Low-latency conversational agent

Tatiana Likhomanenko*

Luke Carlson*, Richard Bai*, Zijin Gu*, Zak Aldeneh*, Yizhe Zhang*, Shiladitya Dutta*, Han Tran*, Ruixiang Zhang, Huangjie Zheng, Navdeep Jaitly*

Conversational AI Reading Meeting | Apple | 11 Dec 2025

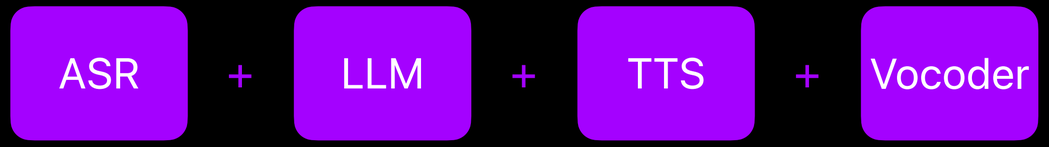
Conversational AI agent design

Is an open problem !

Conversational AI agent design

Is an open problem !

Cascaded



	Low latency	No error propagation	Interpretability	Language understanding	Emergent capabilities
Cascaded					

Conversational AI agent design

Is an open problem !

Cascaded ASR + LLM + TTS + Vocoder

	Low latency	No error propagation	Interpretability	Language understanding	Emergent capabilities
Cascaded					

Conversational AI agent design

Is an open problem !

Cascaded ASR + LLM + TTS + Vocoder

	Low latency	No error propagation	Interpretability	Language understanding	Emergent capabilities
Cascaded					

Conversational AI agent design

Is an open problem !

Cascaded ASR + LLM + TTS + Vocoder

	Low latency	No error propagation	Interpretability	Language understanding	Emergent capabilities
Cascaded		✗	👍	👍	

Conversational AI agