
Gu, Albert, Karan Goel, and Christopher Ré. "**Efficiently modeling long sequences with structured state spaces.**" arXiv preprint arXiv:2111.00396 (2021)

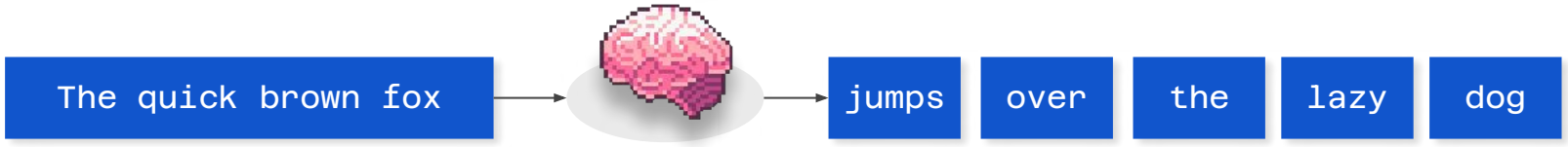
Alternative Architectures for Long Context ASR

- State-Space Models: Quoting Albert Gu's analogy [1] ...

A Coarse Analogy by Albert Gu: Attention vs SSM

Transformers are like a "database"





State space models are like a "brain"

[1] Gu, Albert. **"On the Tradeoffs of State Space Models and Transformers."** *AI Heard That! ICML 2025 Workshop on Machine Learning for Audio*, 27 July 2025

Alternative Architectures for Long Context ASR

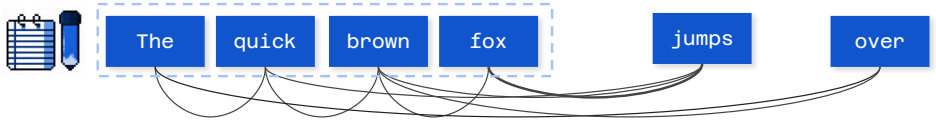
- State-Space Models: Quoting Albert Gu's analogy [1] ...

Tradeoffs of Attention and Transformer

👍 Pros:

Fine-grained attention over past context

Strong at **recall and retrieval**



👎 Cons:

Known to be Efficient?: quadratic scaling training, linear time inference

→ **But Dependent on resolution and semantic content of data**

Reality: Attention is only effective on pre-compressed data at the "right level of abstraction"

[1] Gu, Albert. **"On the Tradeoffs of State Space Models and Transformers."** *AI Heard That! ICML 2025 Workshop on Machine Learning for Audio*, 27 July 2025

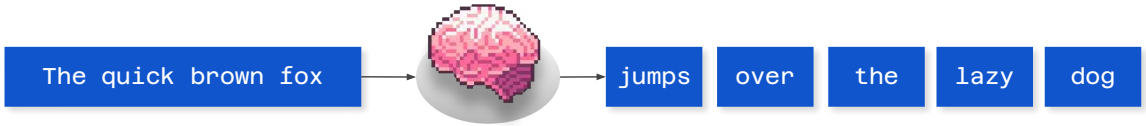
Alternative Architectures for Long Context ASR

- State-Space Models: Quoting Albert Gu's analogy [1] ...

Tradeoffs of State Space Models

👍 **Pros:**

Stateful, compressive model



👎 **Cons:**

Lacks fine-grained recall and retrieval abilities

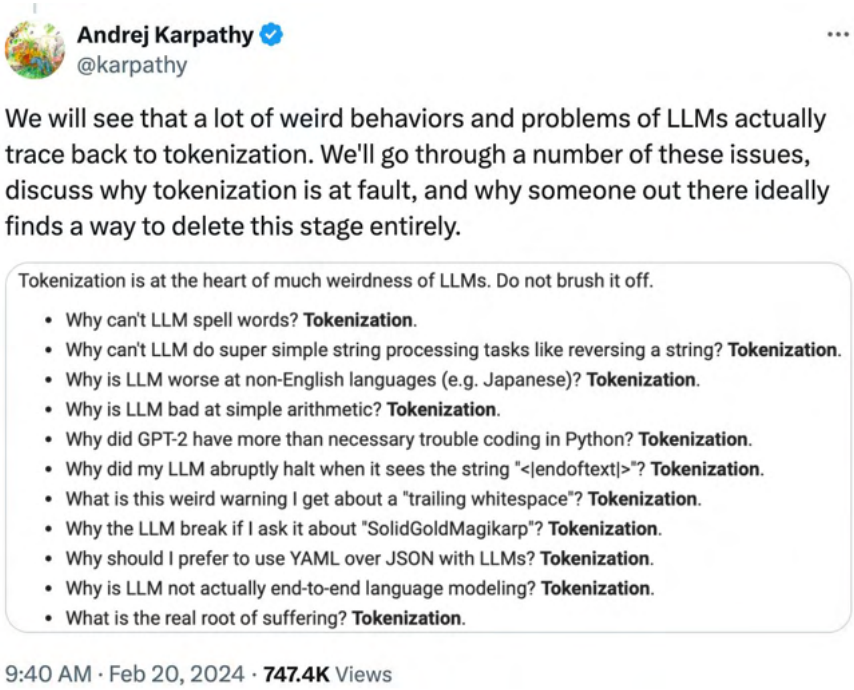
Associative recall, needle-in-a-haystack, general QA.. (Recall and Retrieval)

[1] Gu, Albert. **"On the Tradeoffs of State Space Models and Transformers."** *AI Heard That! ICML 2025 Workshop on Machine Learning for Audio*, 27 July 2025

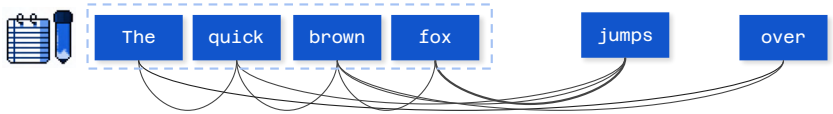
Alternative Architectures for Long Context ASR

- State-Space Models: Quoting Albert Gu's examples [1]

Tokenization is "the problem" of attention



Effective Tokens for Attention



(sub)word tokens ✅
- Semantically meaningful
- Modular, composable

character tokens ❌
DNA tokens ❌
visual "tokens" ?

- Does **hard attention** make sense?
- Does caching a representation for every "token" of data make sense?

Albert Gu says:

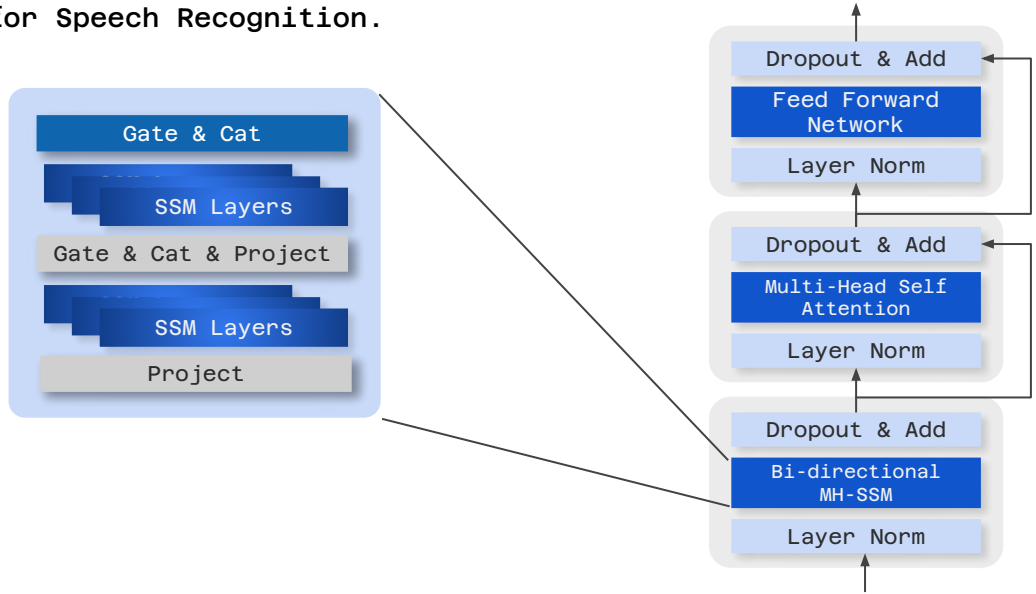
Attention is only effective on pre-compressed data at the "**right level of abstraction**"

[1] Gu, Albert. **"On the Tradeoffs of State Space Models and Transformers."** *AI Heard That! ICML 2025 Workshop on Machine Learning for Audio*, 27 July 2025

Alternative Architectures for Long Context ASR

- State-Space Models: SSM for Speech Recognition

Multi-Head SSM for Speech Recognition.



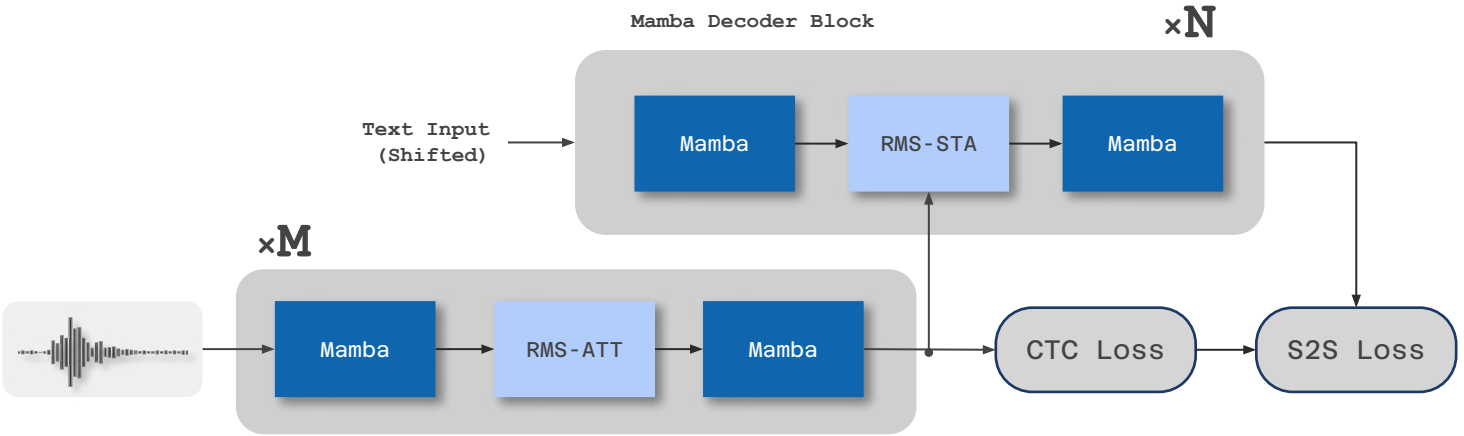
- Drop-in replacement for multihead attention in transformer encoders
- Outperforms the same size transformer transducer

Fathullah, Yassir, et al. **"Multi-Head State Space Model for Speech Recognition."** *INTERSPEECH 2023*

Alternative Architectures for Long Context ASR

- State-Space Models: SSM for Speech Recognition

*RMS-STA: RMS-Norm and source-target multi-head attention
*RMS-ATT: RMS-Norm and multi-head attention

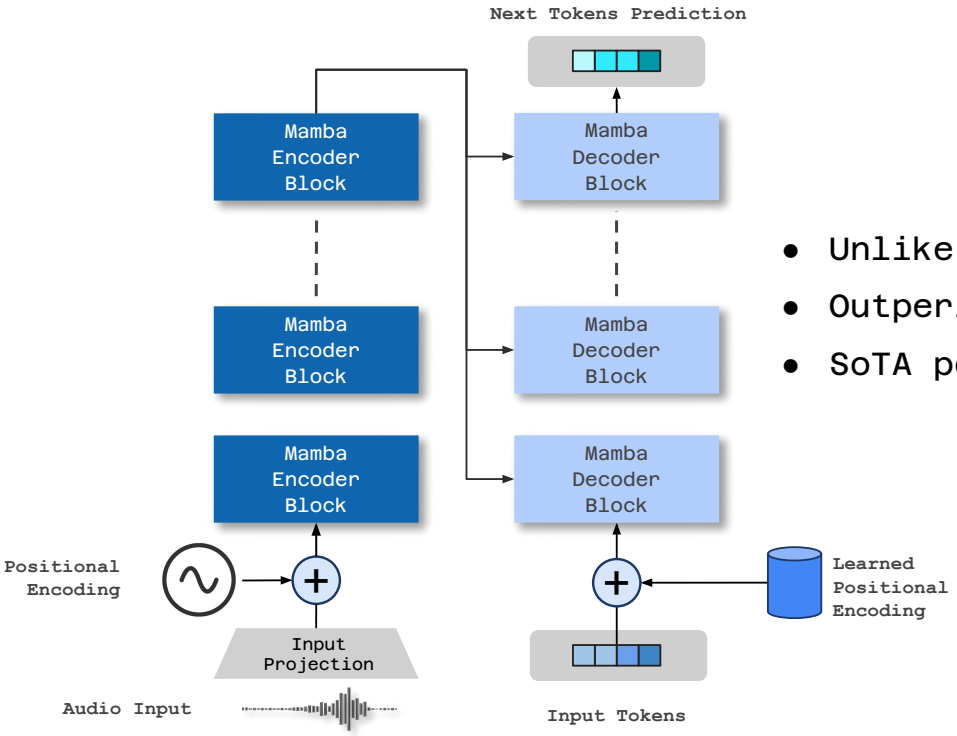


- Incorporates selective SSMs in Transformer architectures with CTC-Loss.
- Outperforms the same size transformer-CTC/Transformer

Gao, Xiaoxue, and Nancy F. Chen. "**Speech-mamba: Long-context speech recognition with selective state spaces models.**" 2024 IEEE Spoken Language Technology Workshop

Alternative Architectures for Long Context ASR

- State-Space Models: SSM for Speech Recognition



- Unlike previous work, fully SSM based model.
- Outperforms the same size Transformer models.
- SoTA performance on public benchmark

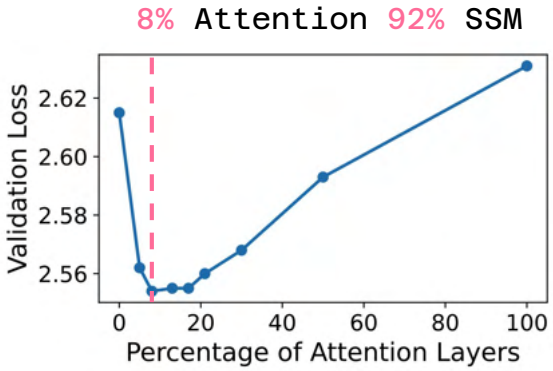
Shakhadri, Syed Abdul Gaffar, Kruthika KR, and Kartik Basavaraj Angadi. **"Samba-ASR: State-Of-The-Art Speech Recognition Leveraging Structured State-Space Models."** *arXiv preprint arXiv:2501.02832* (2025)

Alternative Architectures for Long Context ASR

- **Hybrid Models? Attention + SSM**

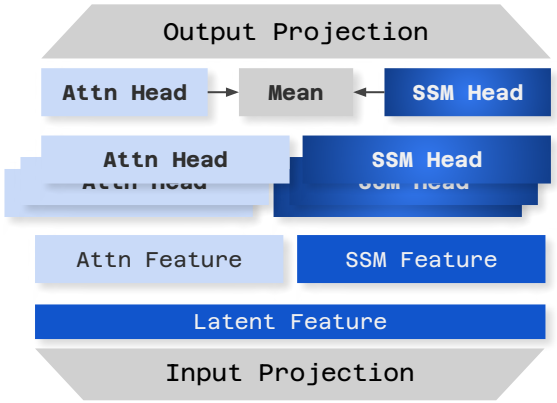
1. Mix Mamba-2, Transformer, MLP and self-attentions. Ablations[1] (130M, 25 Layers) show:

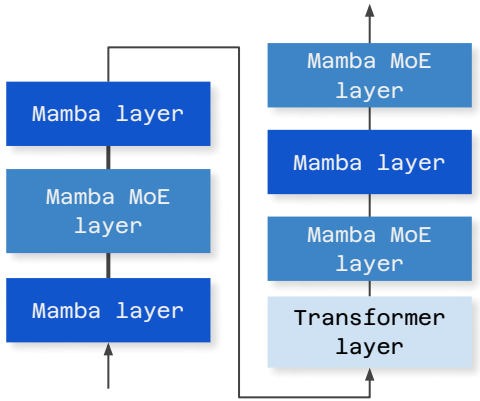
~8% self-attention layers yields the best result.



[1] Waleffe, Roger, et al. **"An empirical study of mamba-based language models."** arXiv preprint arXiv:2406.07887 (2024).

2. Hybrid Approaches: Attention + SSM





[2] Dong, Xin, et al. **"Hymba: A hybrid-head architecture for small language models."** arXiv preprint arXiv:2411.13676 (2024).

[3] Lenz, Barak, et al. **"Jamba: Hybrid transformer-mamba language models."** *The Thirteenth International Conference on Learning Representations*. 2025

Alternative Architectures for Long Context ASR

- Takeaways from SSM

👍 Pros

- 👍 Strengths in time-series data (speech) or non-tokenized data
- 👍 Linear memory scaling - especially efficient at runtime
- 👍 Parameter efficiency in certain types of tasks

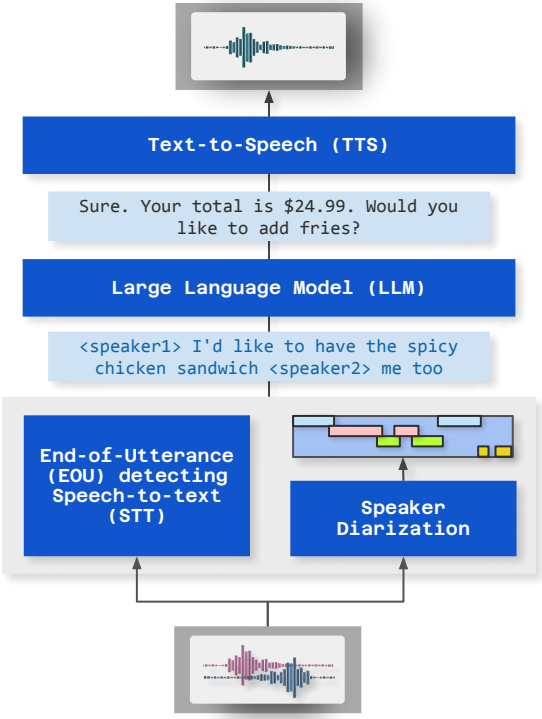
👎 Cons

- 👎 Weak retrieval and recall task performance
- 👎 Underperforms with tokenized data (text) (e.g., Q&A tasks)
- 👎 Less mature; fewer optimized models, tools and silicons than attention-based ones

Voice-agent, Speech-LLM and Acoustic context

Voice-agent, Speech-LLM and Acoustic Context

- Modular (EOU-ASR + LLM + TTS) Systems

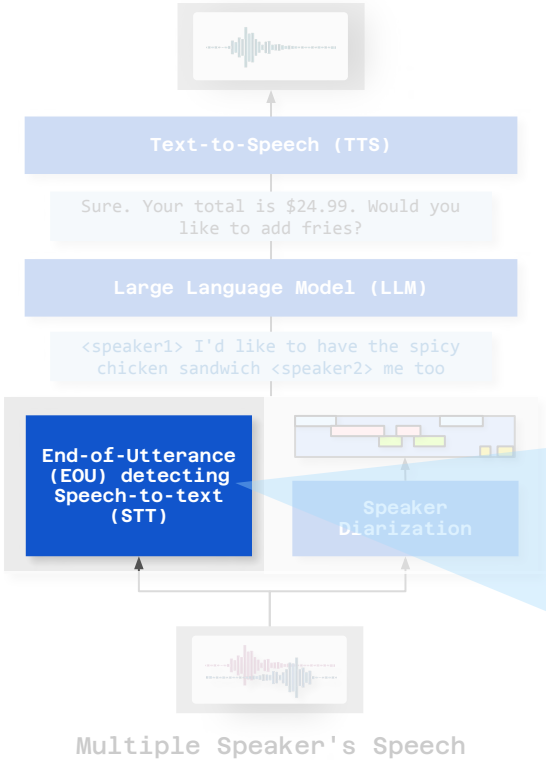


Multiple Speaker's Speech

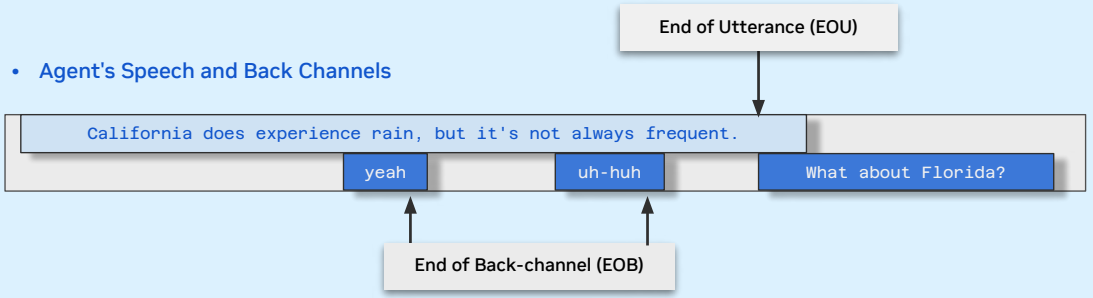
Voice AI agent Framework : Pipecat
STT : Speechmatics (ASR + Diarization)
LLM : OpenAI
TTS : Elevenlabs

Voice-agent, Speech-LLM and Acoustic Context

- Modular (EOU-ASR + LLM + TTS) Systems



- Agent's Speech and Back Channels



Backchannel word example

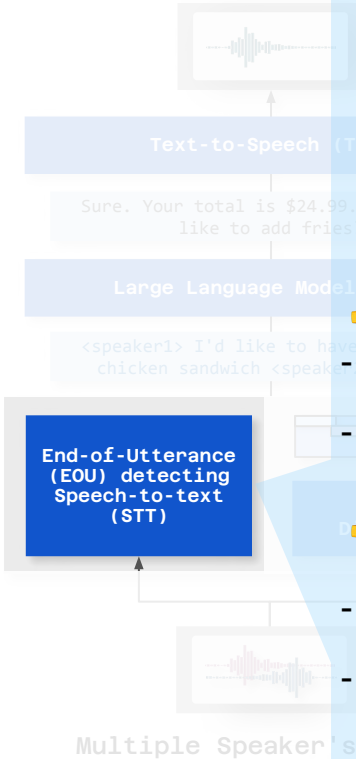

```
{
"text": "yeah",
"offset": 274.13,
"duration": 1.86,
"sou_time": 274.838,
"eou_time": 275.485,
"is_backchannel": true,
"is_end_of_utterance": true,
```


Non-backchannel word example

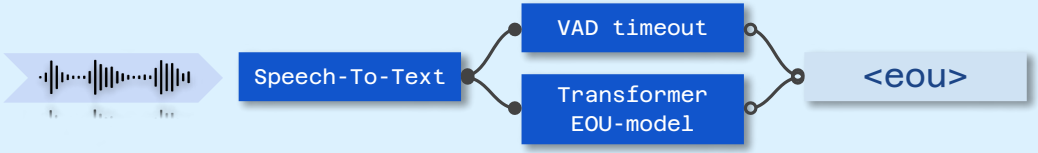

```
{"text": "yeah that would make more meals but youd
only get one if you went out",
"offset": 61.65,
"duration": 4.1,
"sou_time": 62.072,
"eou_time": 65.248,
"is_backchannel": false,
"is_end_of_utterance": true,
```

Voice-agent, Speech-LLM and Acoustic Context

- Modular (EOU-ASR + LLM + TTS) Systems



1. LiveKit EOU: (0:55 w/o EOU, 1:55 with EOU)
   - Do STT then apply EOU model with VAD timeout



👍 Pros:

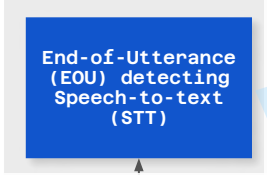
- Use data-rich text EOU datasets
- Can mix and match preferred models

👎 Cons:

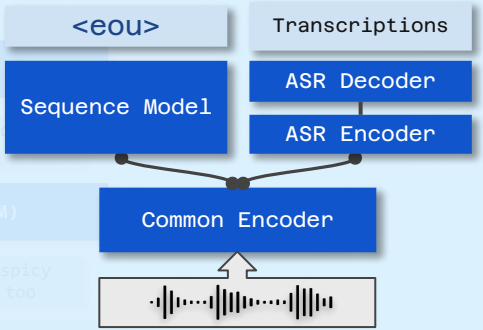
- No use of acoustic information
- Not so development friendly Models

Voice-agent, Speech-LLM and Acoustic Context

- Modular (EOU-ASR + LLM + TTS) Systems



2. Multitask ASR and EP



👍 Pros:

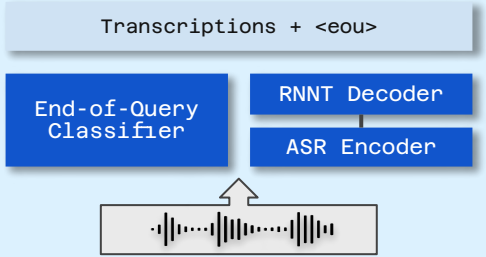
- Share the same encoder state.
- Can be used for speech-to-speech model

👎 Cons:

- Does not utilize ASR transcription
- Still requires both text and audio

Bijwadia, Shaan et. al., "**Unified end-to-end speech recognition and endpointing for fast and efficient speech systems.**" In *2022 IEEE Spoken Language Technology Workshop (SLT)*

3. Joint ASR and EOU



👍 Pros:

- Joint ASR model that transcribes and detects end points
- Simplifies the whole system

👎 Cons:

- With only RNNT-EOU-ASR, end-pointing coverage shrinks
- Still require external EOQ module for the optimized performance

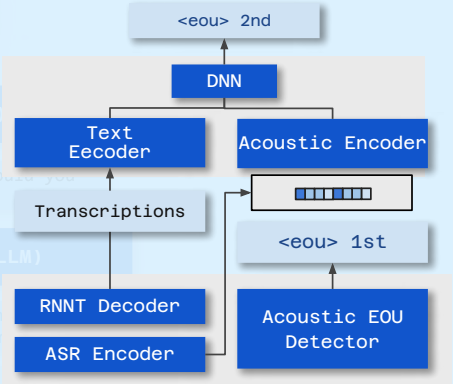
Chang et. al.,**"Joint endpointing and decoding with end-to-end models."** In ICASSP 2019

Voice-agent, Speech-LLM and Acoustic Context

- Modular (EOU-ASR + LLM + TTS) Systems

End-of-Utterance (EOU) detecting Speech-to-text (STT)

3. Two-pass Hybrid EOU ASR



👍 Pros:

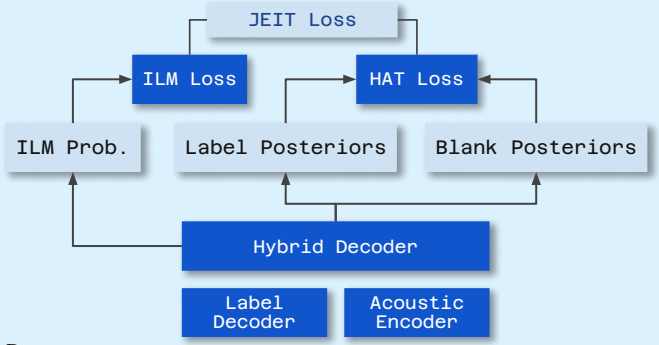
- Leverages both acoustic and linguistic context
- Can train the system with only-text or only-audio.

👎 Cons:

- Additional compute overhead
- Latency burden

Raju, Anirudh et al. **"Two-pass endpoint detection for speech recognition."** In 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 1-8. IEEE, 2023.

4. Hybrid Autoregressive Transducer (HAT) with Internal LM (ILM)



👍 Pros:

- Take advantage of both acoustic and linguistic context
- Powerful internal LM supports EOU

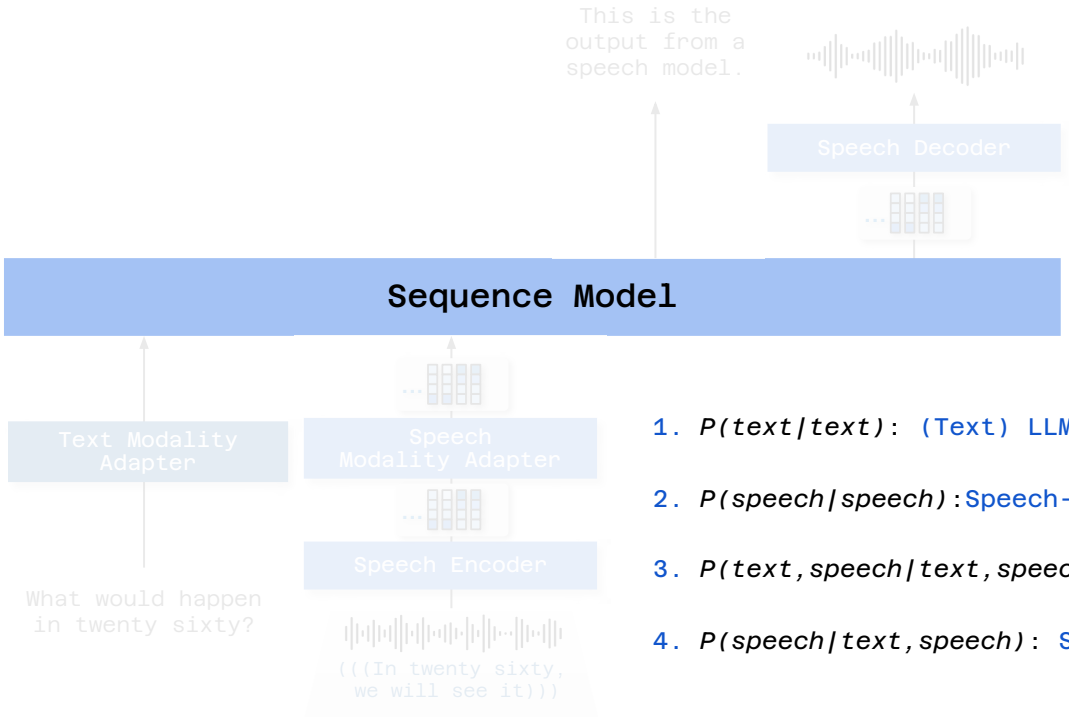
👎 Cons:

- Complicated training steps and architecture

Meng, Zhong et. al.,"**Jeit: Joint End-to-end model and internal language model training for speech recognition.**" In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1-5. IEEE, 2023.

Voice-agent, Speech-LLM and Acoustic Context

- Jointly trained duplex Speech LMs

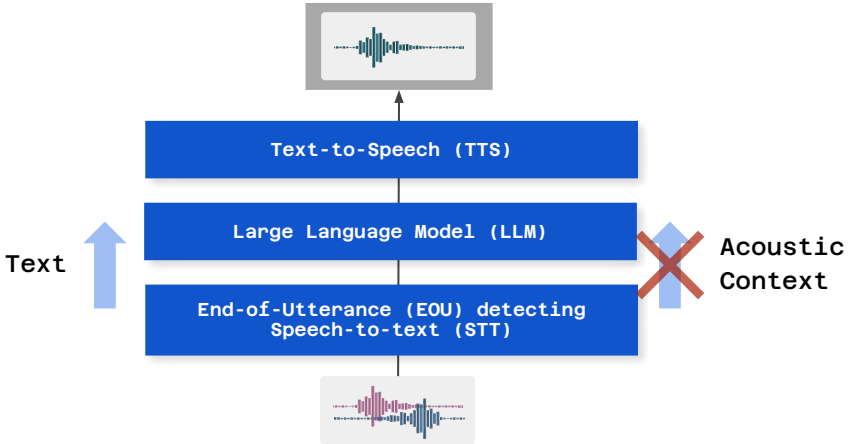


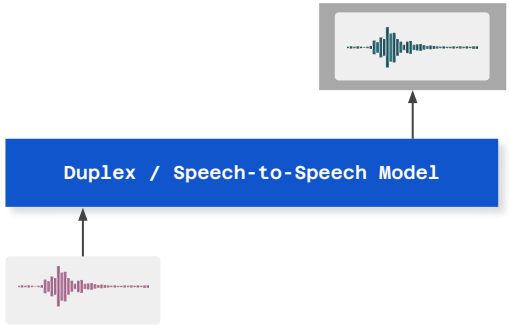
1. *P(text|text)*: (Text) LLM
2. *P(speech|speech)*:Speech-to-Speech
3. *P(text,speech|text,speech)*:Speech Language Model (SLM)
4. *P(speech|text,speech)*: Speech-aware Text LM

Arora, Siddhant, et al. **"On the landscape of spoken language models: A comprehensive survey."** arXiv preprint arXiv:2504.08528 (2025).

Voice-agent, Speech-LLM and Acoustic Context

- Jointly trained duplex Speech LMs





👍 Pros:
- Easily customize LLMs with prompting
- Plug-in different modules (e.g. Diarization)

👎 Cons:
- Compounding errors from ASR → LLM → TTS
- Lack of channel for acoustic and paralinguistic information
- Relatively high latency

👍 Pros:
- Leverages both acoustic and linguistic context
- Relatively Low latency

👎 Cons:
- Hard to customize for the specific need
- Scarcity of the training dataset

Arora, Siddhant, et al. **"On the landscape of spoken language models: A comprehensive survey."** arXiv preprint arXiv:2504.08528 (2025).

Voice-agent, Speech-LLM and Acoustic Context

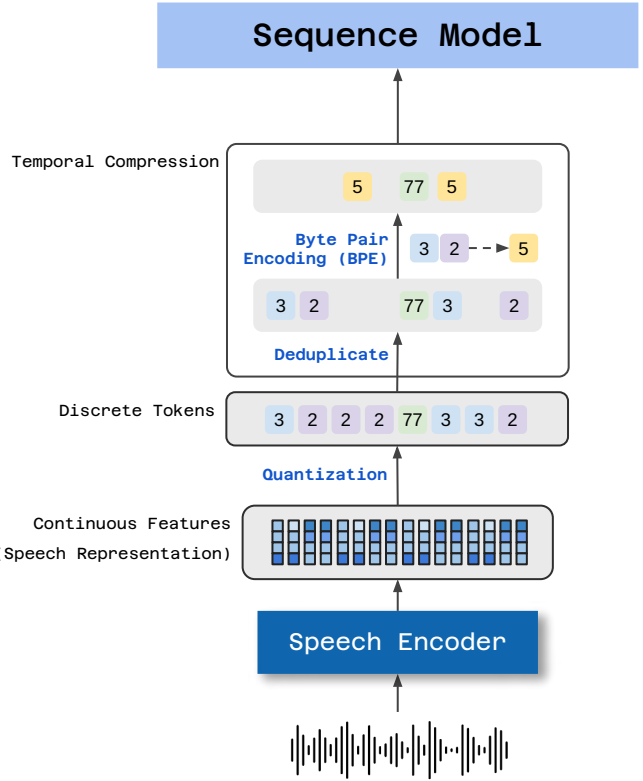
- Jointly trained duplex Speech LMs: Speech Encoder

Continuous Features

1. Spectrogram Features
   a. Mel-Spec Feature
2. Hidden representations from SSL
   a. WavLM, Wav2Vec 2.0, HuBERT
3. Hidden representations from ASR models
   a. Whisper ASR, USM (ASR)
4. Neural audio codecs
   a. EnCodec, SoundStream

Discrete Tokens

1. Phonetic Tokens - Clustering based
2. Audio Codec Tokens: For more detailed acoustic context
   a. Vector Quantization



Défossez, Alexandre, et al. "**High Fidelity Neural Audio Compression.**" *Trans. Mach. Learn. Res.* (2023).

Arora, Siddhant, et al. **"On the landscape of spoken language models: A comprehensive survey."** arXiv preprint arXiv:2504.08528 (2025).

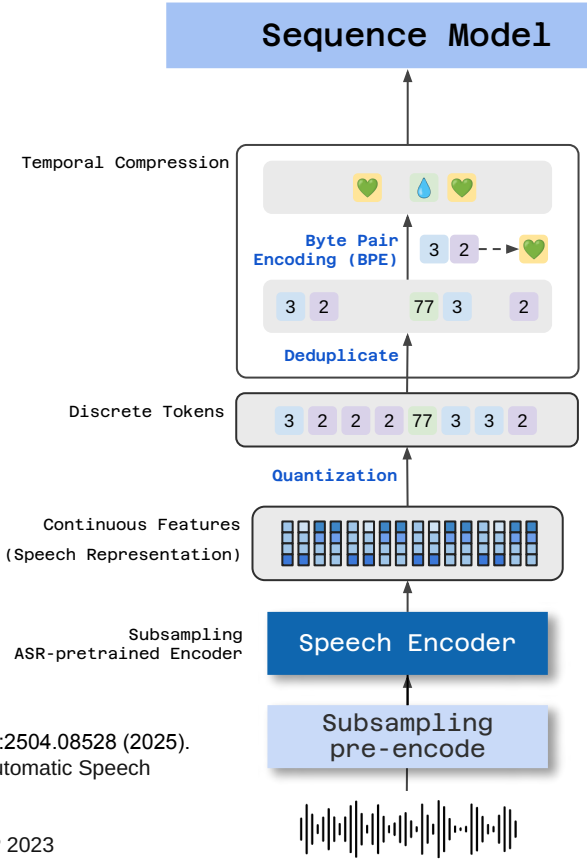
Voice-agent, Speech-LLM and Acoustic Context

- Jointly trained duplex Speech LMs: Speech Encoder

Dealing with a long temporal context

- Speech features have lower information density than text tokens.[1].
- **Temporal compression**: efficiency in runtime speed and memory footprint.
- Techniques for reducing the sequence length.
  - Use **subsampling ASR**[2] model as a speech encoder:
    → Use convnet to increase the frame length (10ms → 80ms
  - **Deduplication**: Merges consecutive identical tokens
  - **Byte Pair Encoding (BPE)**:
    - Merges common co-occurring characters into newly created tokens
    - leverages the morphological information
    - Acoustic BPE[3], Wav2Seq[4]

Diagram from [1]



[1] Arora, Siddhant, et al. **"On the landscape of spoken language models: A comprehensive survey."** arXiv preprint arXiv:2504.08528 (2025).
[2] Rekesh, Dima, et al. **"Fast conformer with linearly scalable attention for efficient speech recognition."** 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2023.
[3] Shen, Feiyu, et al. **"Acoustic bpe for speech generation with discrete tokens."** ICASSP 2024
[4] Wu, Felix, et al. **"Wav2seq: Pre-training speech-to-text encoder-decoder models using pseudo languages."** ICASSP 2023

Speaker-attributed ASR and Long-Context

Takeaways

1. ASR endpointing model plays a crucial role for voice agent.

2. For end point detection ASR taking both acoustic and linguistic context simultaneously.

3. Duplex models also have strength in latency and paralinguistic information but much harder to customize to a specific need.

Table of Contents

Download Slides

Introduction (10 mins) 15:30-15:40

Shinji (30 min) 15:40-16:10 Speech-to-Text Benchmark

Taejin (40 min) 16:10-16:50 Leveraging Long Acoustic Context

Recess (10 min) 16:50-17:00

Huck (40 min) 17:00-17:40 Semantic Context and Speech-Language Modeling

Kyu (30 min) 17:40-18:10 Contextual Biasing and Methods Leveraging Longer Semantic Context for Speech Systems

Closing Remark (10 min) 18:10-18:20

Q&A Session (10 min) 18:20-18:30

10min Recess

16:50-17:00

Table of Contents

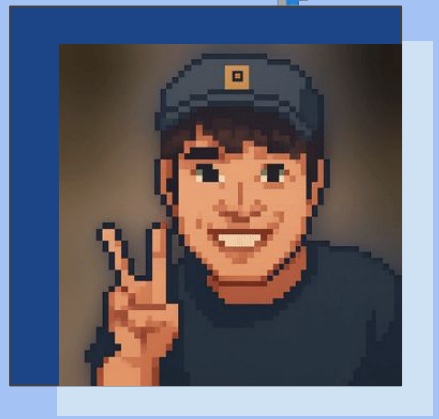
Download Slides

	Introduction (10 mins) 15:30-15:40
Shinji (30 min) 15:40-16:10	Speech-to-Text Benchmark
Taejin (40 min) 16:10-16:50	Leveraging Long Acoustic Context
	Recess (10 min) 16:50-17:00
Huck (40 min) 17:00-17:40	**Semantic Context and Speech-Language Modeling**
Kyu (30 min) 17:40-18:10	Contextual Biasing and Methods Leveraging Longer Semantic Context for Speech Systems
	Closing Remark (10 min) 18:10-18:20
	Q&A Session (10 min) 18:20-18:30

Semantic Context and Speech-Language Modeling

Huck Yang

Huck Yang



SubTitle of talk

1. **Semantic Information Modeling in End-to-End ASR**
   a. Classical ASR-LM
   b. Post-processing ASR correction mechanism

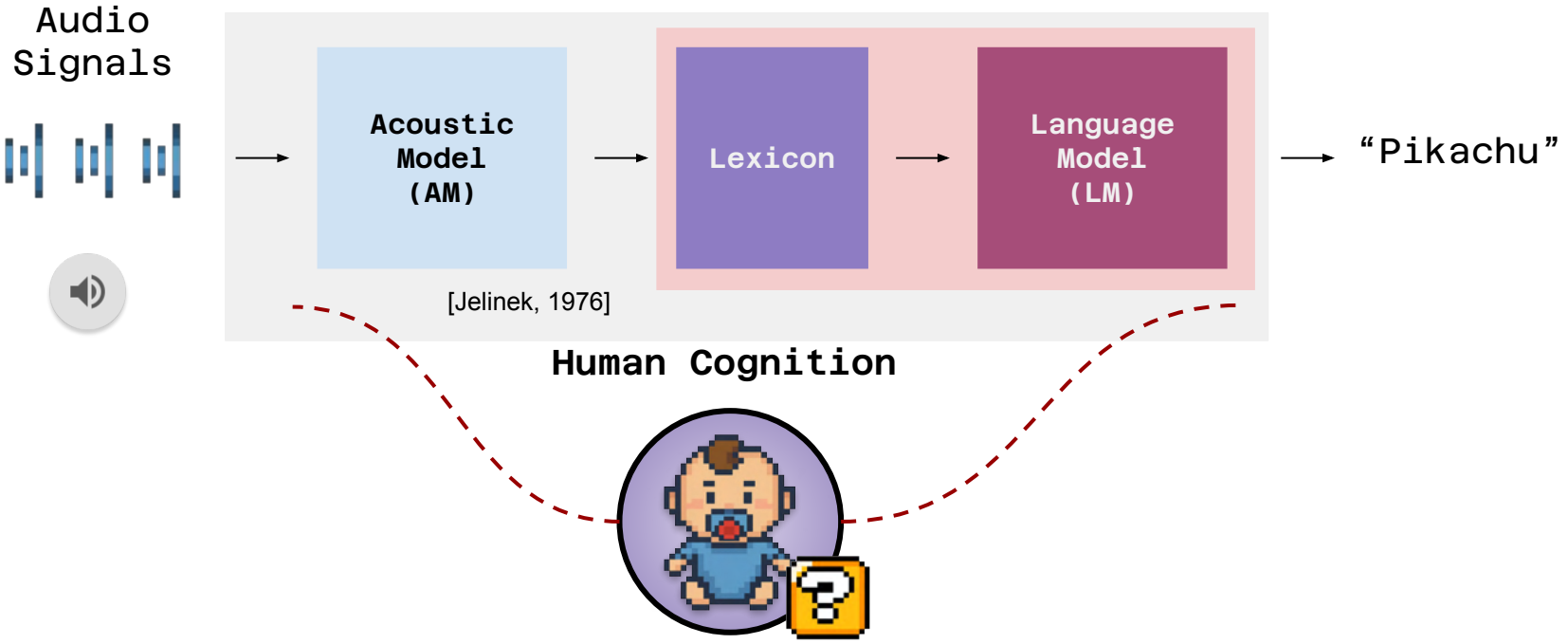
2. **Preference-based Semantic Modeling in Post-training**
   a. LLMs based ASR variants
   b. Post-training mechanism with RLs

3. **Semantic Understanding from ASR to Audio Contexts**
   a. Joint speech and sound semantic modeling (BELU)
   b. Multimodal QA with non-linguistic semantics

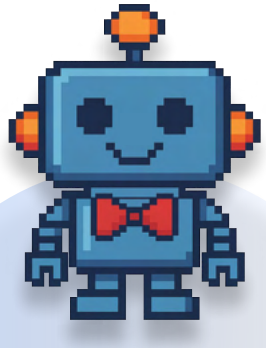
4. **Limitation and New Evaluation of Semantic Modeling**
   a. Data Leakage via Text Pre-training
   b. Agentic and Instruction Evaluation

Speech-Language Modeling: background of ASR-LM

This session with ASR + NLP



Semantic Alignment in Large Models (1/2)



Large Models

(i.e., Superior to Human Abilities)

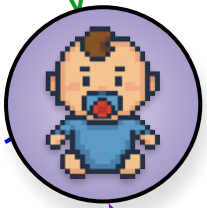
"super listening → i love u"

text

Semantics

Sect 3.1 (10 min)

voice



Human Cognition Model

Cross-modality

Sect 3.2 (10 min)

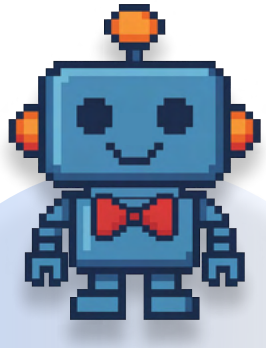
feedbacks

Preference & Values

Sect 3.3 (10 min)

output

Semantic Alignment in Large Models (1/2)



Large Models

(i.e., Superior to Human Abilities)

"super listening → i love u"

text

Semantics

voice

➡️ **Sect 3.1 (10 min)**

Human Cognition Model

Cross-modality

Sect 3.2 (10 min)

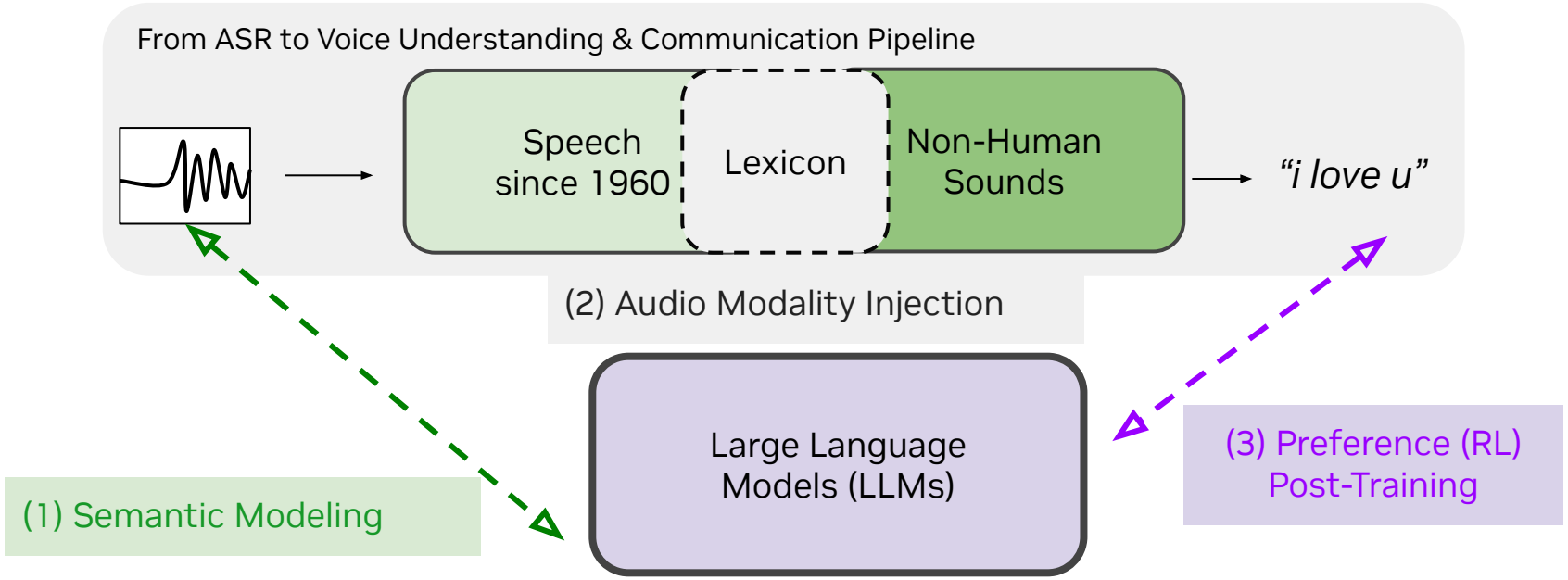
feedbacks

Preference & Values

output

Sect 3.3 (10 min)

Semantic Alignment in Large Models (2/2)



3.1 - Background (1/2) Semantic modeling in ASR-LM

- Language Model (LM) Rescoring over Decoding Outputs (e.g., Utterance Levels)
  - To modify **negative log-likelihood (NLL) scores** from acoustic model **(AM)**

$$\hat{W} = \underset{W}{\operatorname{argmax}} \left[\log P_{\mathrm{AM}}(W|X) + \lambda \cdot \log P_{\mathrm{LM}}(W) \right]$$

ASR words output

NLL from AM to the word utterance (W)

Neural ASR Model

ASR Outputs

Neural LM Rescoring

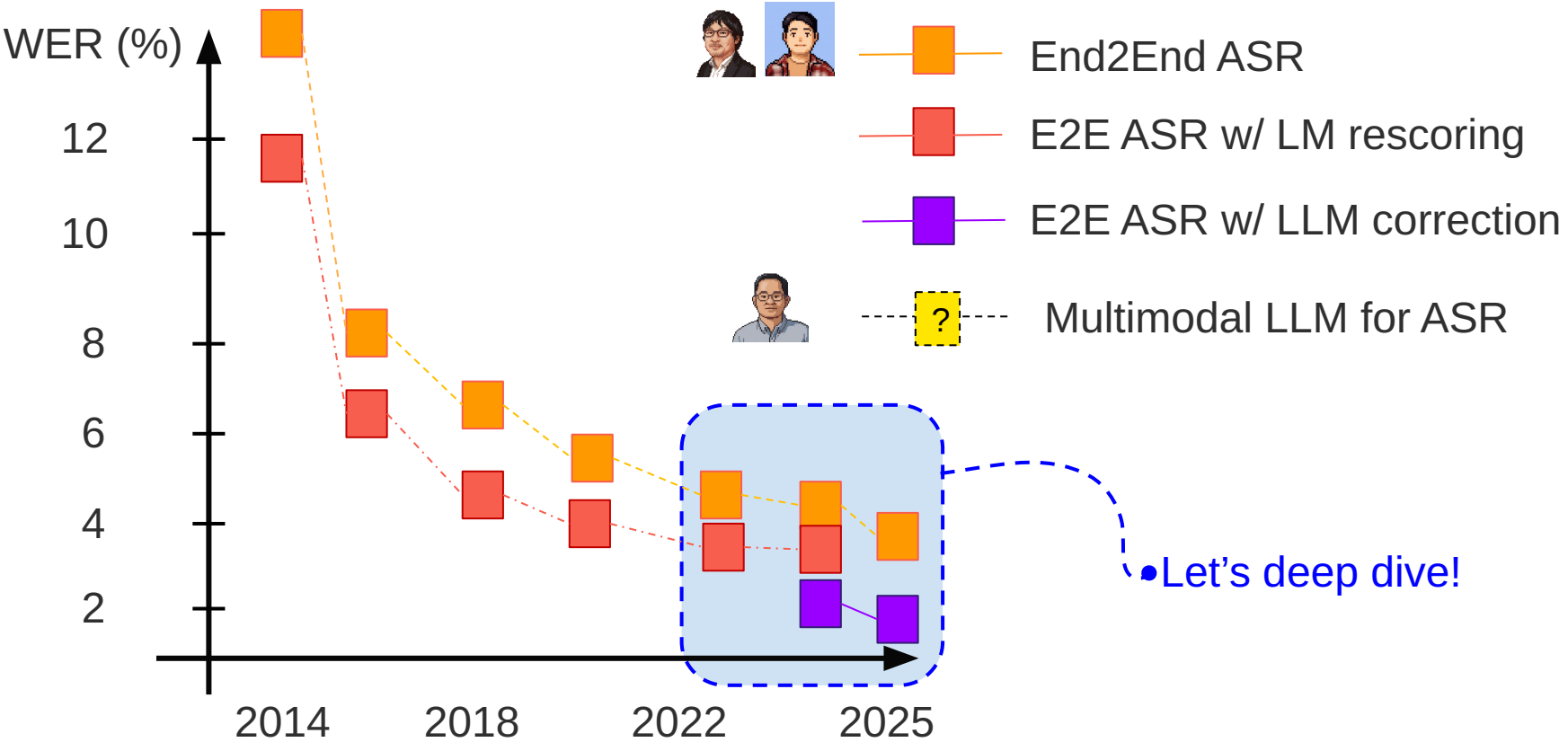
score: 23.6
hyp 1: I flies to Pittsburgh from New York

score: 24.1
hyp 2: I fly to Pot Burger from New York

✅ score: 24.5 → 17.2
hyp 3: I fly to Pittsburgh from New York

- Multi-task LM and NLU training for ASR [ASRU 2021]
- Contextualization and Retrieval (see Part 4)

3.1 - Background (2/2) - End2End ASR vs ASR-LM



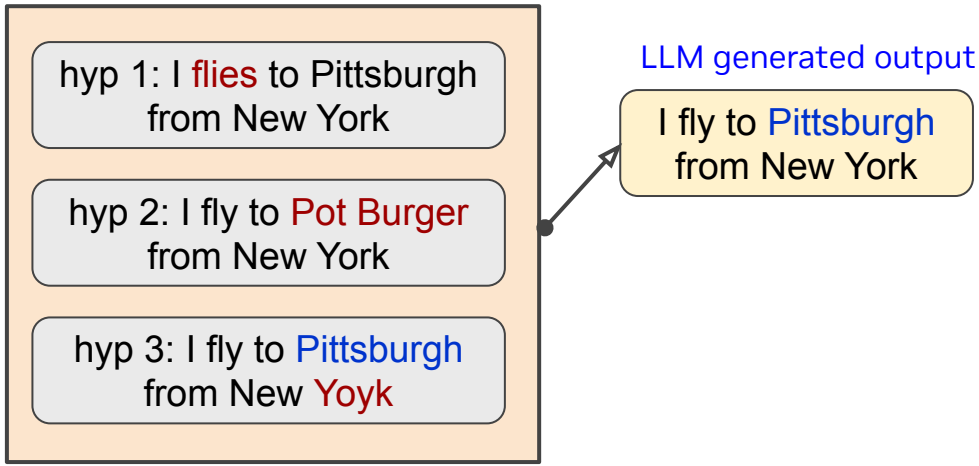
3.1 - Can ASR-LLM beyond ASR correction?

From human recognition to grounded texts

- Input: *N*-best decoding hypotheses
- Output: Re-constructed *1-best* transcription (i.e., ASR, MT, OCR)

➔ **Discriminative** Modeling
- ◆ Utterances Re-Ranking

➔ **Generative** Modeling (GER)
- ◆ Many-to-One Utterance Generation

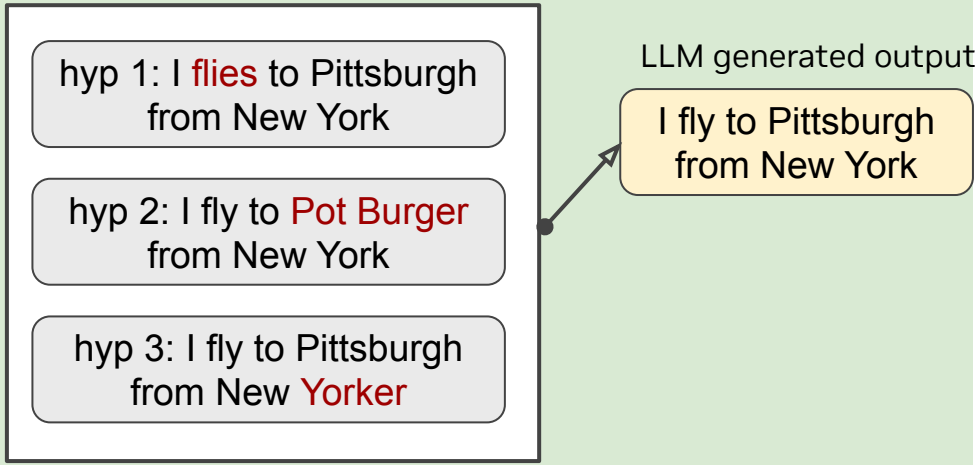


Yang et al. ASRU 23
Chen et al. NeurIPS 23

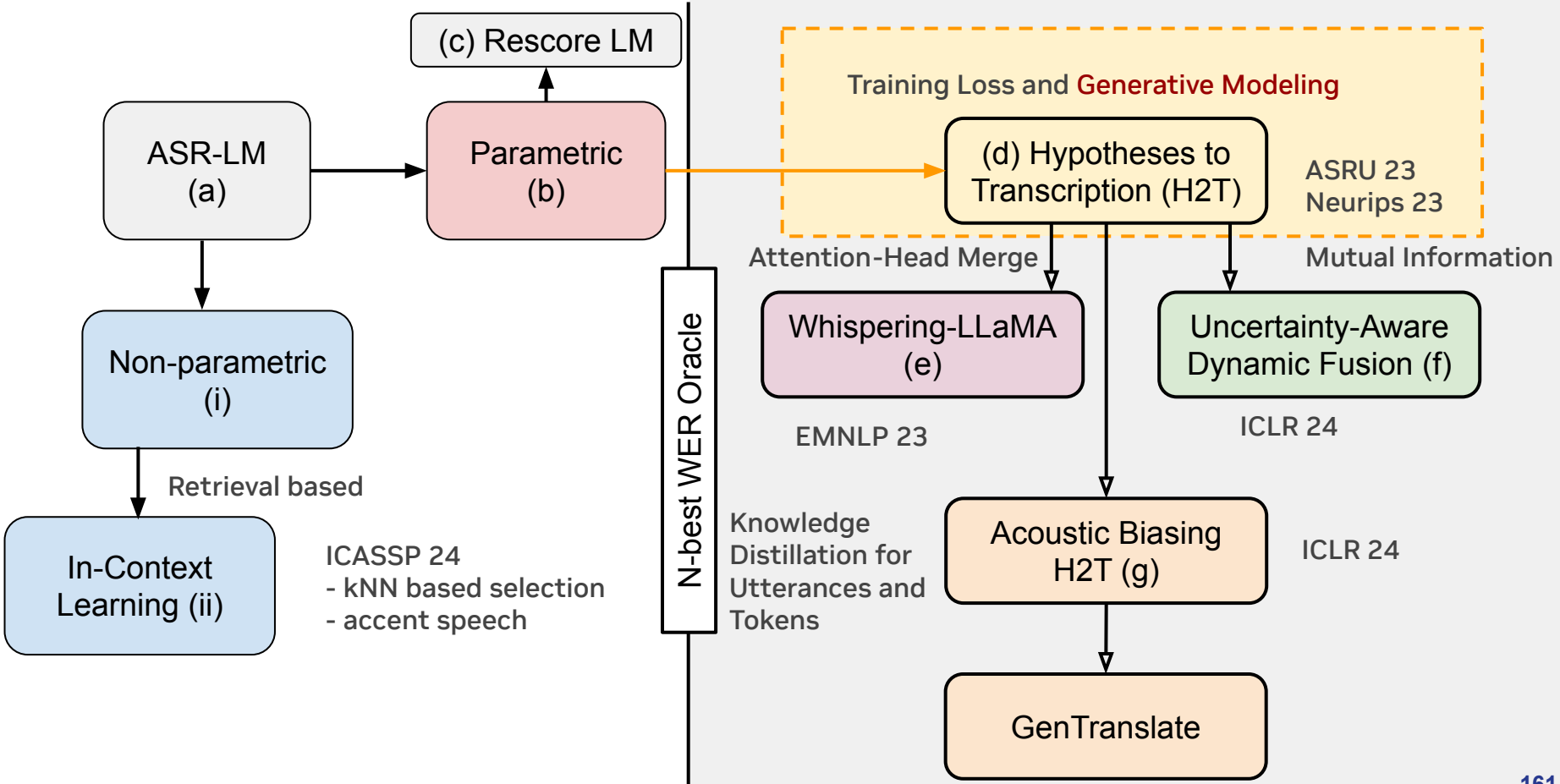
Generative Semantics Modeling (1/3) Post-ASR LM

- **Discriminative** Modeling
  - Utterances Re-Ranking

- **Generative** Modeling (ours)
  - Many-to-One Utterance Generation

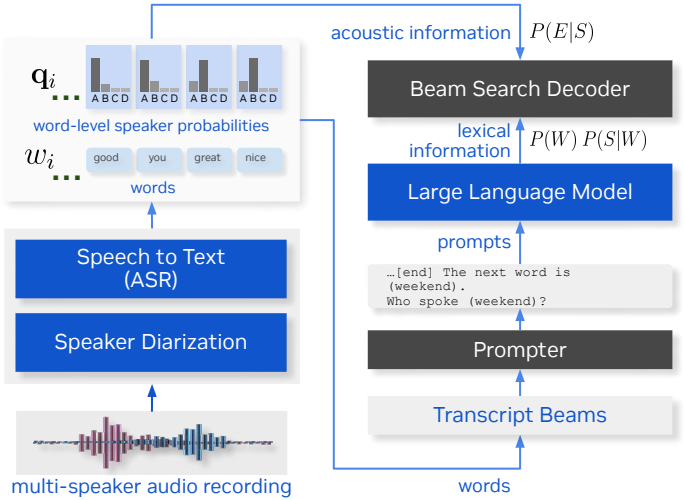


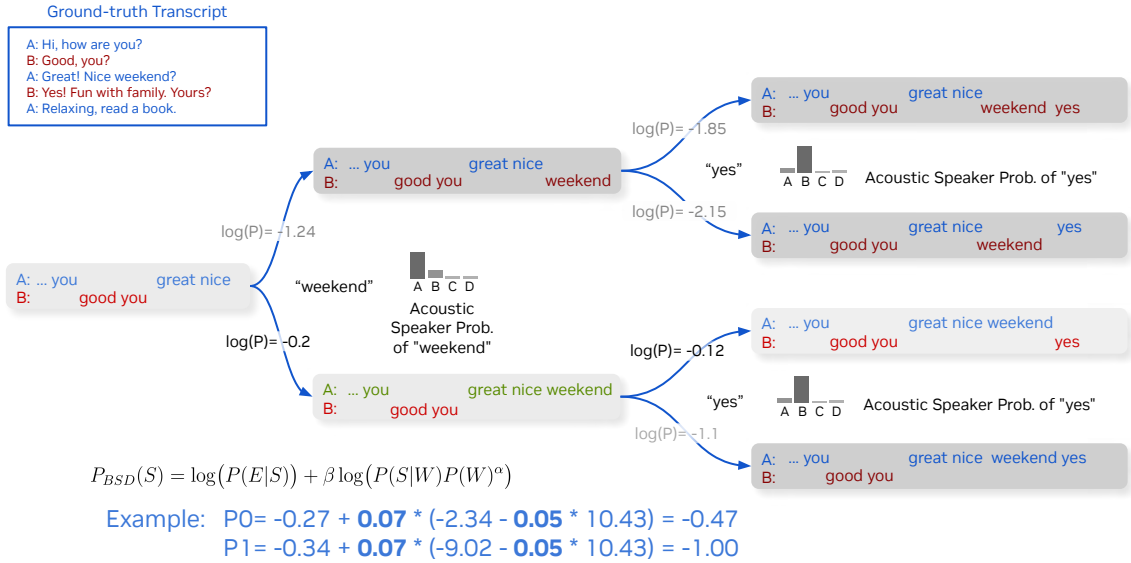
3.2 Roadmap of ASR-LLM (2022 to 24)



Post-ASR Correction (2/3) Multi-Speakers

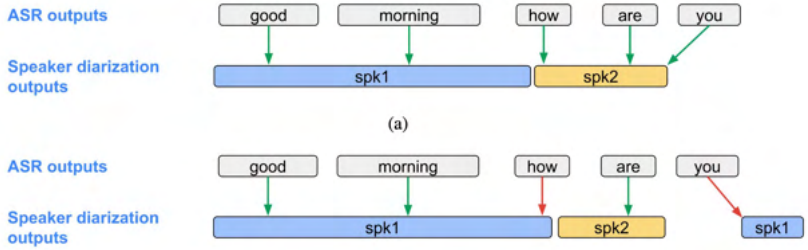
- Multi-speaker ASR Correction with LLM - Without Fine-tuning LLMs





Post-ASR Correction (3/3) Multi-Speakers LMs

- Multi-speaker ASR Correction with LM - By Fine-tuning LLMs



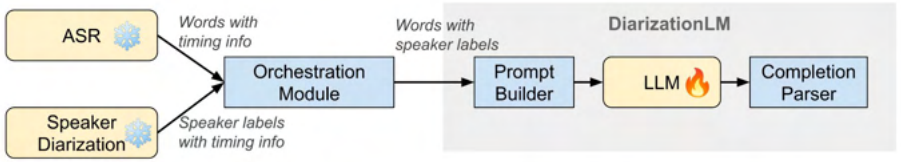


Figure 2: Diagram of the proposed DiarizationLM framework.

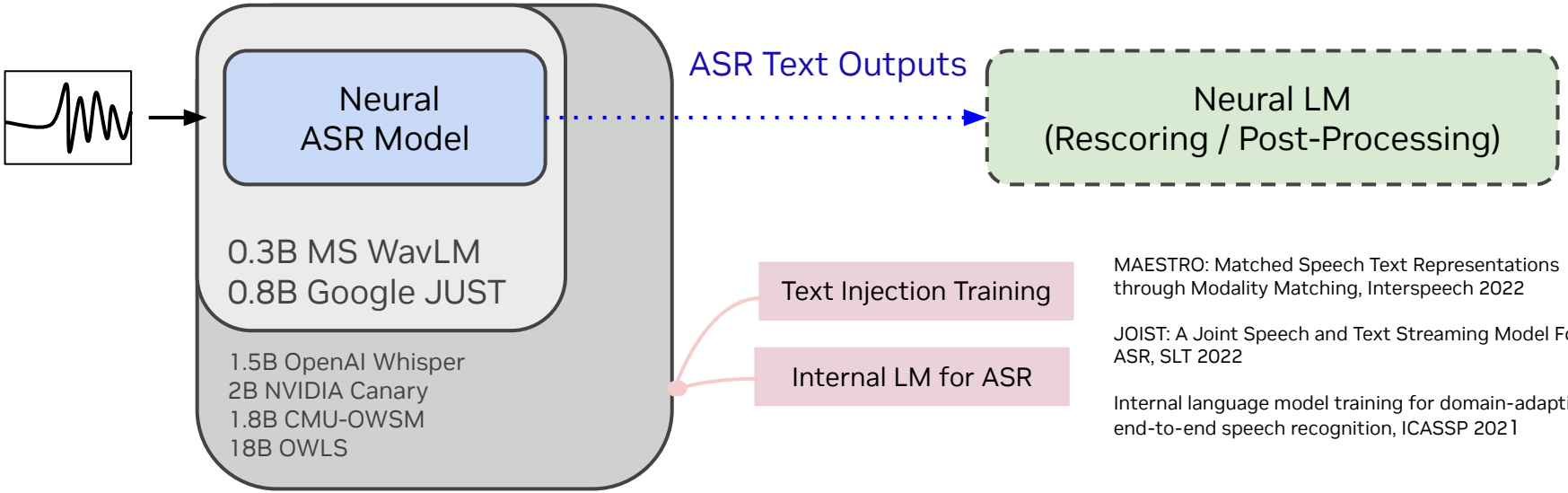
Example text representation of the prompt

```
Word sequence:        ["good", "morning", "how", "are", "you"]
Speaker sequence:     [1, 1, 2, 2, 2]
Text representation:  "<spk:1> good morning <spk:2> how are you"
```

Wang, Quan, et al. "DiarizationLM: Speaker Diarization Post-Processing with Large Language Models." Interspeech 2024.

3.1 - ASR-LM Designs (1/2) from 2017 to 2023

- Do we really need a cascaded LM in end-to-end **streaming ASR** model?
  - **Arguments**: *"Transformer-Transducer based **ASR-decoder** already contains contextual information"*

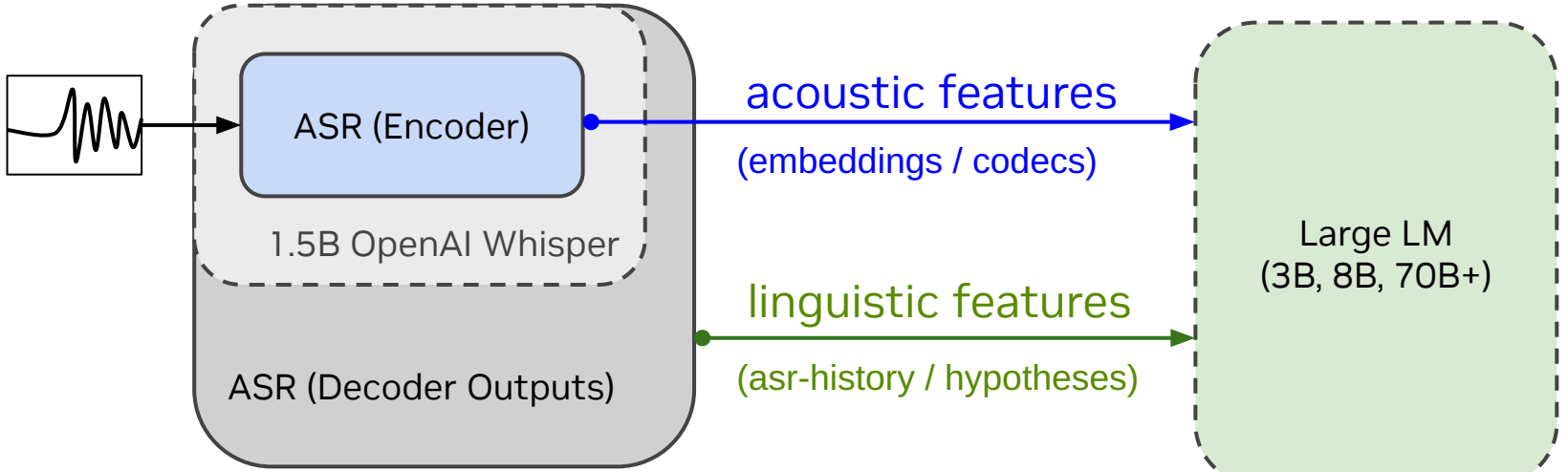


MAESTRO: Matched Speech Text Representations through Modality Matching, Interspeech 2022

JOIST: A Joint Speech and Text Streaming Model For ASR, SLT 2022

Internal language model training for domain-adaptive end-to-end speech recognition, ICASSP 2021

3.1 - ASR-LLM Designs (2/2) from 2023 to now



🧠 Agentic ASR-LLM
→ best performance

Task-activating ASR-LLM [ASRU 23]
Whispering-LLaMA [EMNLP 23]
GenTranslate [ACL 24]
NeKo MoE [ACL 25]

🚀 Duplex SpeechLM
→ real time interaction

SoundStorm, Google 2023
Moshi, Kyutai 2024
Synchronous LLMs [EMNLP 24]

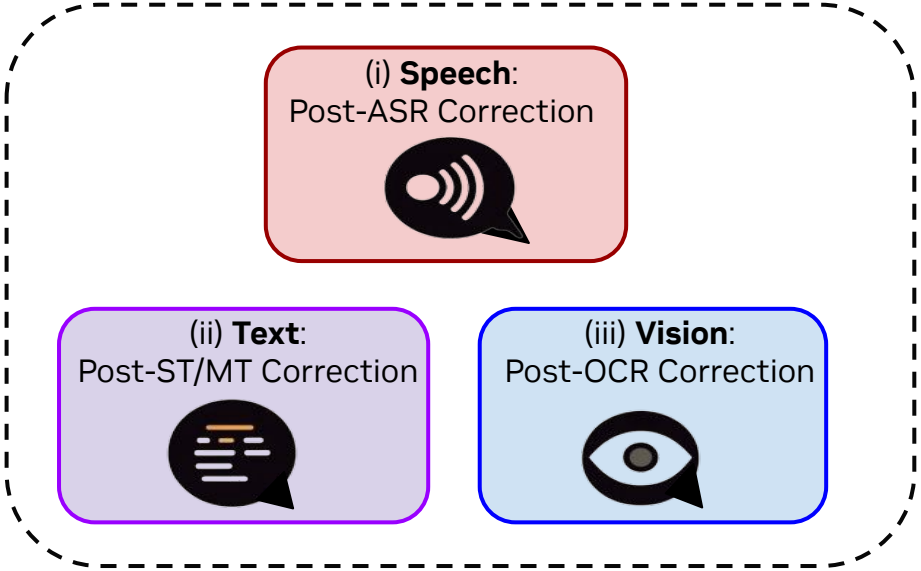
🍭 Audio-Injector LM
→ multi-task modeling

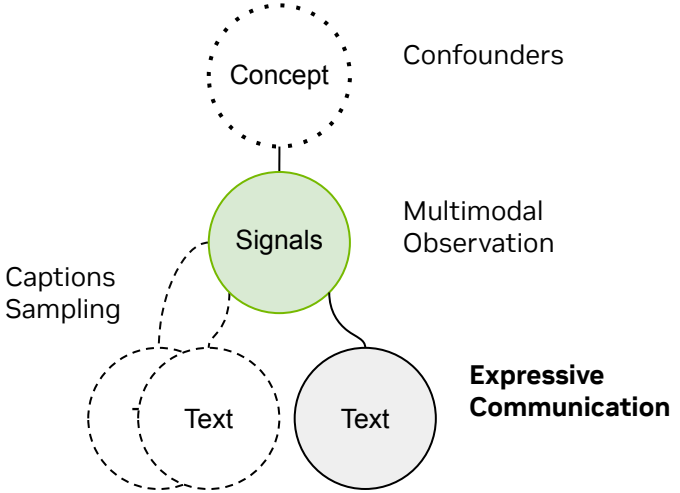
SALMON [ICLR 24]
AudioFlamingo [ICML 24]
LLaMA-3.1, Meta 2024
Gemma, Google 2024
Canary-Qwen, NVIDIA 2025

3.1 - Can ASR-LLM beyond ASR correction?

- X-to-Text Communication Tasks with **“Grounded Transcription”**
  - Automatic Speech Recognition **(ASR)**
  - Multilingual Speech Translation **(ST)** and Machine Translation **(MT)**
  - Optical character recognition **(OCR)**

Generalized Cross-modal N-best correction



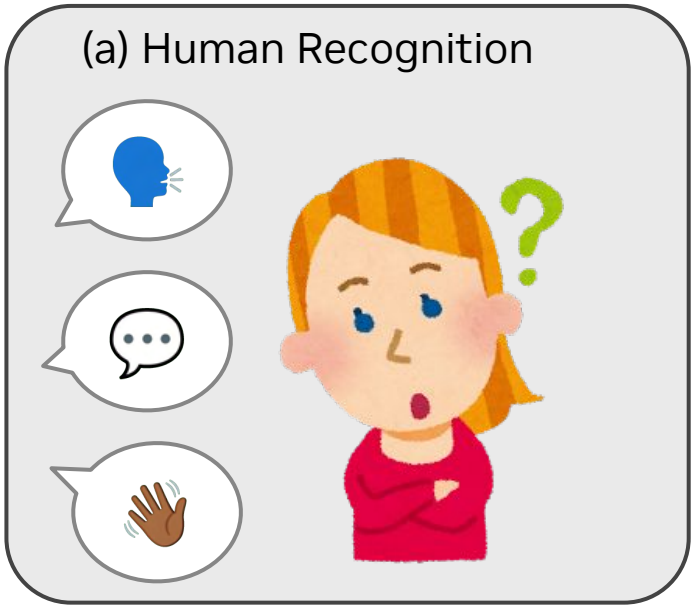


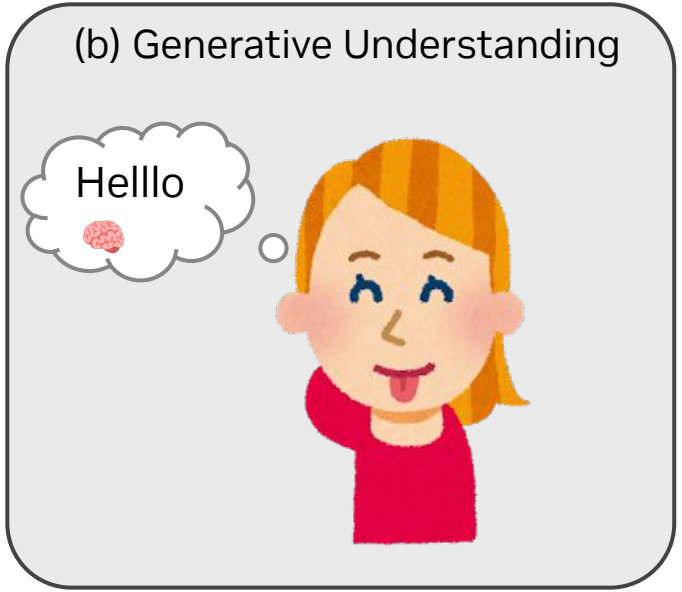
EMNLP 24 Y. Hirota et al.
EMNLP 23 D. Chan et al.

ACL 24 Y. Hu et al.
ASRU 23 C.-H. Yang et al.

3.1 - Can ASR-LLM beyond ASR correction?

- Input: *N*-best decoding hypotheses
- Output: Re-constructed 1-best transcription (i.e., ASR, MT, OCR)





SC Levinson and N Evans Time for a sea-change in linguistics: Response to comments on "The myth of language universals", Lingua, 2010

3.1 - Can ASR-LLM beyond ASR correction?

(i) **Speech** to Text

➔ Word Error Rate (the lower; the better)

Domain Shift	Test Set	Baseline	GPT-3.5 Turbo 0-shot	GPT-3.5 Turbo 5-shot	Claude-Opus 0-shot	Claude-Opus 5-shot	0-shot w/ NEKO NEKO-FFT	0-shot w/ NEKO NEKO-BTX	0-shot w/ NEKO NEKO-MoE	*Oracle* N-best
Specific Scenario	WSJ-*dev93*	9.0	$8.5_{-5.6\%}$	$7.7_{-14.4\%}$	$8.2_{-8.9\%}$	$7.4_{-17.8\%}$	$8.6_{-4.4\%}$	$7.5_{-16.7\%}$	**6.8**$_{-24.4\%}$	6.5
	WSJ-*eval92*	7.6	$7.3_{-3.9\%}$	$6.6_{-13.2\%}$	$7.0_{-7.9\%}$	$6.3_{-17.1\%}$	$7.4_{-2.6\%}$	$6.4_{-15.8\%}$	**5.8**$_{-23.7\%}$	5.5
	ATIS	5.8	$5.5_{-5.2\%}$	$5.0_{-13.8\%}$	$5.2_{-10.3\%}$	$4.7_{-19.0\%}$	$5.6_{-3.4\%}$	$4.8_{-17.2\%}$	**4.2**$_{-27.6\%}$	3.5
Common Noise	CHiME4-*bus*	18.8	$17.6_{-6.4\%}$	$16.2_{-13.8\%}$	$17.1_{-9.0\%}$	$15.7_{-16.5\%}$	$17.7_{-5.9\%}$	$15.9_{-15.4\%}$	**14.5**$_{-22.9\%}$	16.8
	CHiME4-*caf*	16.1	$14.7_{-8.7\%}$	$13.7_{-14.9\%}$	$14.2_{-11.8\%}$	$13.2_{-18.0\%}$	$14.8_{-8.1\%}$	$13.4_{-16.8\%}$	**12.2**$_{-24.2\%}$	13.3
	CHiME4-*ped*	11.5	$10.9_{-5.2\%}$	$9.7_{-15.7\%}$	$10.5_{-8.7\%}$	$9.3_{-19.1\%}$	$11.0_{-4.3\%}$	$9.5_{-17.4\%}$	**8.6**$_{-25.2\%}$	8.5
	CHiME4-*str*	11.4	$10.9_{-4.4\%}$	$9.7_{-14.9\%}$	$10.5_{-7.9\%}$	$9.3_{-18.4\%}$	$11.0_{-3.5\%}$	$9.4_{-17.5\%}$	**8.5**$_{-25.4\%}$	9.0
Speaker Accent	MCV-*af*	25.3	$24.9_{-1.6\%}$	$23.6_{-6.7\%}$	$24.4_{-3.6\%}$	$23.0_{-9.1\%}$	$25.0_{-1.2\%}$	$23.3_{-7.9\%}$	**21.0**$_{-17.0\%}$	23.6
	MCV-*au*	25.8	$25.1_{-2.7\%}$	$24.0_{-7.0\%}$	$24.6_{-4.7\%}$	$23.4_{-9.3\%}$	$25.2_{-2.3\%}$	$23.7_{-8.1\%}$	**21.4**$_{-17.1\%}$	24.9
	MCV-*in*	28.6	$27.6_{-3.5\%}$	$25.0_{-12.6\%}$	$27.0_{-5.6\%}$	$24.3_{-15.0\%}$	$27.8_{-2.8\%}$	$24.6_{-14.0\%}$	**22.2**$_{-22.4\%}$	27.1
	MCV-*sg*	26.4	$26.5_{+0.4\%}$	$25.1_{-4.9\%}$	$25.9_{-1.9\%}$	$24.5_{-7.2\%}$	$26.6_{+0.8\%}$	$24.7_{-6.4\%}$	**22.3**$_{-15.5\%}$	25.5

3.1 - Can ASR-LLM beyond ASR correction?

(i) **Speech** to Text

➔ Word Error Rate (the lower; the better)

Model	Inference Para.	Avg. ↓	AMI	Earnings22	Gigaspeech	LS Clean	LS Other	SPGI	Tedlium	Voxp.	MCV9
ASR or SpeechLMs: End-to-end Voice Understanding Models											
Distil-Whisper-V2-L (Gandhi et al., 2023)	0.75B	8.31	14.65	12.12	10.31	2.95	6.39	3.28	4.30	8.22	12.60
Whisper-V2-L (Radford et al., 2022)	1.5B	8.06	16.82	12.02	10.57	2.56	5.16	3.77	4.01	7.50	10.11
Canary (NVIDIA, 2024)	2B	6.67	14.00	12.25	10.19	**1.49**	**2.49**	**2.06**	3.58	**5.81**	7.75
Bestow Speech LM (Chen et al., 2024c)	1.8B	**6.50**	**12.58**	12.86	10.06	1.64	3.07	2.11	**3.41**	5.84	**6.97**
Qwen2-Audio (Chu et al., 2024)	8B	7.43	-	-	-	1.6	3.6	-	-	-	-
Gemini-2.0-Flash	-	8.56	-	-	-	-	-	-	-	-	-
ASR+LLM: Frozen Whisper-v2-L (1.5B) + Voice Correction LMs											
+ Gemma 2B (Team et al., 2024) FFT	3.5B (2B)	6.61	13.20	12.30	10.40	1.60	2.60	2.20	3.70	6.00	7.50
+ Gemma 8x2B FFT	3.5B (2B)	6.51	13.10	12.20	10.30	1.50	2.50	2.10	3.60	5.90	7.40
+ NEKO (Ours) Gemma 8x2B	3.5B (2B)	6.41	13.00	12.10	10.20	1.40	2.40	2.00	3.50	5.80	7.30
+ NEKO (Ours) Qwen1.5-MoE	4.2B (2.7B)	**5.90**	12.60	**11.82**	**9.95**	**1.30**	**2.32**	**1.94**	**3.20**	5.80	7.30
+ Mistral 7B (Jiang et al., 2023) FFT	8.5B (7B)	6.40	13.07	11.87	10.09	1.48	2.46	2.04	3.55	**5.75**	7.29
+ Mixtral 8x7B (Jiang et al., 2024b) FFT	8.5B (7B)	6.51	12.91	12.19	10.34	1.54	2.55	2.12	3.64	5.89	7.43
+ Mixtral 8x7B Lora	8.5B (7B)	6.60	12.96	12.24	10.38	1.55	2.56	2.13	3.66	5.92	7.47
+ Mistral 8x7B BTM (Sukhbaatar et al., 2024)	8.5B (7B)	6.43	13.13	11.93	10.14	1.49	2.47	2.05	3.57	5.78	7.33
+ NEKO (Ours) Mixtral 8x7B	8.5B (7B)	**6.34**	**12.55**	**11.82**	10.02	1.49	2.47	2.05	3.52	5.76	**7.25**
+ NEKO (Ours) Mixtral 8x22B	23.5B (22B)	6.40	12.61	11.93	10.15	1.52	2.51	2.09	3.58	5.82	7.33

3.1 - Can ASR-LLM beyond ASR correction?

(ii) **Text to Text**

Cascaded agentic LLMs

En→X	FLEURS							CoVoST-2				MuST-C			
	Es	Fr	It	Ja	Pt	Zh	Avg.	Fa	Ja	Zh	Avg.	Es	It	Zh	Avg.
End-to-end ST Methods															
SeamlessM4T-Large (Barrault et al., 2023a)	23.8	41.6	23.9	21.0	40.8	28.6	30.0	18.3	24.0	34.1	25.5	**34.2**	**29.9**	16.2	26.8
GenTranslate (Hu et al., 2024c)	**25.4**	**43.1**	**25.5**	**28.3**	**42.4**	**34.3**	**33.2**	**21.1**	**29.1**	**42.8**	**31.0**	33.9	29.4	**18.5**	**27.3**
SeamlessM4T-Large-V2 (Barrault et al., 2023b)	23.8	42.6	24.5	21.7	43.0	29.5	30.9	16.9	23.5	34.6	25.0	32.1	**27.5**	15.6	25.1
GenTranslate-V2 (Hu et al., 2024c)	**25.5**	**44.0**	**26.3**	**28.9**	**44.5**	**34.9**	**34.0**	**19.4**	**29.0**	**43.6**	**30.7**	**32.2**	27.3	**18.1**	**25.9**
Cascaded ASR+MT Methods															
Whisper + NLLB-3.3b (Costa-jussà et al., 2022)	25.1	41.3	25.0	19.0	41.5	23.5	29.2	13.6	19.0	32.0	21.5	35.3	29.9	13.5	26.2
SeamlessM4T-Large (ASR+MT) (Barrault et al., 2023a)	24.6	44.6	25.4	22.5	41.9	31.2	31.7	18.8	24.0	35.1	26.0	35.1	30.8	17.7	27.9
SeamlessM4T-V2 (ASR+MT) (Barrault et al., 2023b)	24.7	44.1	25.1	20.6	43.6	30.6	31.5	17.4	23.8	35.4	25.5	33.0	27.8	14.5	25.1
Cascaded ASR+GEC Methods															
GenTranslate	26.8	45.0	26.6	29.4	43.1	36.8	34.6	21.8	30.5	43.3	31.9	35.5	31.0	19.6	28.7
GenTranslate-V2	27.0	44.3	26.4	27.8	44.5	36.1	34.4	20.8	29.7	43.5	31.3	33.2	28.3	16.9	26.1
NEKO-Gemma-2B-FT	26.9	44.2	26.3	27.7	44.4	36.0	34.3	20.7	29.6	43.4	31.2	33.1	28.2	16.8	26.0
NEKO-Gemma-8x2B–BTX	27.2	44.5	26.7	28.0	44.7	36.3	34.6	21.0	29.9	43.8	31.6	33.4	28.5	17.1	26.3
NEKO-Gemma-8x2B-MoE	**28.5**	**46.2**	**28.0**	**30.1**	**46.3**	**38.7**	**36.3**	**23.4**	**32.6**	**46.5**	**34.2**	**37.2**	**32.8**	**21.5**	**30.5**

NeKo

➔ GPT-4o-mini-2411 BLEU CoVoST-2 (~33.2) | MuST-C (~29.2)

3.1 - Vision - OCR Correction

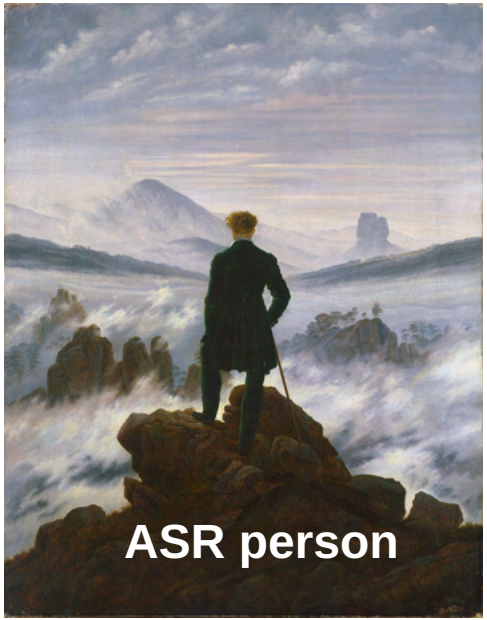
Cascaded agentic LLMs

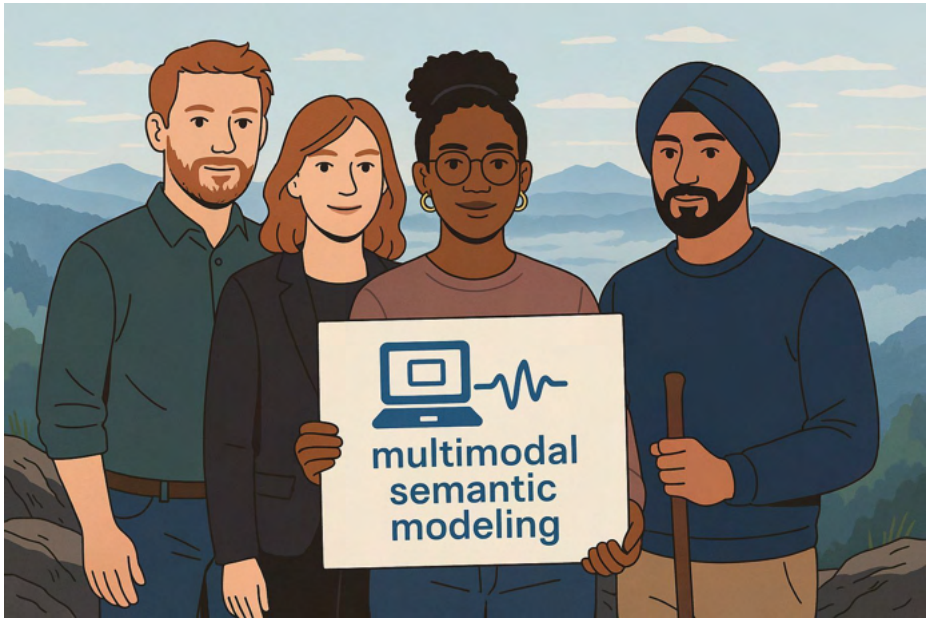
(iii) **Vision**:
Post-OCR Correction

NeKo 🐱

Task / WER ↓	Grammar Correction	Coherence Improv.	OCR
Mixtral-MoE (frozen)	31.41	13.48	71.03
GPT-3.5-turbo	17.43	12.25	39.45
Mixtral-MoE-FFT	10.73	12.05	45.32
NEKO-Mixtral-MoE	9.42	9.71	14.43

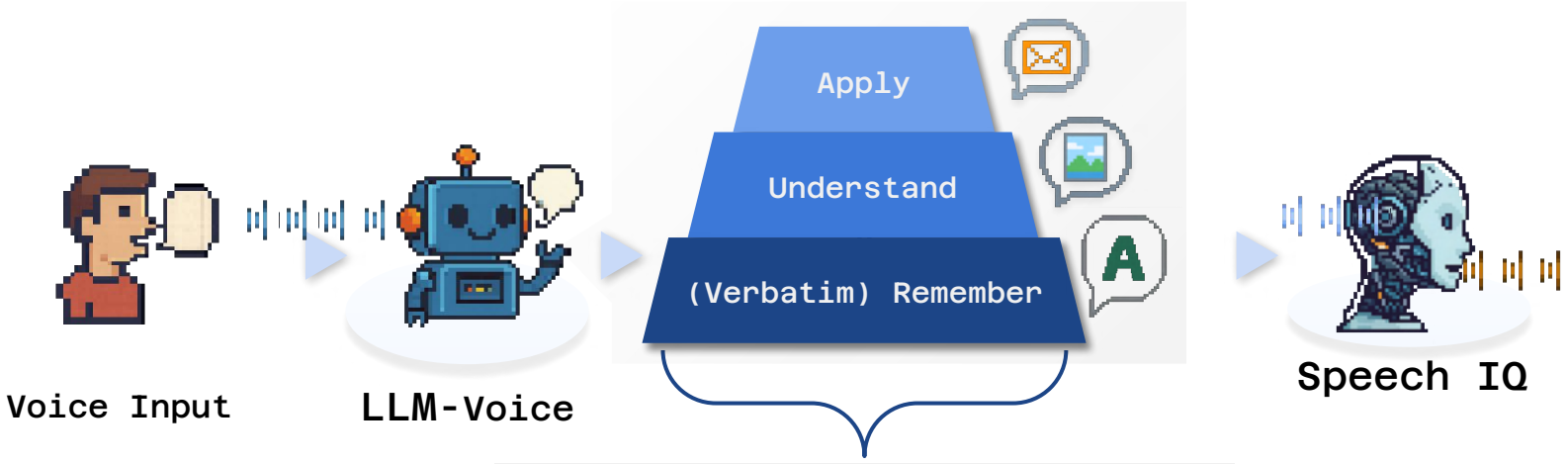
3.1 semantic ASR-LM: what's the next





Der Wanderer über dem Nebelmeer

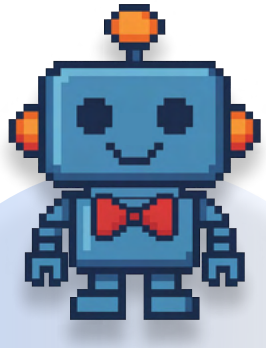
Recent Benchmark-LLM Responses



Relative Levels of Bloom's Taxonomy

- **SpeechIQ:** Speech Intelligence Quotient Across Cognitive Levels in Voice Understanding Large Language Models, ACL 2025
- **Dynamic-superb phase-2:** A collaboratively expanding benchmark for measuring the capabilities of spoken language models with 180 tasks, ICLR 2025

Semantic Alignment in Large Models II



Large Models

(i.e., Superior to Human Abilities)

"super listening → i love u"

text

Semantics

Sect 3.1 (10 min)

voice

Human Cognition Model

Cross-modality

➡️ **Sect 3.2 (10 min)**

feedbacks

Preference & Values

Sect 3.3 (10 min)

output

3.2 How to Inject Acoustic Info into LLM?

- Hidden **semantic information in the audio** modality
  - How to ground sound, accent speech, and music into text?

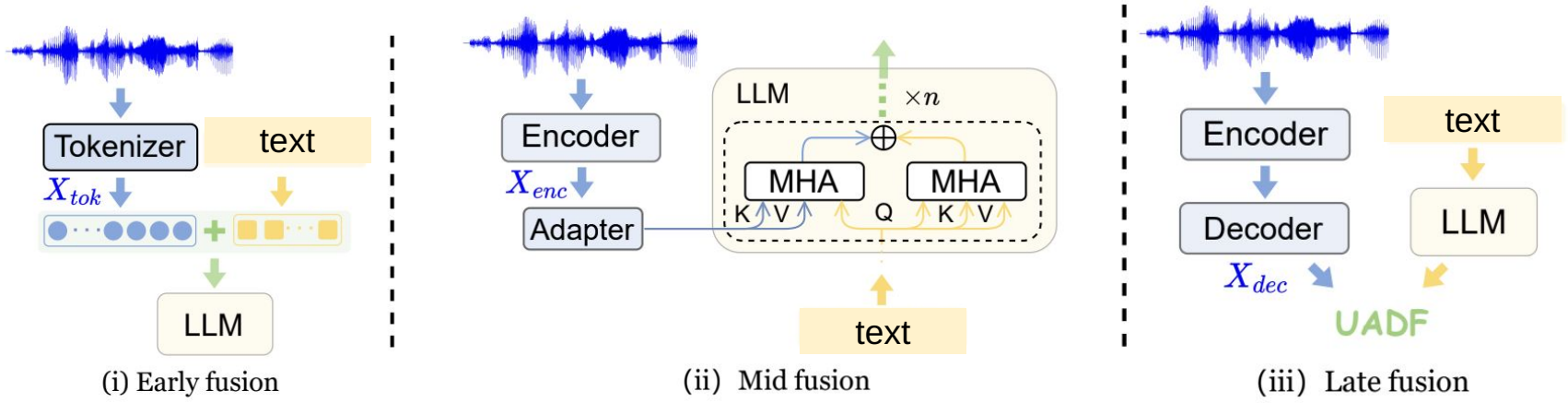


Image Source: It's Never Too Late: Fusing Acoustic Information into Large Language Models for Automatic Speech Recognition, ICLR 24

3.2 Different Semantic Aligned Speech-LMs

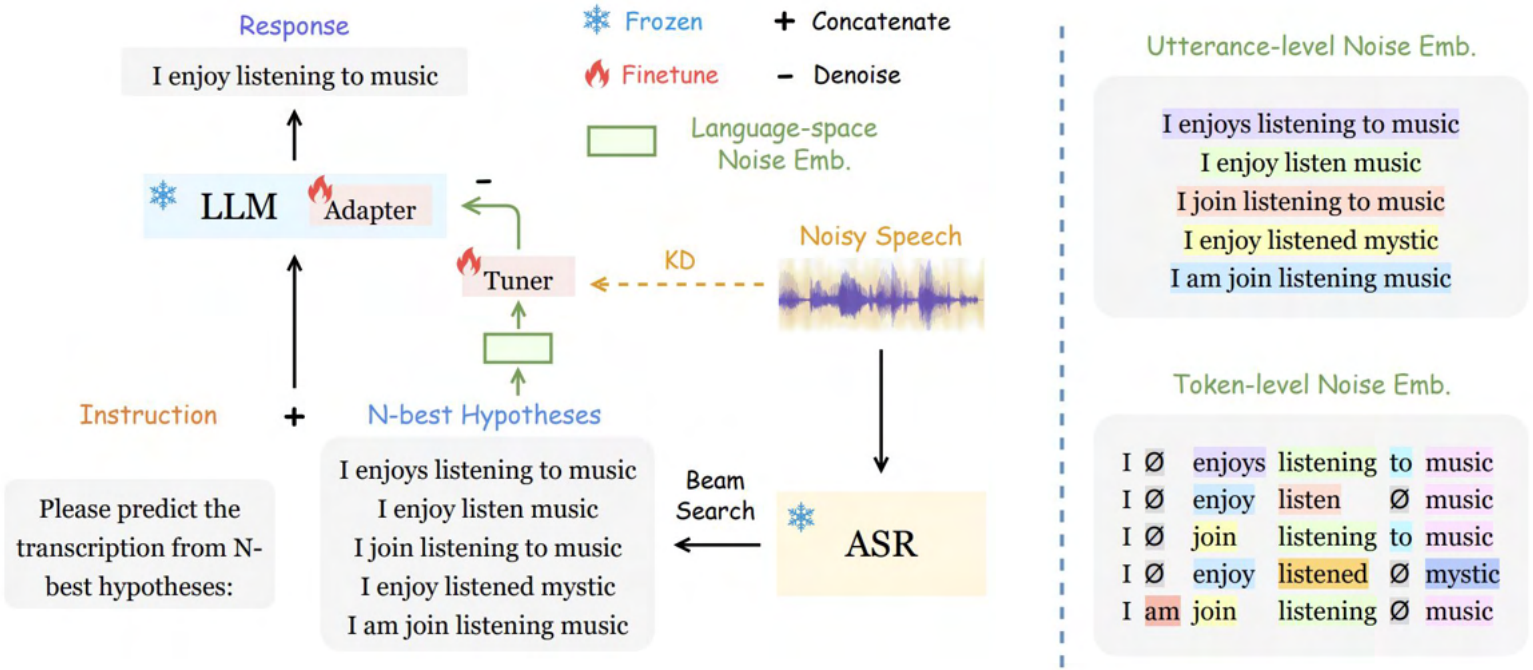
- **early fusion**: audio embedding injection into text space
  - Qwen-audio, and most of the audioLMs
  - more post-training fusion methods

- **codec based**: compressing signals into discrete codecs

- **cascaded agent**: audio model chained LM

Generative Semantics Modeling (2/3) Language

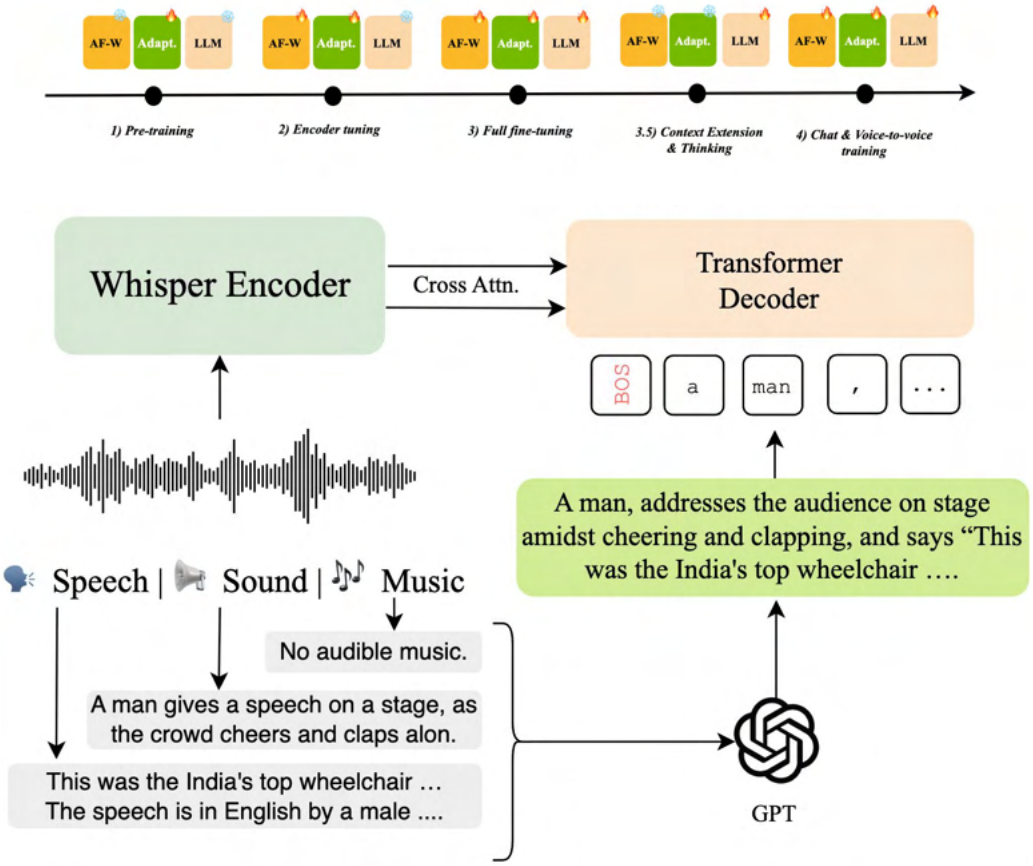
Denoising



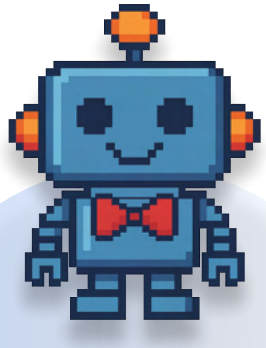
Y. Hu, C. Chen, **C.-H. H. Yang**, et al. LLMs are Efficient Learners of Noise-Robust Speech Recognition, *ICLR 2024*

Generative Semantics Modeling (3/3) Audio-LMs

- Audio Flamingo 3
  - Applied Deep Learning Research (ADLR)
  - Arushi Goel, Sreyan Ghosh, et al.
  - Open AudioLM
    - https://huggingface.co/nvidia/audio-flamingo-3



Semantic Alignment in Large Models III



Large Models

(i.e., Superior to Human Abilities)

"super listening → i love u"

text

Semantics

Sect 3.1 (10 min)

voice

Human Cognition Model

Cross-modality

Sect 3.2 (10 min)

feedbacks

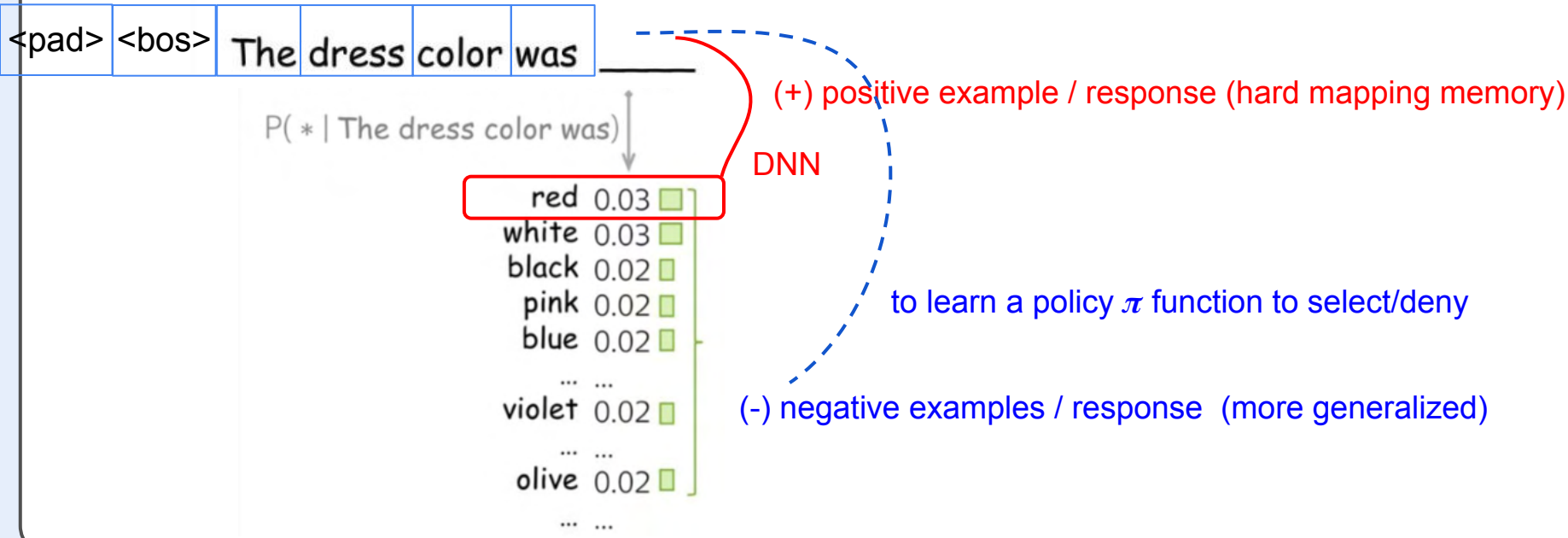
Preference & Values

➡️ **Sect 3.3 (10 min)**

output

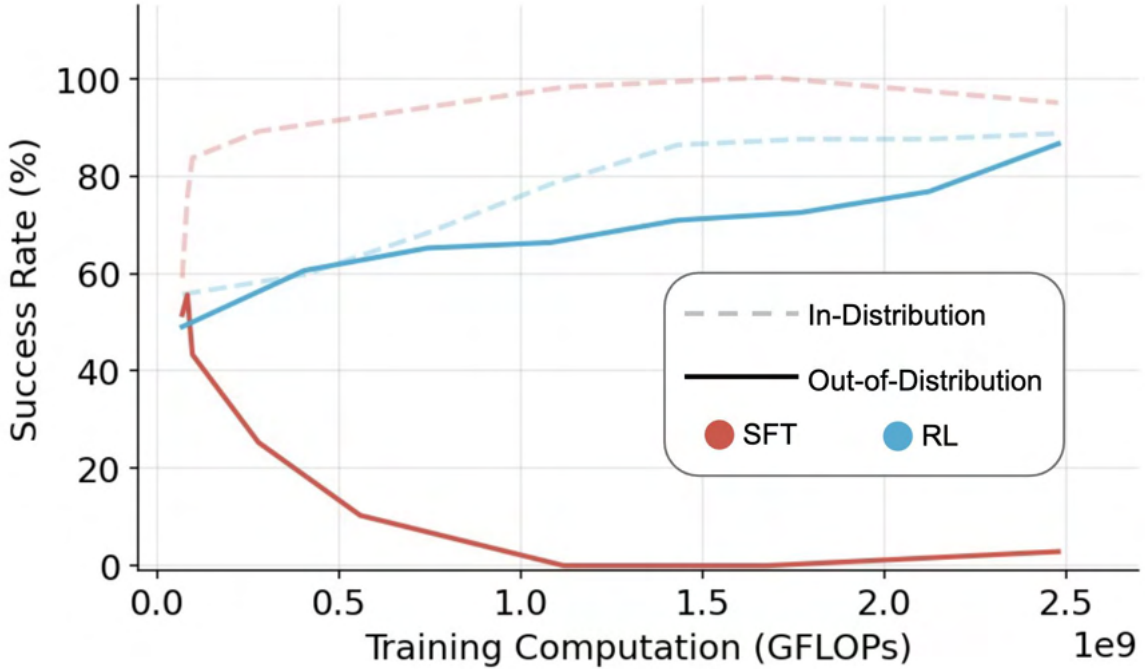
Supervised Fine-Tuning (SFT) (1/3): positive examples

- SFT has “limit” on predicting the next token only on "positive examples"
  - an example on causal LM training



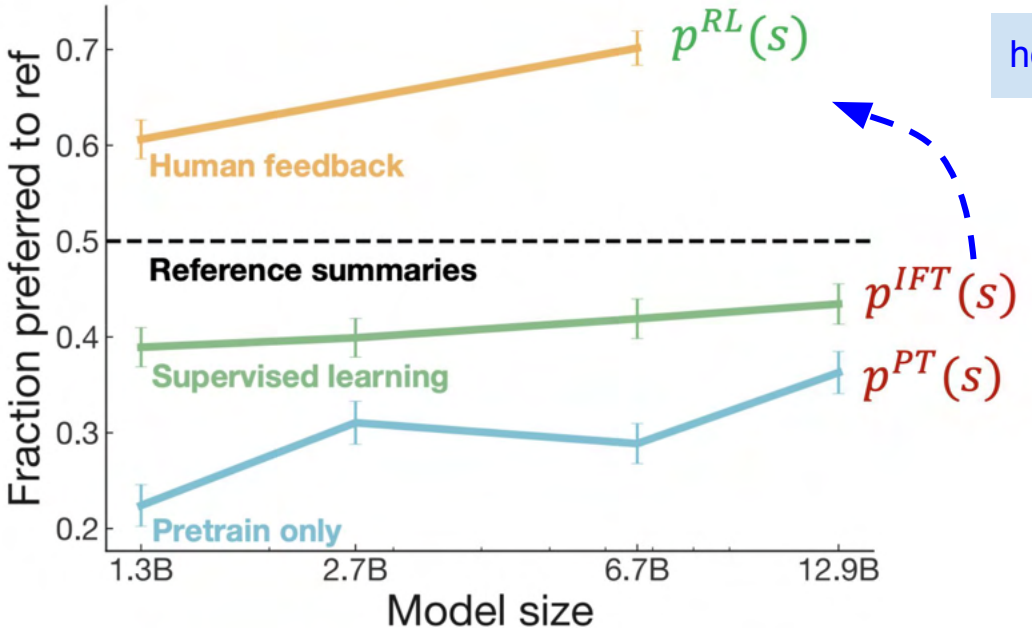
Causal LM Training (2/3): negative examples

- SFT Memorizes, RL Generalizes [ICML 25]
  - We need both SFT and RL, which generalizes in SFT-DPO or SFT-GRPO.



Causal LM Training (3/3): stage-wise training

- *"Learning to summarize from human feedback [NeurIPS 20]"* OpenAI
  - stage-wise training is essential before RL-stage



how reward model works?

stage 2 to stage 3

stage 1 to stage 2

How Reward Model (RM) works in reasoning LMs?

- How preference information can be learned from negative information.

$$J_{RM}(\phi) = -\mathbb{E}_{(s^w, s^l)\sim D}\left[\log \sigma(RM_\phi(s^w) - RM_\phi(s^l))\right]$$

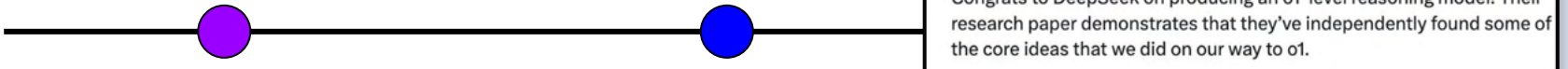
"winning" sample

"losing" sample

s^w should score higher than s^l

→ Reward Model needs preference data but how data further generalizes the reasoning ability

How Data work in the Reasoning LMs



gpt-o1-preview

Sep 24: unknown method

deepseek-r1

Jan 25: <thinking> format compresses CoTs and GRPO

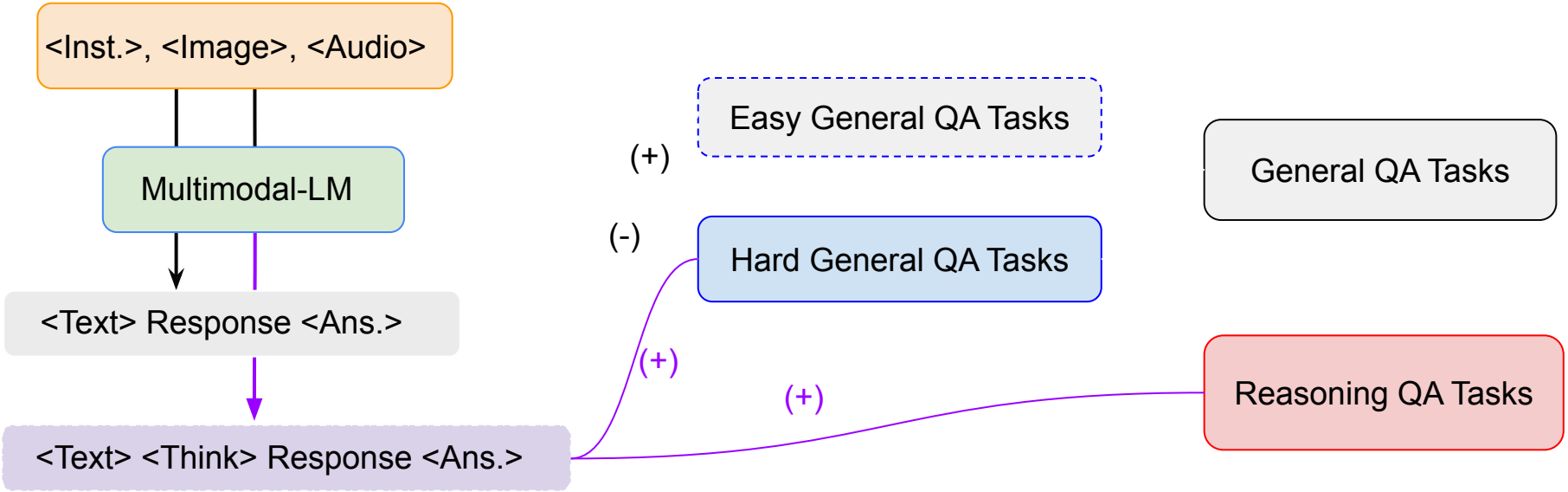
> Congrats to DeepSeek on producing an o1-level reasoning model! Their research paper demonstrates that they've independently found some of the core ideas that we did on our way to o1.
>
> 1:11 PM · Jan 28, 2025 · **8.1M** Views

A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively, i.e., <think> reasoning process here </think> <answer> answer here </answer>. User: prompt. Assistant:

Compressing Chain of Thoughts (CoTs; demo searching traces)

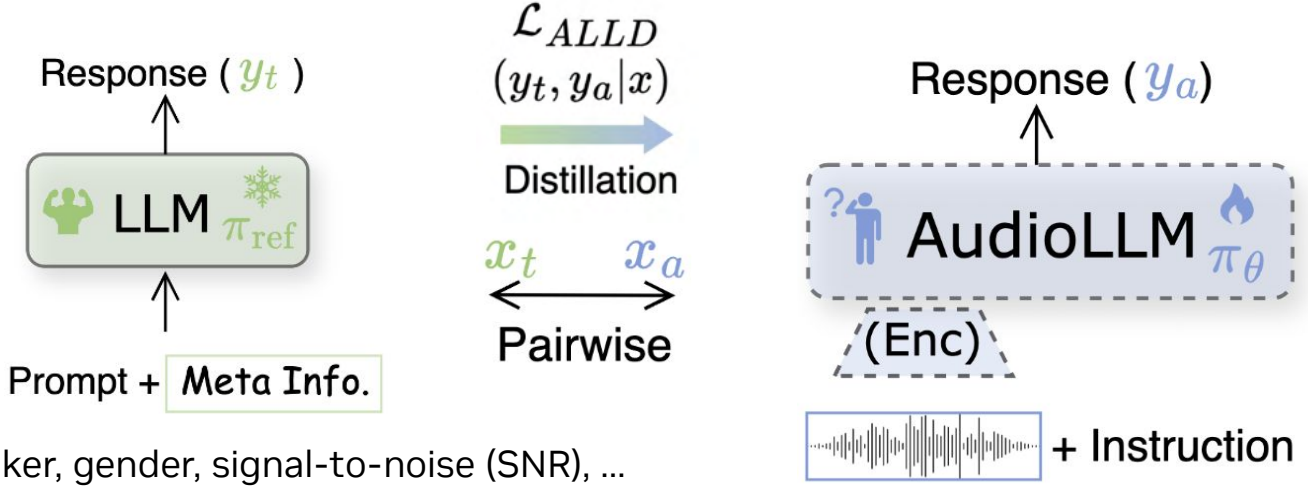
Why we need “Reasoning Ability”?

- Training Objectives
  - (1) Long <think> responses helps on hard and reasoning QA tasks
  - (2) When generate <think> token



Audio LLMs can be Voice Quality Descriptors (1/4)

- Can Audio LLM learning from synthesized feedback data
  - Textual Caption view of audio signals
  - A/B testing for reward format data

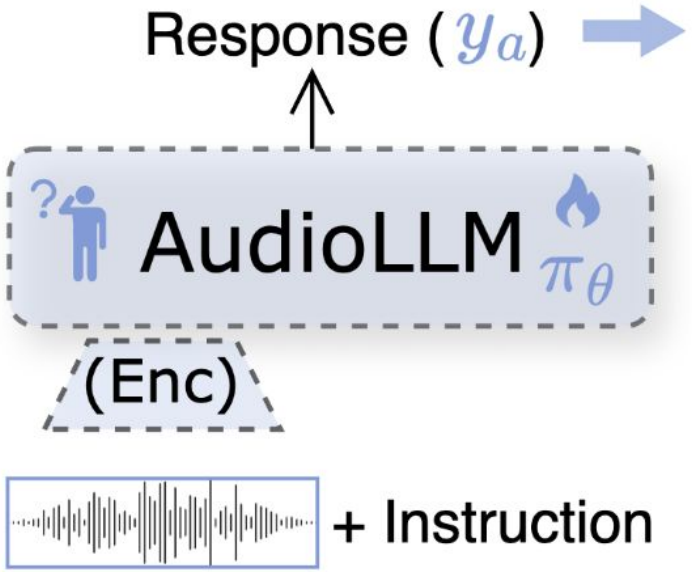


speaker, gender, signal-to-noise (SNR), ...

Source: Audio Large Language Models Can Be Descriptive Speech Quality Evaluators, ICLR 25

Audio LLMs can be Voice Quality Text Descriptors (2/4)

- How to make proper data format?



SFT-target

Exp1: Evaluate the quality of 

Res1: This given speech has slight distortion, but the ... ..., the overall MOS score is about ***2.4****.*

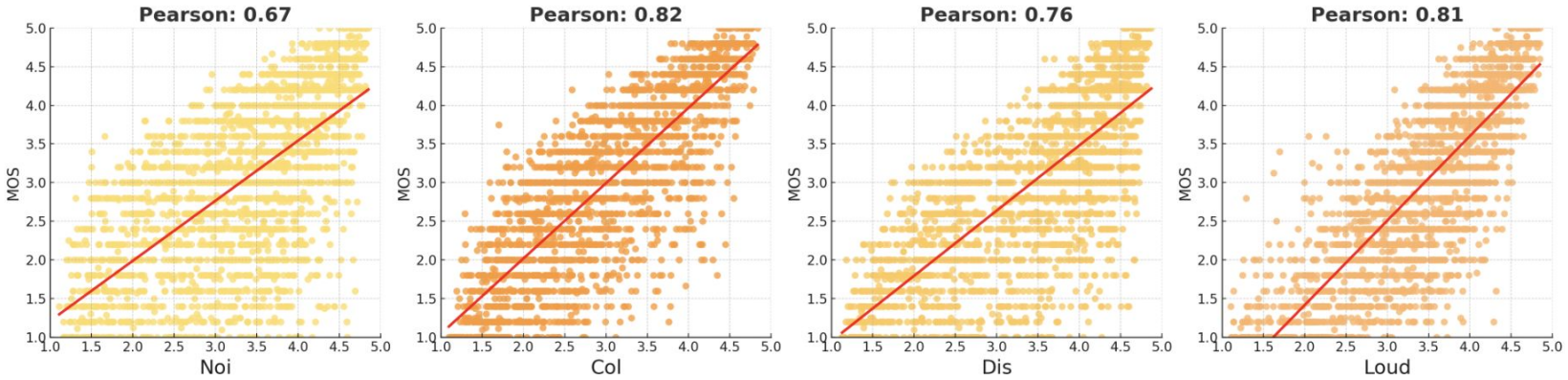
DPO-target

Exp2: A/B test for  and 

Res2: The noise level of Speech A is slighter than ... ... , Therefore, I think Speech A has better quality.

Audio LLMs can be Voice Quality Descriptors

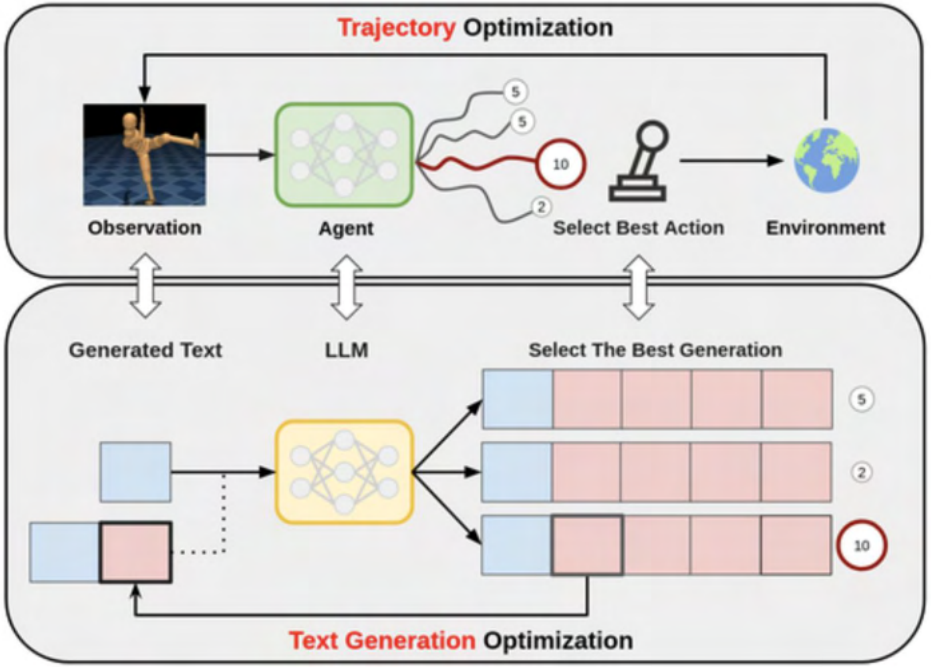
- How Audio LLM correlates to human Mean opinion score (MOS)?



Joint Training	A/B Test		MOS			
	BLEU	Acc (%)	LCC	SRCC	MSE	BLEU
✗	29.02	95.6	0.92	0.92	0.20	25.22
✓	30.17	98.6	0.92	0.91	0.20	26.08

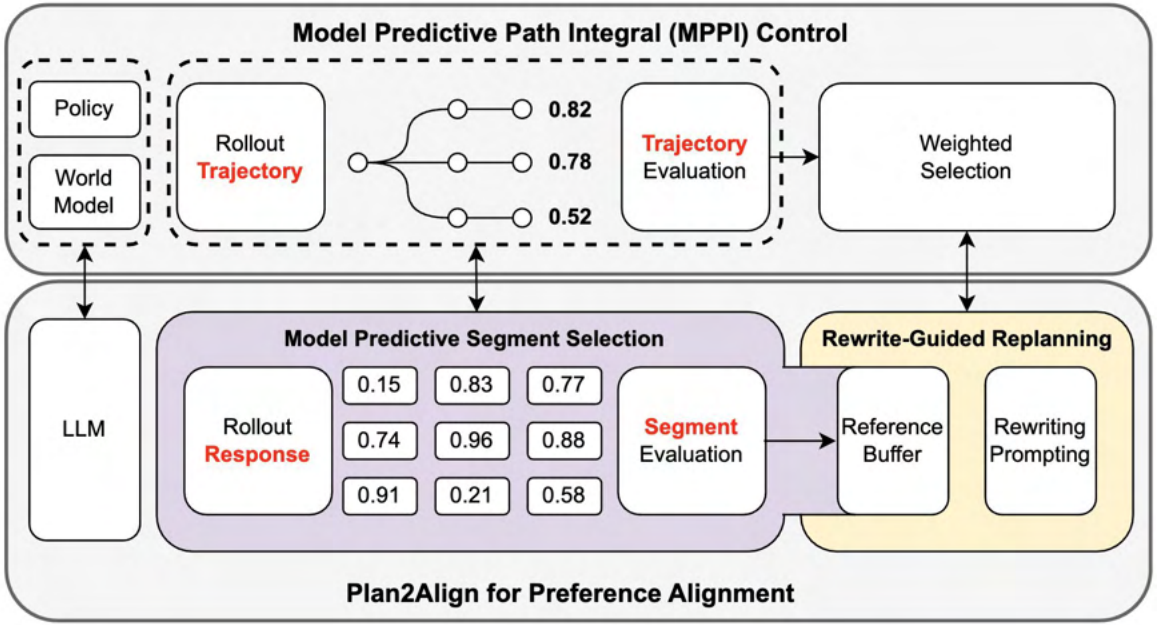
Test-Time Adaptation with Semantic Modeling (1/3)

- Plan2Align: Semantic Trajectory Modeling During Test-Time
  - Improved Machine Translation Results by X%



Test-Time Adaptation with Semantic Modeling (2/3)

- Plan2Align: Semantic Trajectory Modeling During Test-Time
  - Improved Machine Translation Results by X%



Test-Time Adaptation with Semantic Modeling (3/3)

- Plan2Align: Semantic Trajectory Modeling During Test-Time
  - Improved Machine Translation Results by X%

Methods	Test-Time	zh → en		zh → ru		zh → de	
		CW-COMET ↑	CW-KIWI ↑	CW-COMET ↑	CW-KIWI ↑	CW-COMET ↑	CW-KIWI ↑
GPT-4o $_{2024\text{-}08\text{-}06}$	-	94.58	73.06	93.74	54.20	94.54	54.55
Qwen-2.5 (14B)	-	94.43	72.44	90.47	50.13	92.98	50.02
Llama-3.1 (8B)	×	84.36	58.27	86.28	34.32	88.97	39.49
Llama-3.1$_{SFT}$	×	**93.54**	**67.16**	**89.11**	**43.99**	93.47	**47.39**
Llama-3.1$_{SimPO}$	×	91.74	62.31	84.56	41.66	93.40	46.47
Llama-3.1$_{DPO}$	×	90.23	62.09	82.15	38.91	**93.48**	46.09
Llama-3.1$_{RAIN}$	✓	58.52	36.39	66.29	31.75	67.43	31.69
Llama-3.1$_{ARGS}$	✓	63.99	41.62	43.03	19.57	51.97	23.23
Llama-3.1$_{Best\text{-}of\text{-}60}$	✓	90.97	65.41	84.86	**48.93**	82.74	38.55
Llama-3.1$_{Vanlia\ MPC}$	✓	88.50	57.06	73.84	31.45	90.32	**42.20**
Llama-3.1$_{Plan2Align}$	✓	**94.62**	**68.00**	**91.53**	41.62	**91.73**	41.50

Takeaways: Semantic Modeling beyond E2E-ASR

- Semantic models can recover meaning from noisy or partially corrupted speech by leveraging both text and audio cues.
- Large language models (LLMs) improve ASR outputs with post-ASR correction, speaker attribution, and emotion tagging—even without clean input.
- Contextual and multi-modal retrieval bridges gaps in ASR, supporting domain adaptation and code-switching.
- Integration of semantic modeling with **retrieval-augmented** approaches could be further explored for context-aware speech understanding.

Table of Contents

Download Slides

Introduction (10 mins) 15:30-15:40	
Shinji (30 min) 15:40-16:10	Speech-to-Text Benchmark
Taejin (40 min) 16:10-16:50	Leveraging Long Acoustic Context
Recess (10 min) 16:50-17:00	
Huck (40 min) 17:00-17:40	Semantic Context and Speech-Language Modeling
Kyu (30 min) 17:40-18:10	**Contextual Biasing and Methods Leveraging Longer Semantic Context for Speech Systems**
Closing Remark (10 min) 18:10-18:20	
Q&A Session (10 min) 18:20-18:30	

Contextual Biasing and Methods Leveraging Longer Semantic Context for ASR Systems

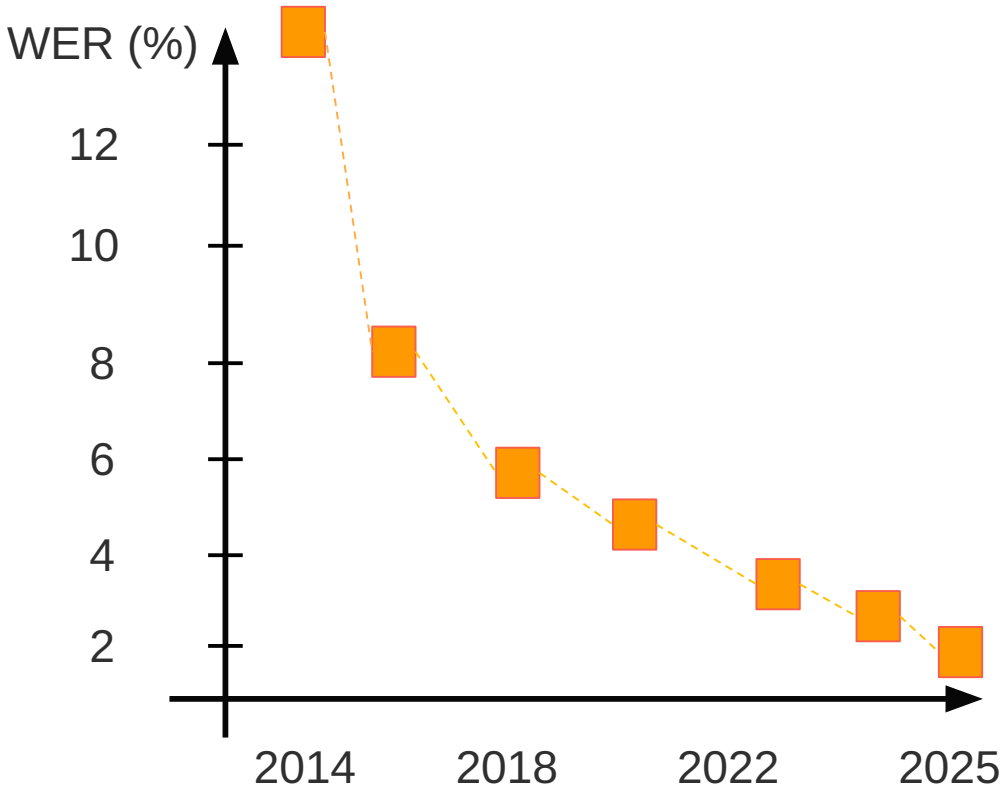
Kyu J. Han

Contextual Biasing w/ Longer Semantic Context

1. **Contextual Biasing for E2E ASR**
   a. Decoder Biasing
   b. Encoder Biasing
   c. Hybrid Biasing
   d. Retrieval Toward Large-Scale Biasing

2. **Retrieval Augmented Generation (RAG)**
   a. What is RAG?
   b. RAG for Contextual Biasing

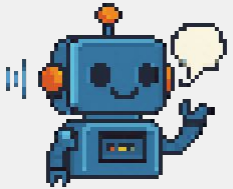
Is ASR a Solved Problem?



How Is ASR Used, Any Example?

ASR

Query



AI Drive-Thru

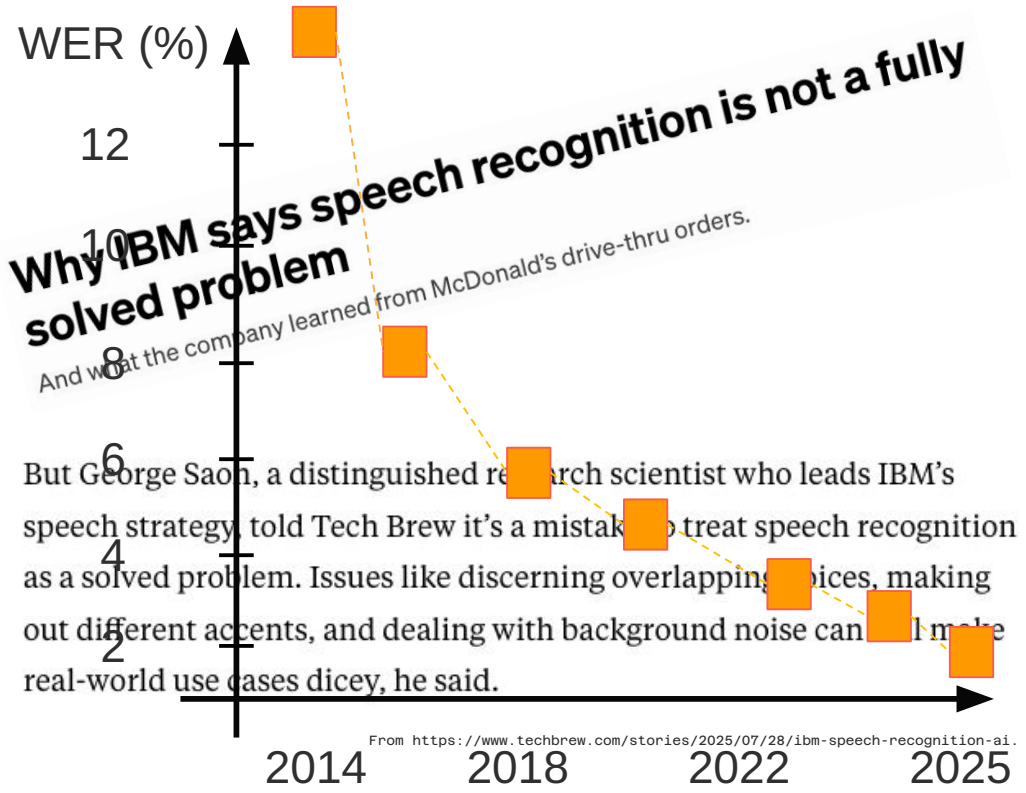
LLM

Database

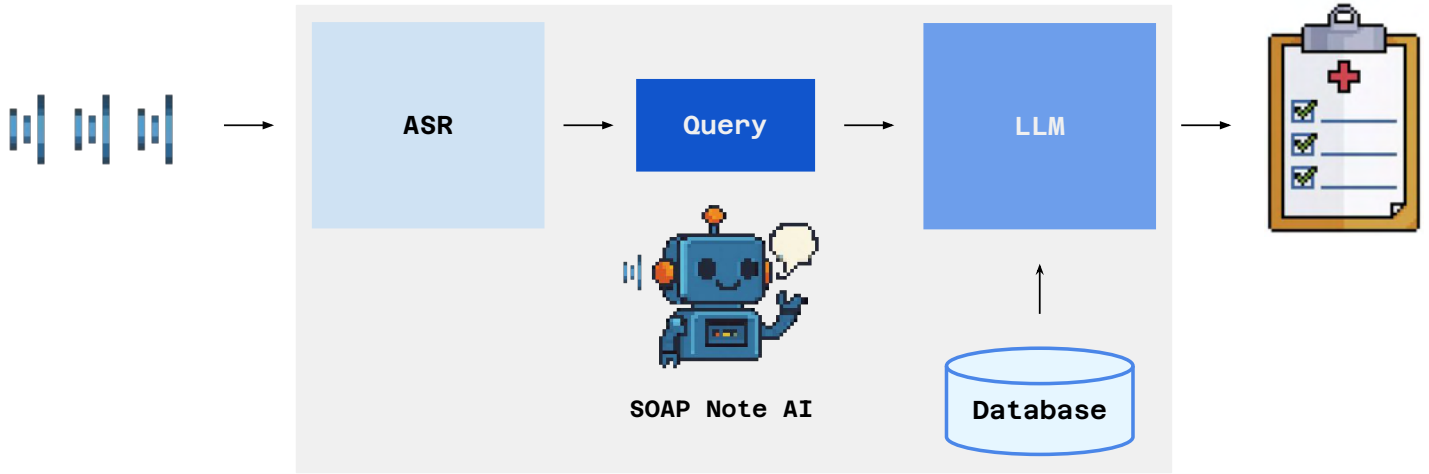
TTS

Response

Is ASR a Solved Problem?



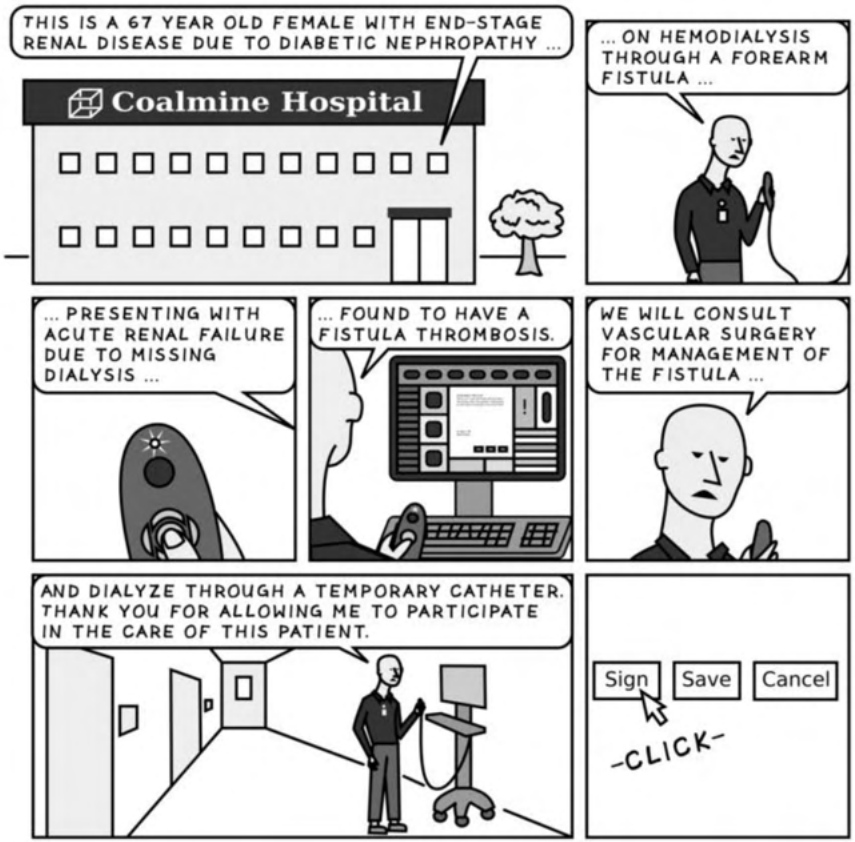
How Is ASR Used, Another Example?



Clinical note generation

- ASR + LLM for post-consultation transcription and clinical note summary
- Could save over 45% of doctors' time.

Is ASR a Solved Problem?



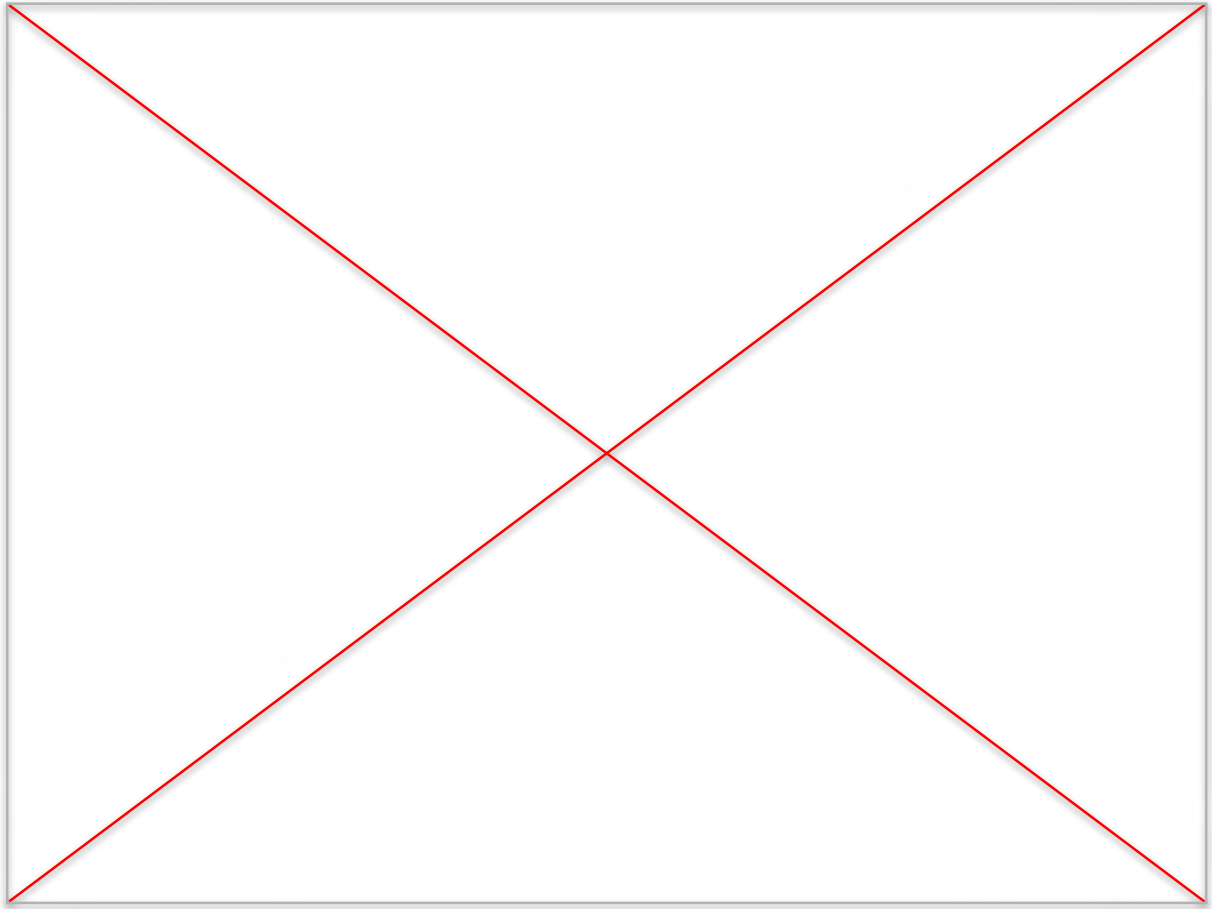
ASSESSMENT AND PLAN:
This is a 6-7 year old female with end days real disease due 2 die pathetic nephropathy, on hero dial a sis through a 4-arm fist EULA, presenting with a cute real fail your do 2 missing dial a sis appointments and found to have a fist EULA trombone sis. We will insult vascular surgery for management of the fist EULA and dial eyes through a temp urinary catheter. Thank you for allowing me to produce pain in the care of this patient.

Ian Bean, MD
Nephrologist

Electronically signed by beanian01 on 01/16/2023 at 1930.

WHATEVER. TOO LATE TO CHANGE IT NOW.

Is ASR a Solved Problem? (AWS Transcribe Example)



Contextual Biasing in E2E ASR

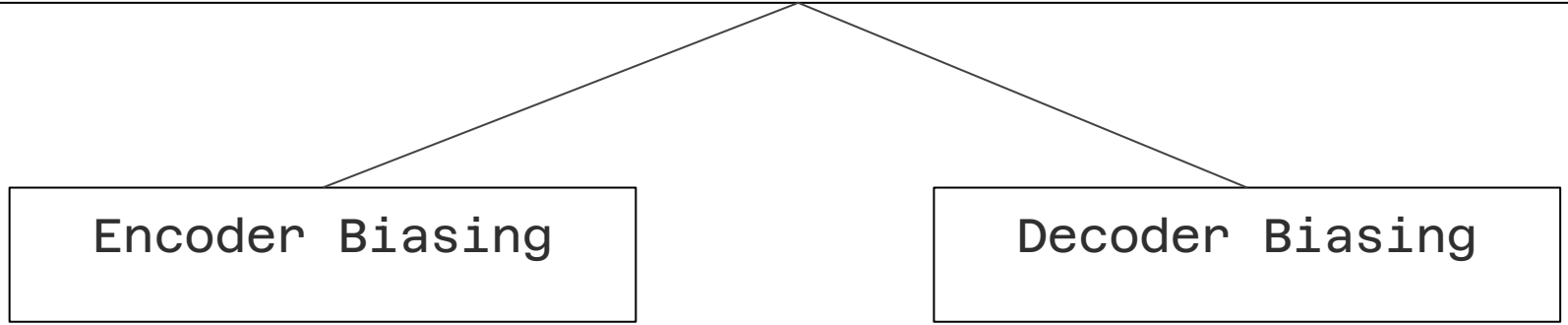
Context Biasing

Technique of influencing an ASR system ***to favor specific words or phrases*** *during decoding, often* ***because they are relevant to a particular context or user****. This is particularly useful when dealing with* ***rare words, named entities, or phrases specific to a domain****, which might otherwise be poorly recognized by a general ASR model.*

Contextual Biasing in E2E ASR

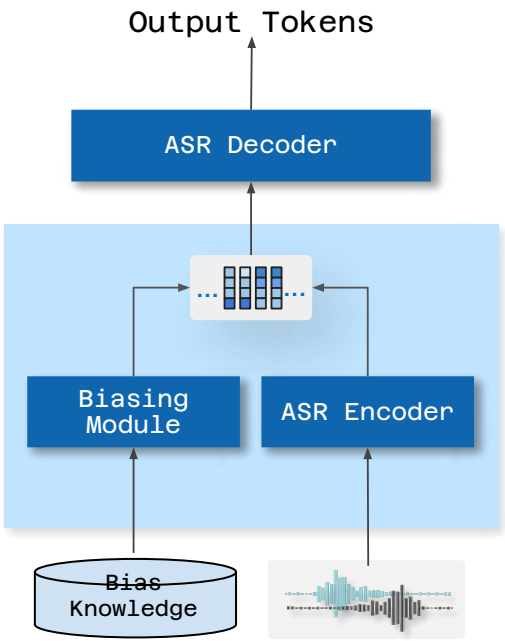
Context Biasing

Technique of influencing an ASR system ***to favor specific words or phrases*** *during decoding, often* ***because they are relevant to a particular context or user****. This is particularly useful when dealing with* ***rare words, named entities, or phrases specific to a domain****, which might otherwise be poorly recognized by a general ASR model.*

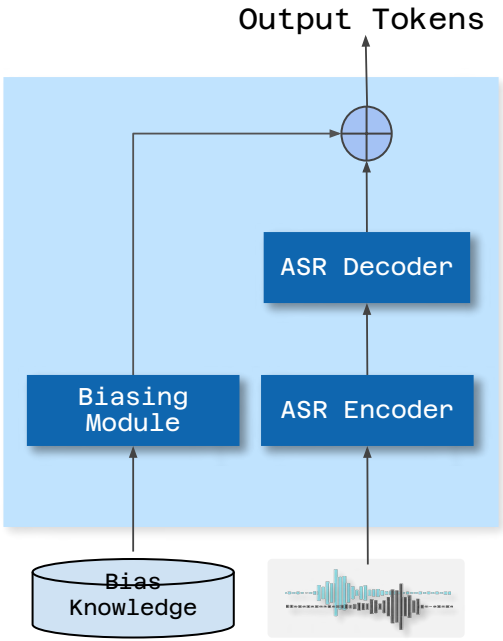


Encoder vs. Decoder Biasing

Encoder biasing



Decoder biasing

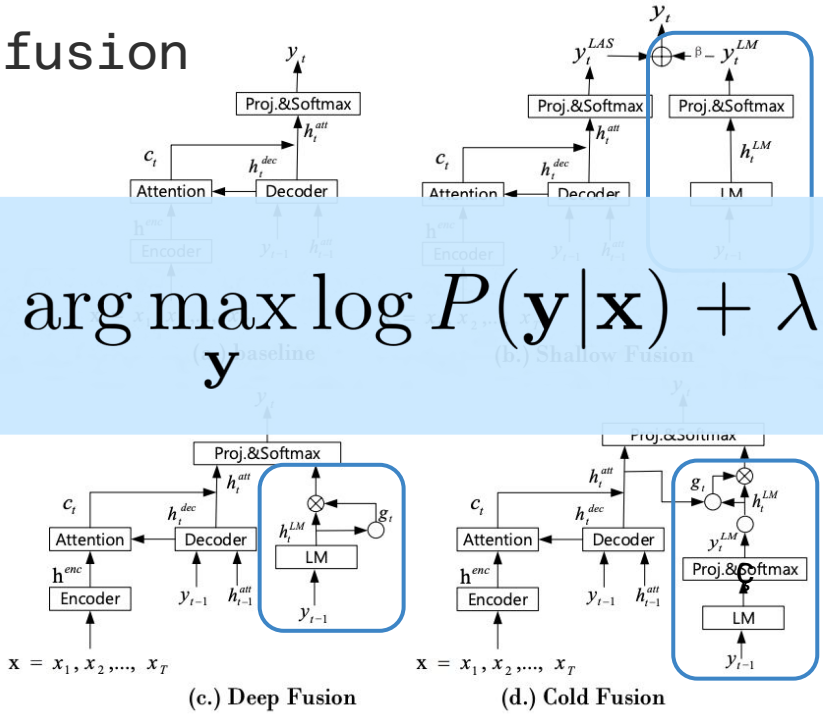


Decoder Biasing

Integrating static knowledge to ASR models to influence decoding scores on bias words

- LM shallow fusion

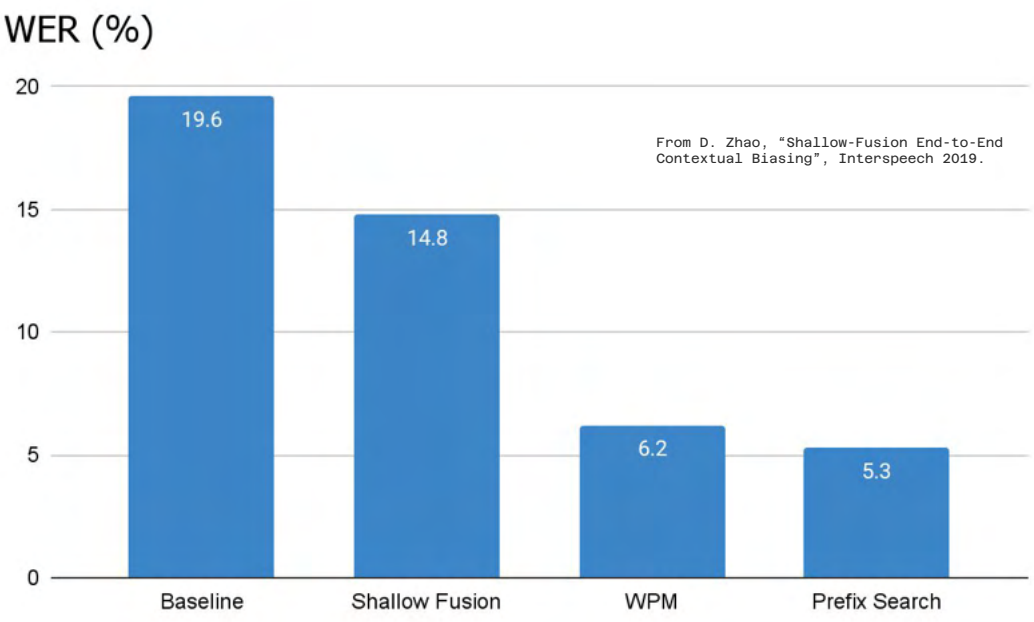
$$\mathbf{y}^* = \arg\max_{\mathbf{y}} \log P(\mathbf{y}|\mathbf{x}) + \lambda \log P_C(\mathbf{y})$$



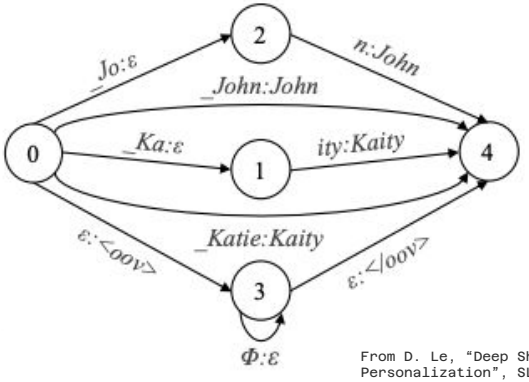
From C. Shan, et al., "Component Fusion: Learning Replaceable Language model component for End-to-End Speech Recognition System", ICASSP 2019.

Decoder Biasing

- Class-LM WFST
- Prefix tree search
- Wordpiece modeling (WPM)

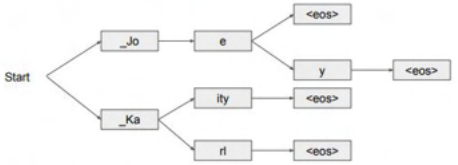


From D. Zhao, "Shallow-Fusion End-to-End Contextual Biasing", Interspeech 2019.



From D. Le, "Deep Shallow Fusion for RNN-T Personalization", SLT 2021.

Prefix tree for bias phrases [Joe, Joey, Kaity, Karl]:



Previously emitted tokens: [_call _Ka]

PLM output:

[_Jo, _Ka], Results for empty prefix

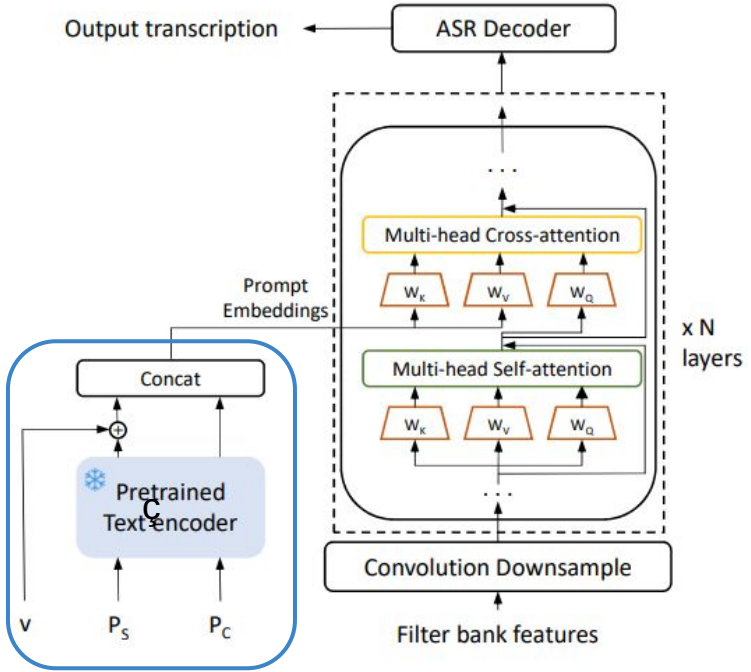
[ity, rl], Results for prefix: _Ka

[ ], Results for Prefix: _call _Ka

Encoder Biasing

Trying to affect embedding vectors directly to influence posteriori probs on bias words

Bias words	Stella, Froedtert Hospital, fistula, amoxicillin, etc.
Previous transcripts	This is a 67 year old female with end…
Prompts/Context	Today's game is between Real Madrid and Liverpool.



From X. Yang, "PromptASR for Contextualized ASR with Controllable Style", ICASSP 2024.

Contextual Biasing in E2E ASR

Context Biasing

Encoder Biasing

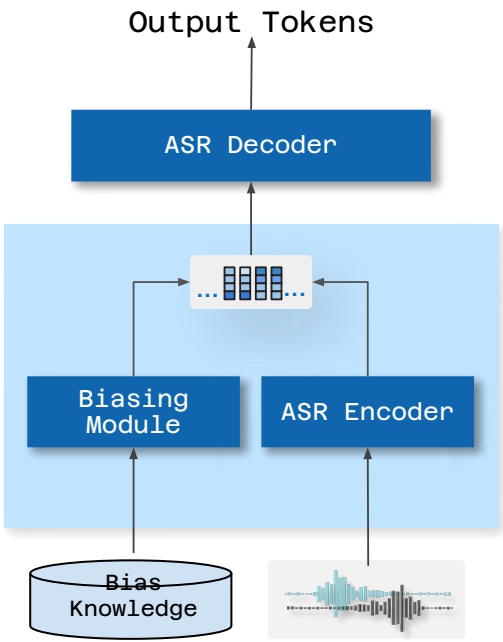
- Intrinsic
- Could lead to more modest accuracy improvement
- Reliable across datasets, languages, etc.

Decoder Biasing

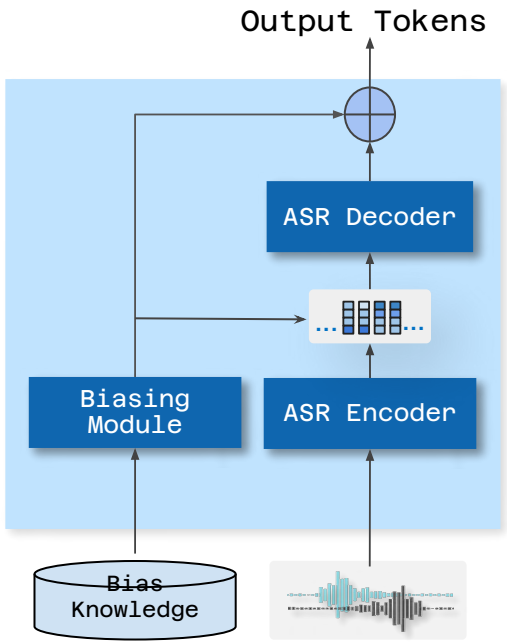
- Extrinsic
- Could result in more false positives
- Could be sensitive to datasets, languages, etc.

Hybrid Biasing

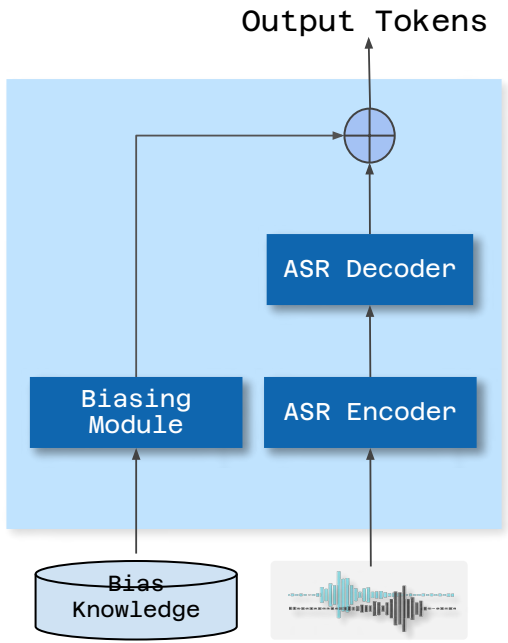
Encoder Biasing



Hybrid Biasing



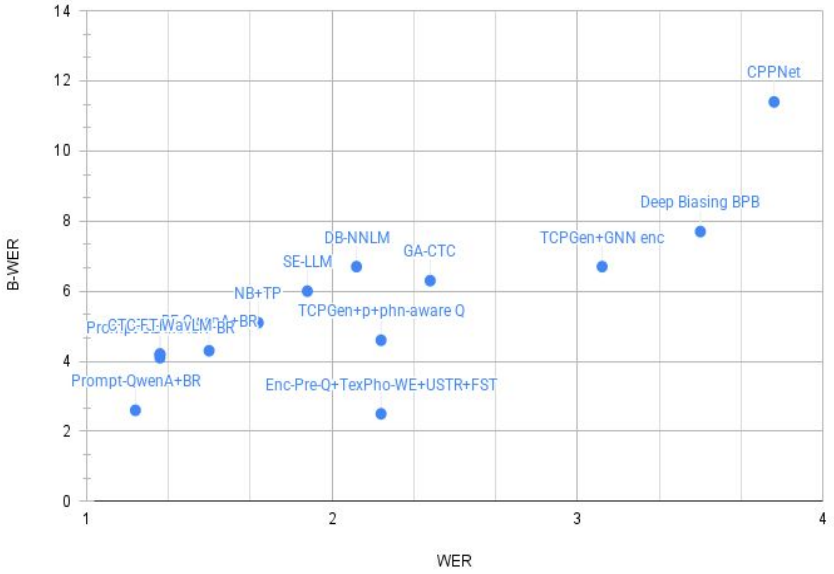
Decoder Biasing



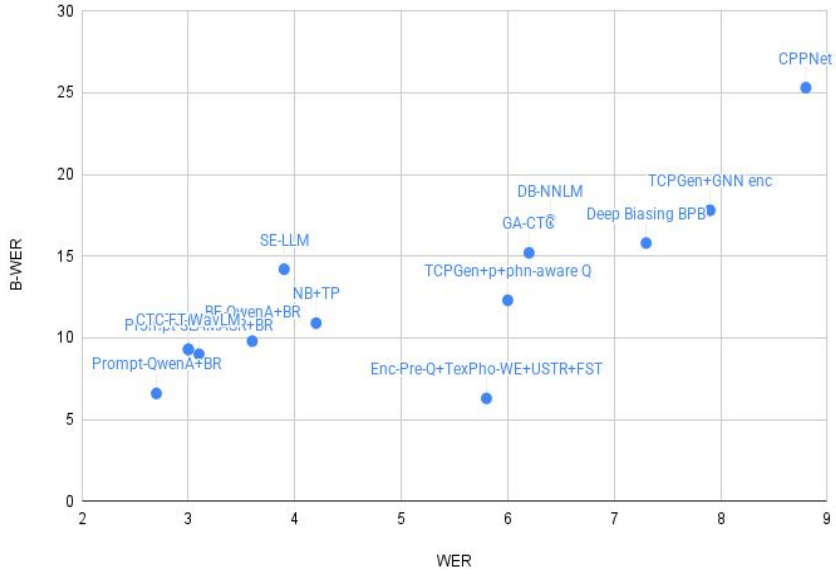
WER vs B-WER Benchmark on Hybrid Biasing (mostly)

B-WER: WER on bias (rare) words (N=1000)

LibriSpeech: test-clean



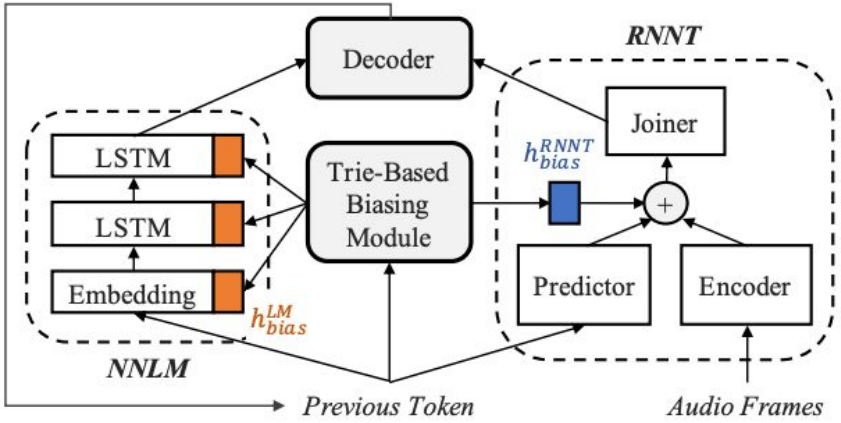
LibriSpeech: test-other



Hybrid Biasing

Leveraging advantages of both encoder and decoder biasing

- Shallow fusion
- Trie-based deep biasing
- Deep integration
  - Biasing module to encoder/LM

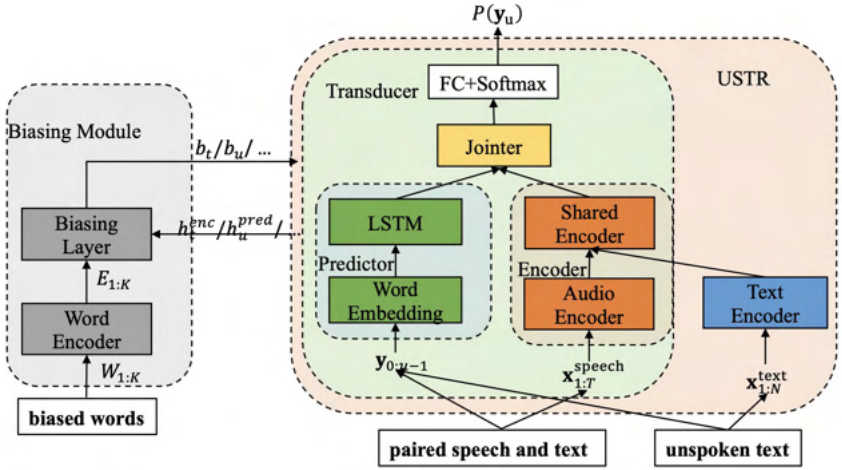


From D. Le, "Contextualized Streaming End-to-End Speech Recognition with Trie-Based Deep Biasing and Shallow Fusion", Interspeech 2021.

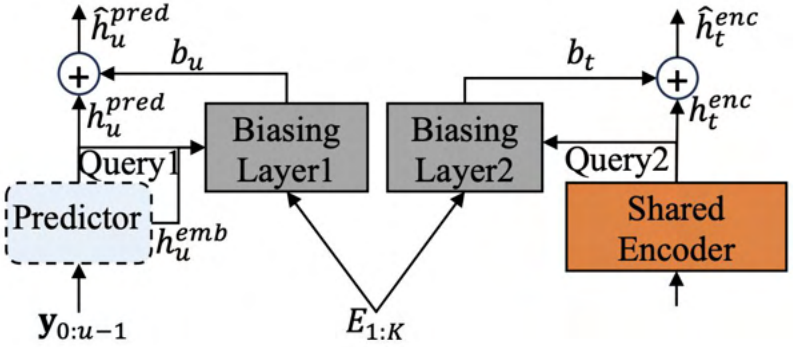
Hybrid Biasing

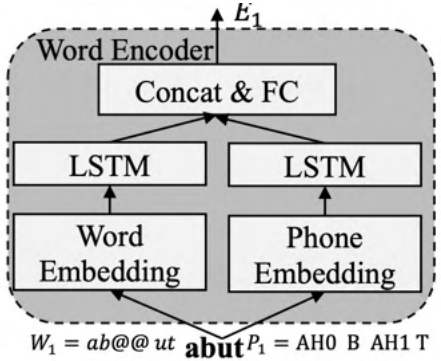
Leveraging advantages of both encoder and decoder biasing

- Text only domain adaptation
- Text and phoneme embedding
- Encoder-predictor biasing



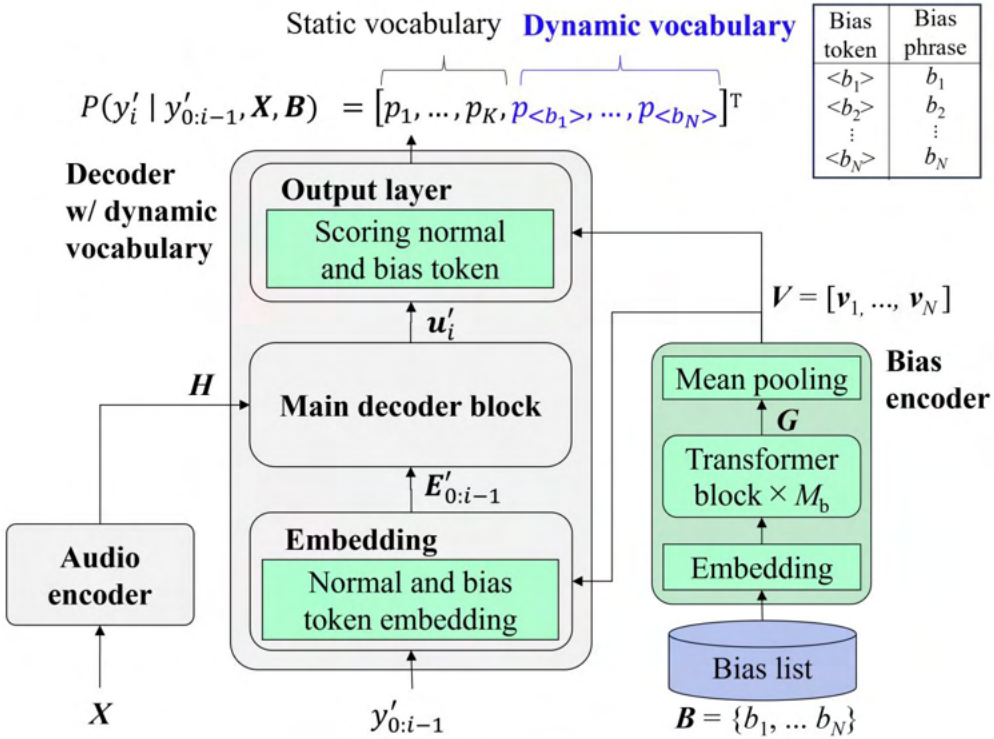
From J. Qiu, "Improving Large-Scale Deep Biasing with Phoneme Features and Text-Only Data in a Streaming Transducer", ASRU 2023.





Hybrid Biasing

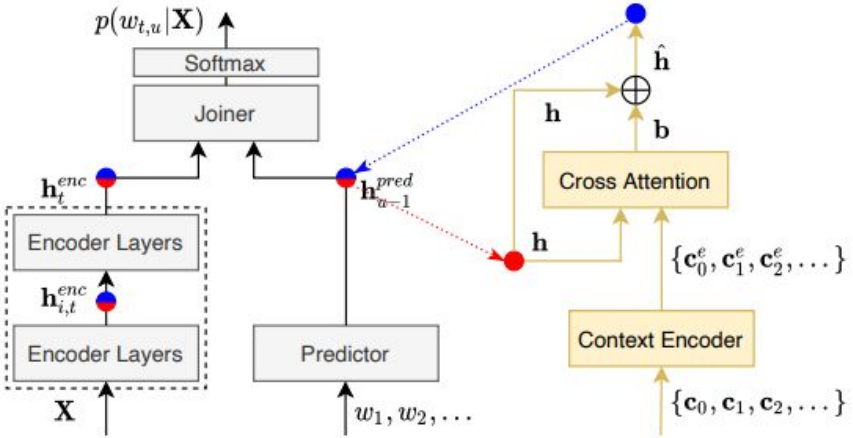
- Expanding output token layer



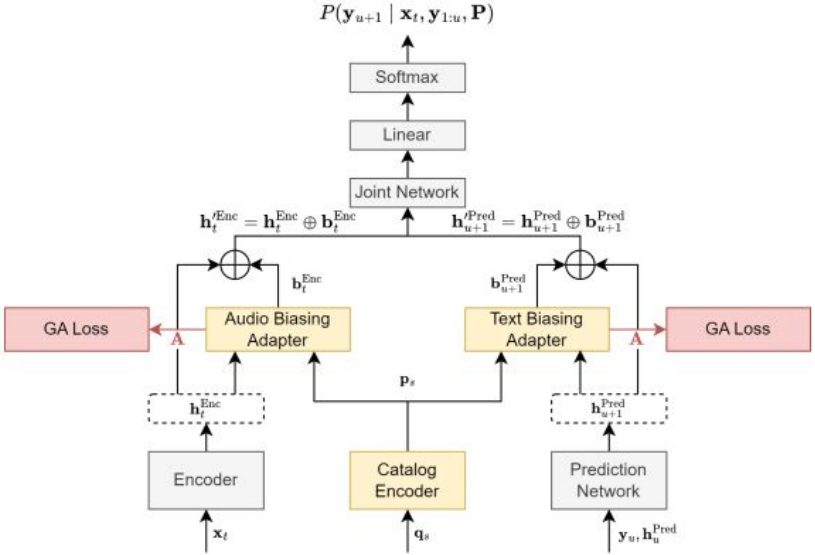
Hybrid Biasing

Applying diverse mechanisms to incorporate external knowledge for bias words to ASR models

- Guided attention on biasing adapters
- Neural biasing w/ attention



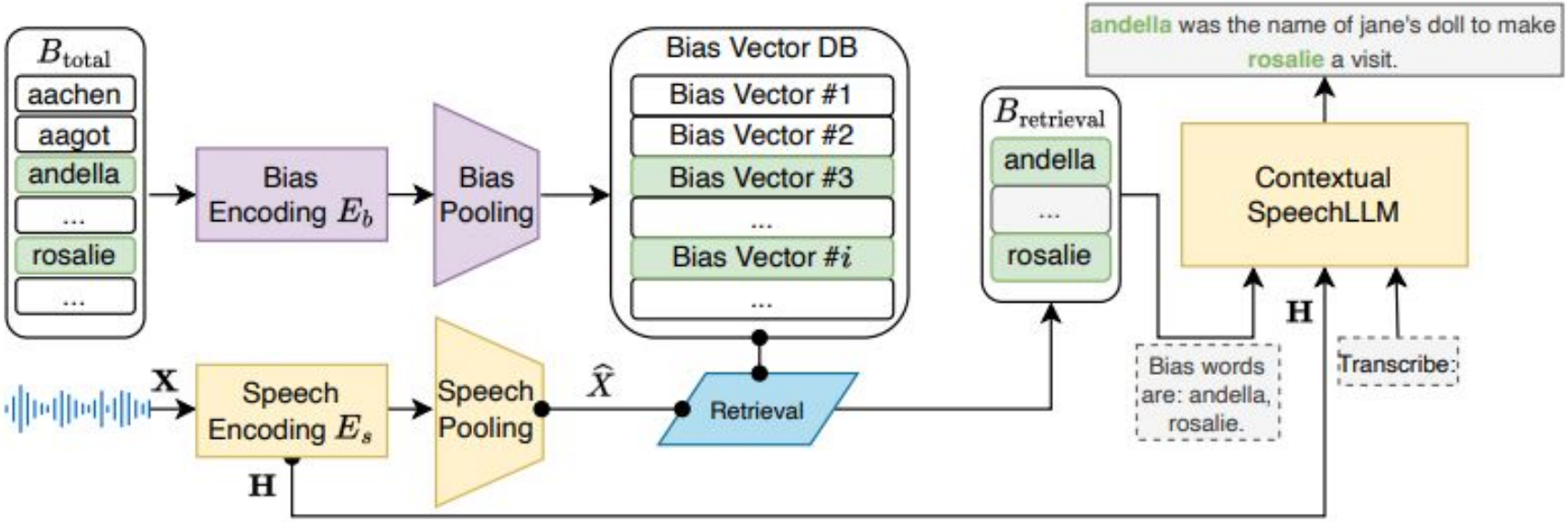
From R. Huang, "Improving Neural Biasing for Contextual Speech Recognition by Early Context Injection and Text Perturbation", Interspeech 2024.



From J. Tang, "Improving ASR Contextual biasing with Guided Attention", ICASSP 2024.

Contextual Biasing for LLMs

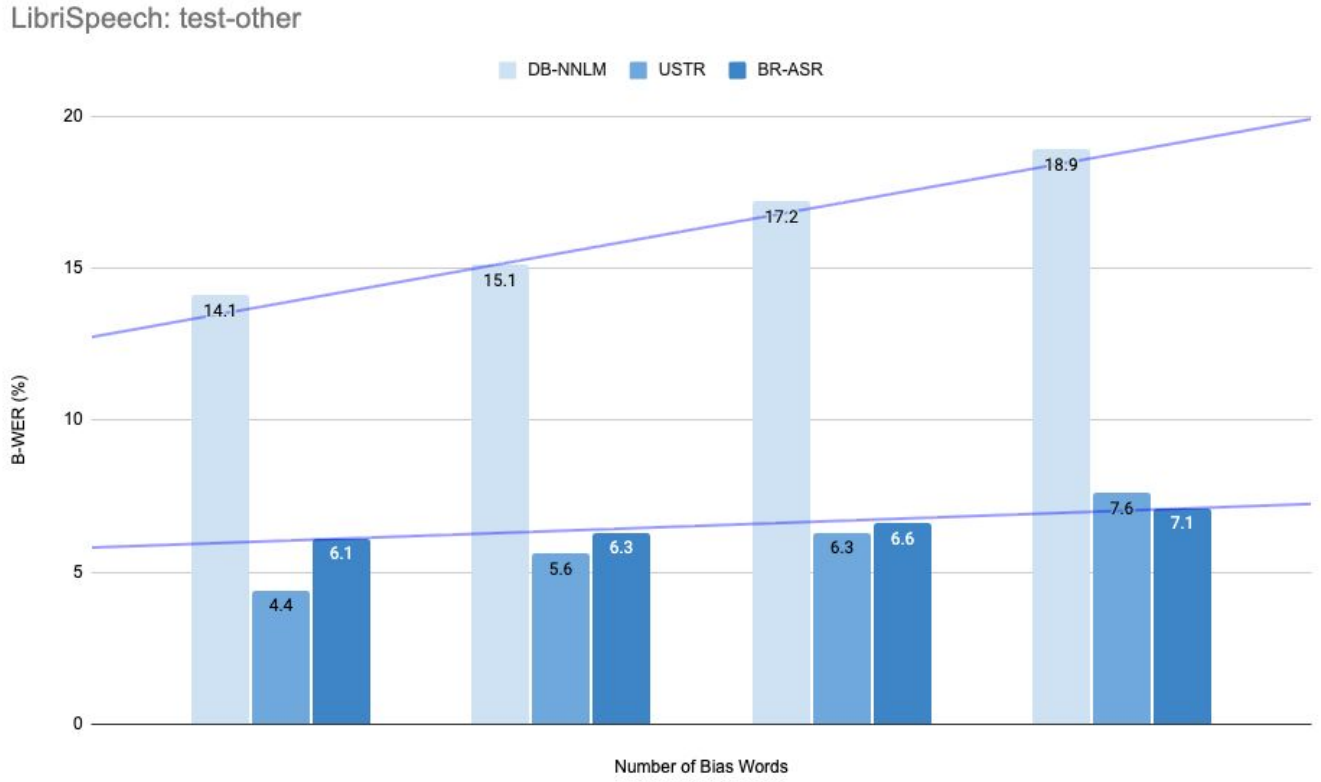
Prompting LLMs to enable contextual biasing



From X. Gong, "BR-ASR: Efficient and Scalable Bias Retrieval Framework for Contextual Biasing ASR in Speech LLM", **Interspeech 2025: Wed 13:30-15:30 Area12-Poster3**.

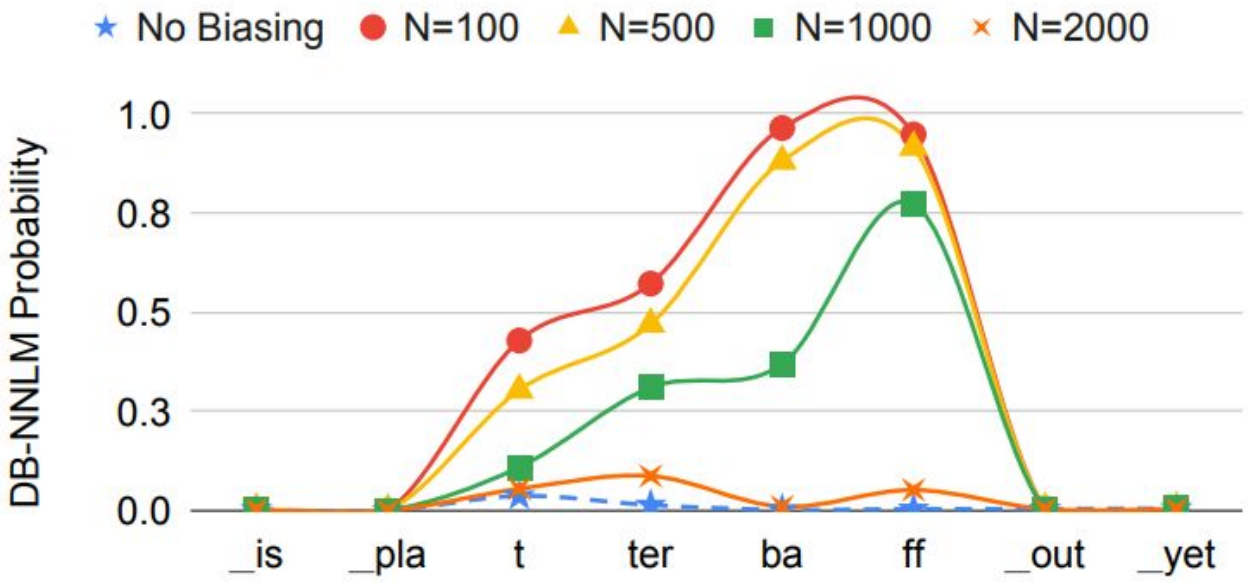
How Scalable is Contextual Biasing?

B-WER, increasing as a list of bias (rare) words gets larger



How Scalable is Contextual Biasing?

B-WER, increasing as a list of bias (rare) words gets larger and w/ more distractors



From D. Le, "Contextualized Streaming End-to-End Speech Recognition with Trie-Based Deep Biasing and Shallow Fusion", Interspeech 2021.

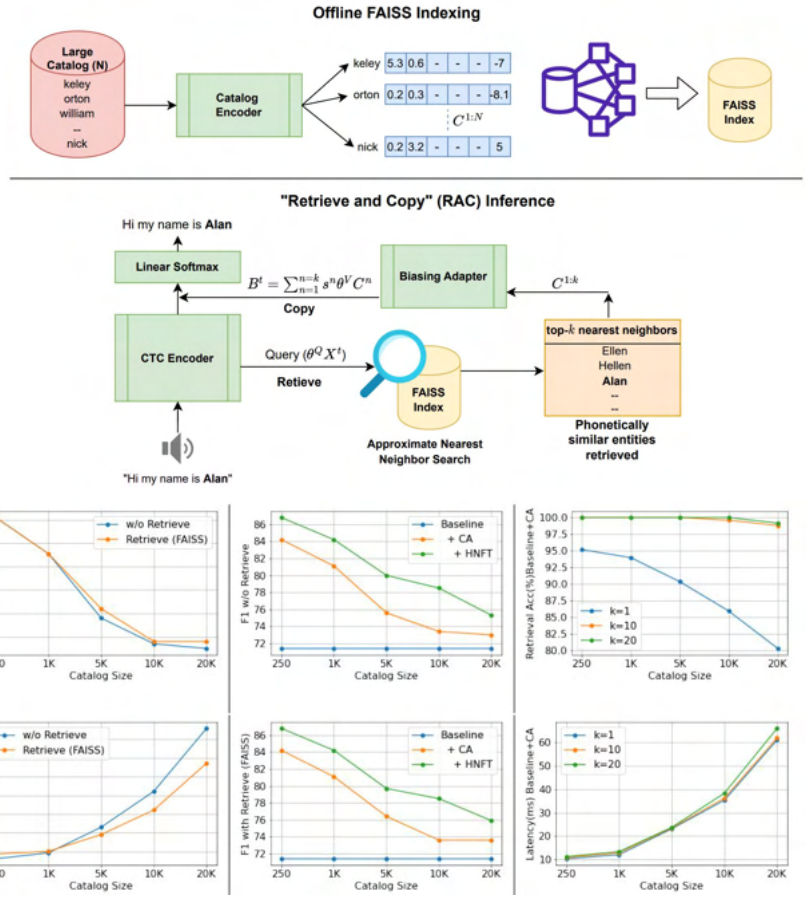
Retrieval Toward Large-Scale Contextual Biasing

Contextual biasing up to 200K

- Nearest search w/ FAISS
- Hard negative fine-tuning
- Encoder biasing w/ CTC

Model	Transcription
Baseline	my name is Ruben
+ CA	my name is Ruben
+ HNFT	my name is Rueben
Baseline	my name is Wally
+ CA	yes it's Wy
+ HNFT	yes it's Wally

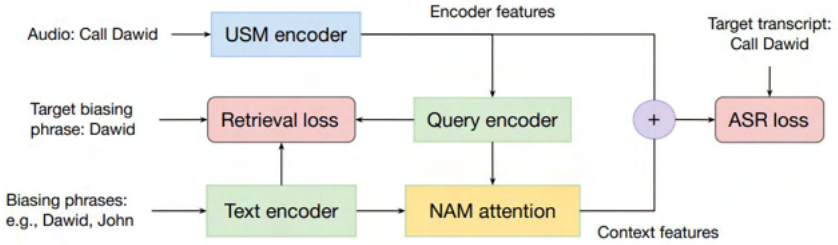
From S.Jayanthi, "Retrieve and Copy: Scaling ASR Personalization to Large Catalogs", EMNLP 2023.



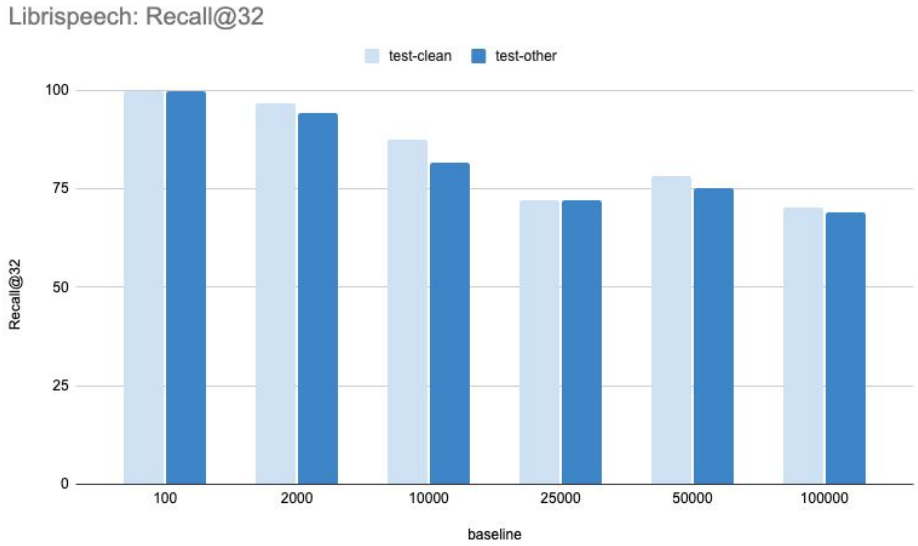
Retrieval Toward Large-Scale Contextual Biasing

Contextual biasing up to 100K

- Neural attention memory

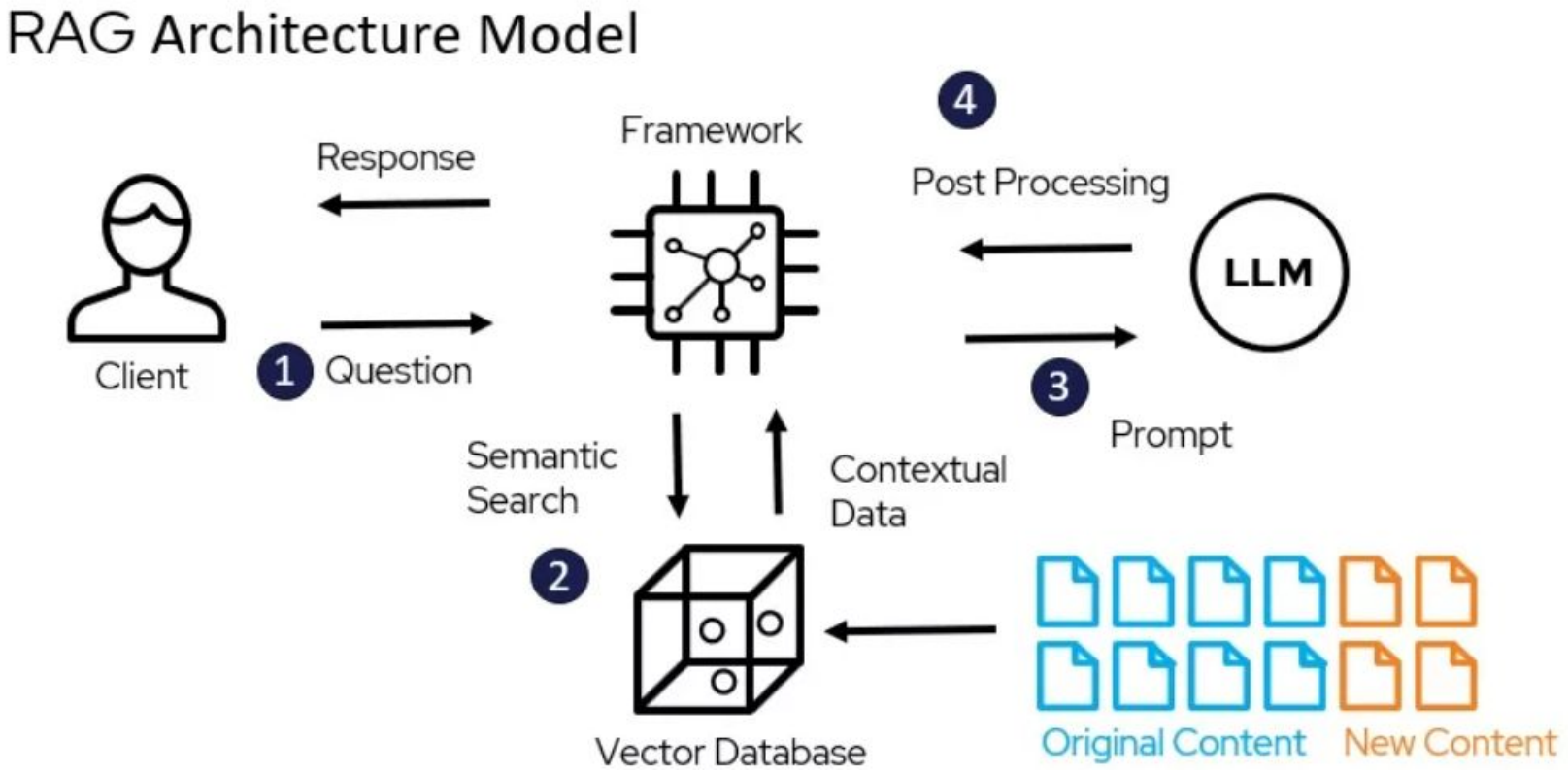


From Z. Huang, "Optimizing Large-Scale Context Retrieval for End-to-End ASR", Interspeech 2024.



Transcript truth	Top 5 phrases
nought worse than many others i **reckon**	nought, beckon, reckoning, reckoned, more's, er
margaret was **collecting** her mother's working materials and preparing to go to bed	collecting, milking, mature, marking, lurking
just as she was leaving the room she hesitated she was inclined to make an **acknowledgment** which she thought would please her father but which to be full and true must include a little **annoyance**	annoyance, acknowledgement, hesitating, incur, liberal allowance

What is Retrieval Augmented Generation (RAG)?



What is Retrieval Augmented Generation (RAG)?

Benefits of RAG

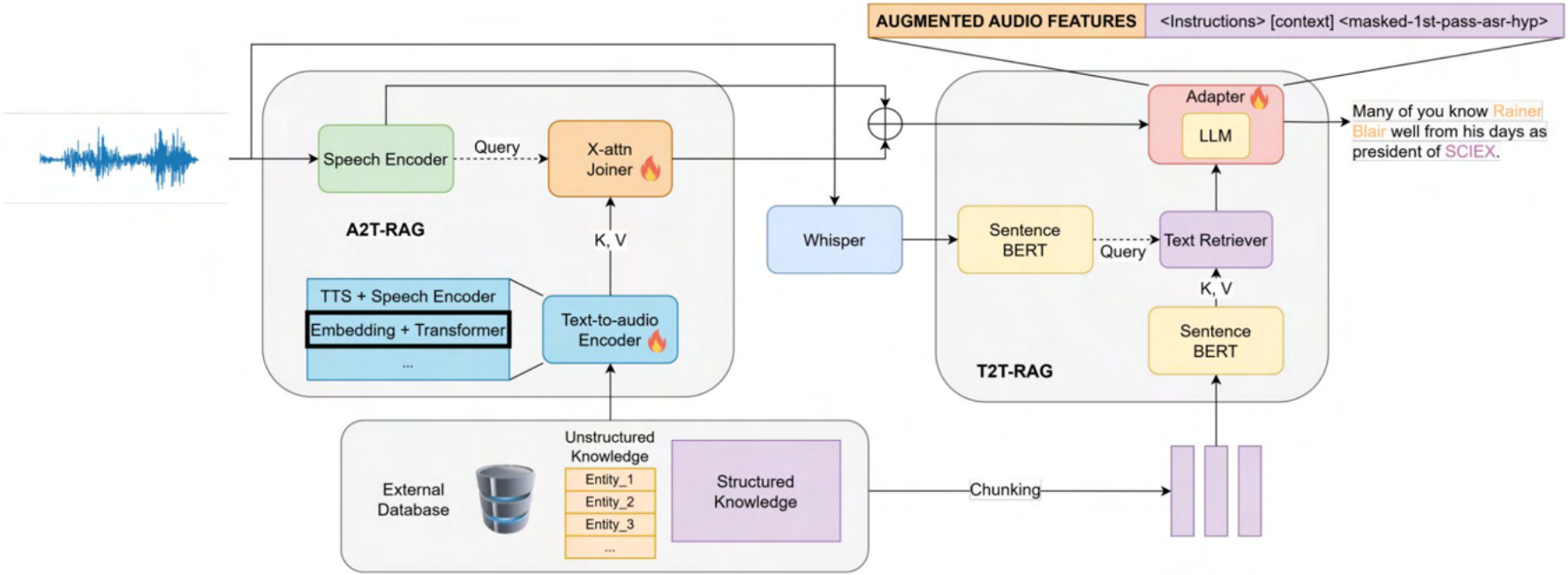
- Cost-effective implementation
- Current information
- Enhanced user trust
- More developer control

Benefits of RAG for contextual biasing for ASR

- Using innate contextual information in knowledge bases to provide better contextual prompts for LLMs
- Mitigating ASR output hallucinations

RAG-based Contextual Biasing for E2E ASR

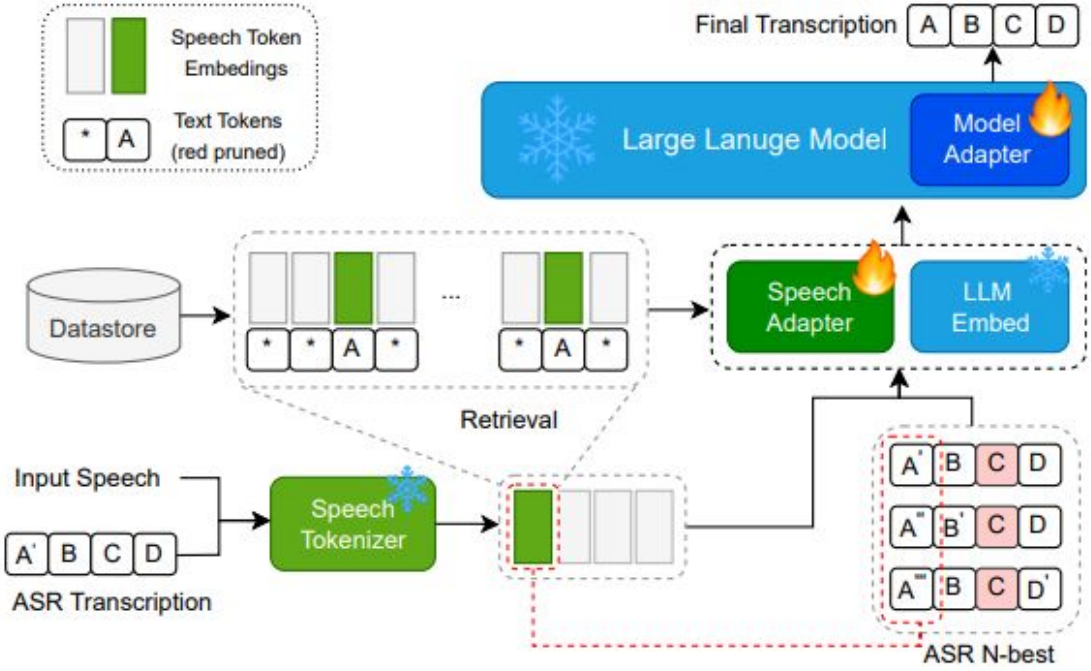
Text-based RAG + cross attention-based encoder biasing



From C. Xiao, "Contextual ASR with Retrieval Augmented Large Language Model", ICASSP 2025.

RAG-based Contextual Biasing for E2E ASR

RAG w/ multimodal embeddings for contextual biasing



From S. Li, "Enhancing LLM-based ASR Accuracy with Retrieval-Augmented Generation, https://arxiv.org/abs/2409.08597.

Takeaways: Contextual Biasing w/ Longer Context

- Contextual biasing
  - Encoder biasing can be a little conservative than decoder biasing in terms of biasing ASR outputs, but decoder biasing could lead to more false positives, resulting in substitution errors.
  - Hybrid biasing has been popular, but is still challenged by large-scale biasing, e.g., 100k bias words.
- Large-scale biasing w/ retrieval
  - FAISS or something similar to enable efficient retrieval of contextual information from large-scale knowledge source has been recently explored with decent results.
- Retrieval augmented generation
  - Can be another biasing solution when it comes to LLMs being more utilized for ASR tasks.

Table of Contents

Download Slides

Introduction (10 mins) 15:30-15:40	
Shinji (30 min) 15:40-16:10	Speech-to-Text Benchmark
Taejin (40 min) 16:10-16:50	Leveraging Long Acoustic Context
Recess (10 min) 16:50-17:00	
Huck (40 min) 17:00-17:40	Semantic Context and Speech-Language Modeling
Kyu (30 min) 17:40-18:10	Contextual Biasing and Methods Leveraging Longer Semantic Context for Speech Systems
Closing Remark (10 min) 18:10-18:20	
Q&A Session (10 min) 18:20-18:30	

Closing

Kyu J. Han

Long-Context Acoustic and Linguistic Insights

Acoustic Insights

- How **intrinsic variability** (Physiological traits, speaker characteristics) and **extrinsic variability** (mic, codecs) construct acoustic context for ASR
- How streaming ASR application and speaker attributed ASR takes advantage of acoustic context
- **Alternative architectures** for speech recognition
- Speech encoders and **ASR endpointing** in voice agents applications.

Linguistic Insights

- **Semantic Information** Modeling in End-to-End ASR: Classical ASR-LM, Post-processing ASR corrections
- Preference-based Semantic Modeling in Post-training
- **Semantic Understanding** from ASR to Audio Contexts
- Limitation and New Evaluation of Semantic Modeling: Data Leakage via Text, Pre-training Agentic & Instruction Evaluation

Context Biasing

- **Contextual Biasing** for E2E ASR Decoder, Encoder, Hybrid Biasing
- Retrieval Toward Large-Scale Biasing
- **Retrieval Augmented Generation (RAG)** and how RAG is used for contextual Biasing

Summary

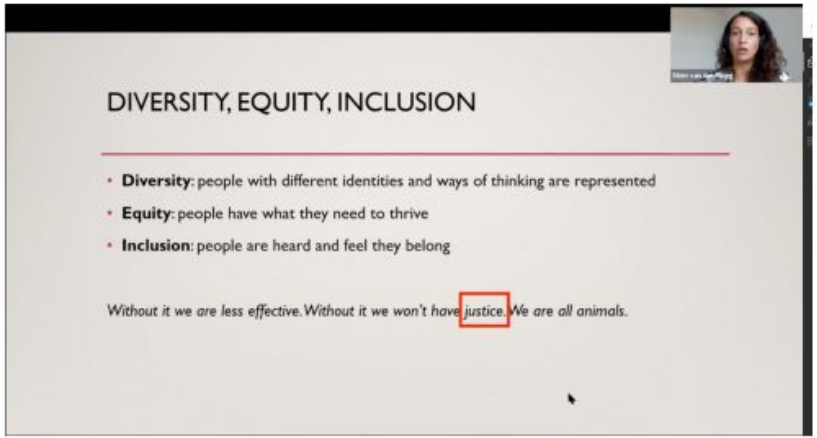
Integrating long-context acoustic and linguistic insights

- Evaluation metrics
- Datasets
- Modeling frameworks leveraging acoustic and linguistic insights
- Trends in recent LLMs with long contextual information
- Semantic contextual biasing
- RAG-based contextual bising for LLMs on ASR

Future Directions

- New tasks/datasets
  - Audio question answering, joint ASR/audio captioning, etc.
- More acoustic and semantic insights
  - Environmental, personal, regional, etc.
- Error correction and optimization with downstream tasks
  - For example, WER improvement vs. more useful clinical note generation, doe the former guarantee the latter all the time?
- RAG vs. long-context LLMs
  - Whether to use RAG or use long-context LLMs as is.
- Multimodal biasing for multimodal LLMs
  - How to incorporate multimodal context biasing.

Future Directions

Slide		
GT	okay so let me introduce our 1st speaker who is stien van der ploeg	without it we are less effective without it we will not have justice we are all animals
AV	okay so let me introduce our 1st speaker who is stein van der ploeg	without it we are less effective without it we will not have justice we are all animals
A	okay so let me introduce our 1st speaker who is steam funder plu	without it we are less effective without it we will not have guests we are all animals

Thank you for your Attention!

Questions And Answers




Backup Appendix

Appendix: Privacy and LLM Data Leakage

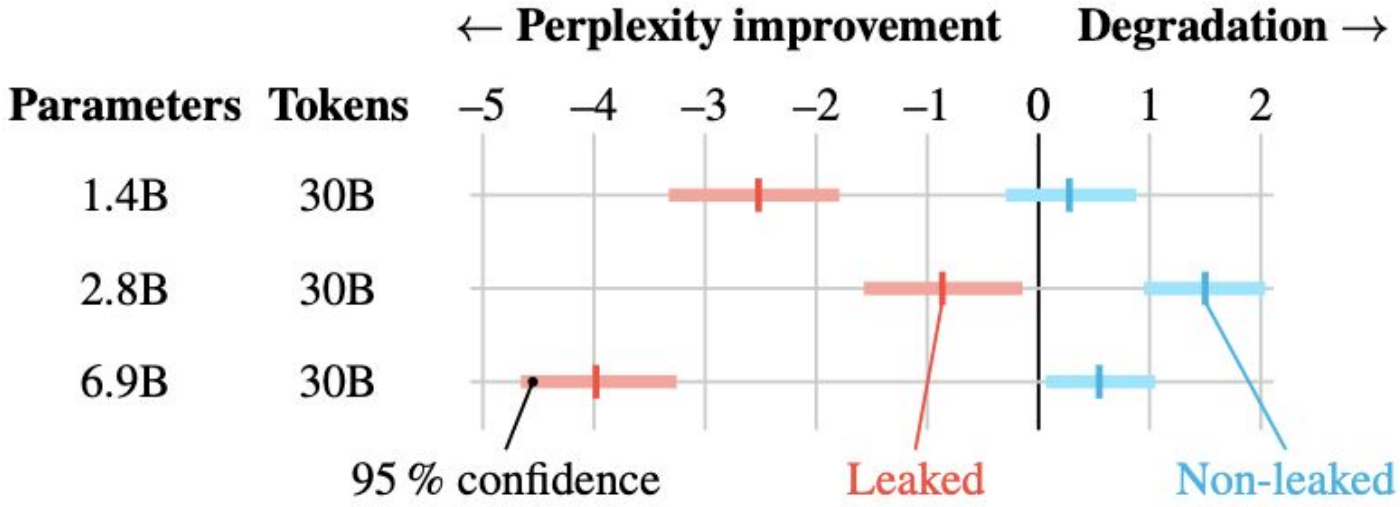
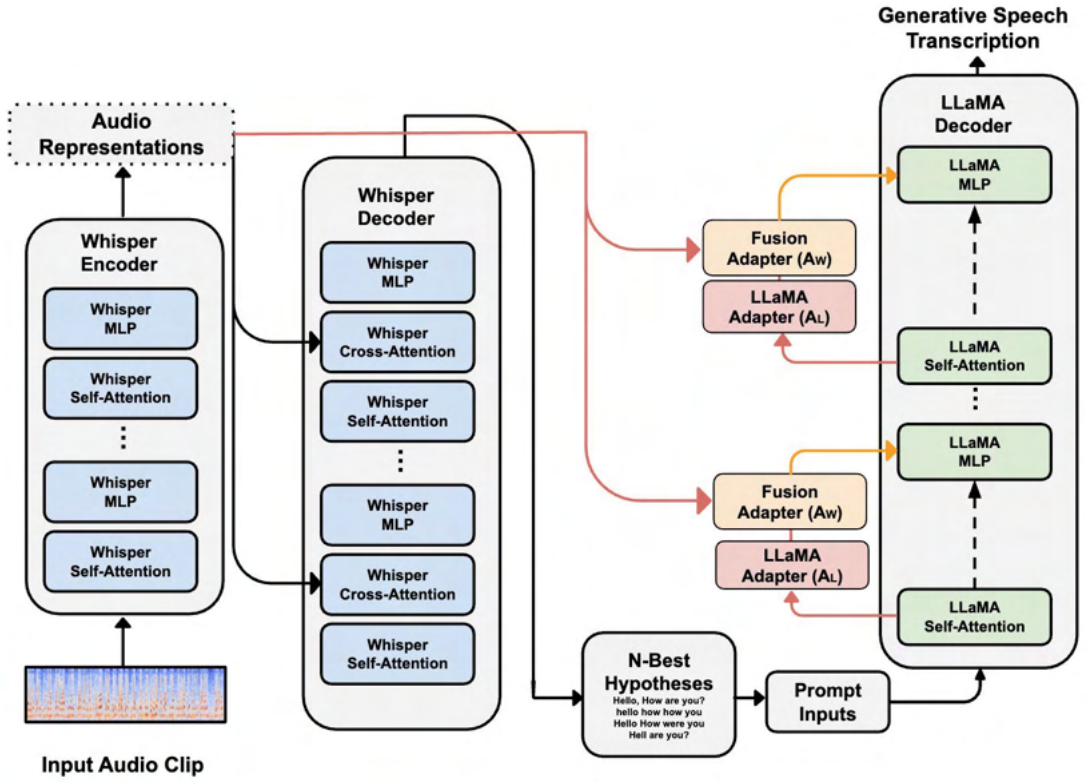
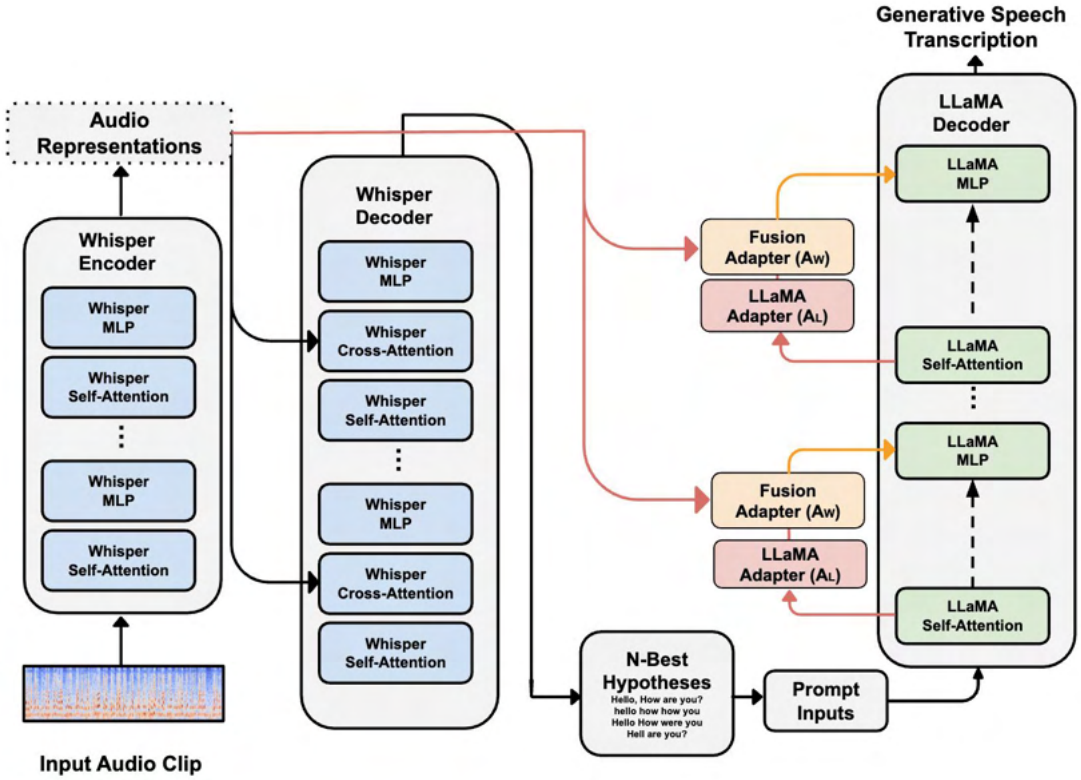


Image Source: Evaluation of LLMs in Speech is Often Flawed: Test Set Contamination in Large Language Models for Speech Recognition, Yuan Tseng et al. 2025

Appendix: Whispering-LLaMA on the attention merging



Pipeline


Generative Speech Transcription
Audio Representations
Whisper Encoder
Whisper MLP
Whisper Self-Attention
Whisper MLP
Whisper Self-Attention
Input Audio Clip
Whisper Decoder
Whisper MLP
Whisper Cross-Attention
Whisper Self-Attention
Whisper MLP
Whisper Cross-Attention
Whisper Self-Attention
Fusion Adapter (Aw)
LLaMA Adapter (AL)
Fusion Adapter (Aw)
LLaMA Adapter (AL)
LLaMA Decoder
LLaMA MLP
LLaMA Self-Attention
LLaMA MLP
LLaMA Self-Attention
N-Best Hypotheses
Hello, How are you?
hello how how you
Hello How were you
Hell are you?
Prompt Inputs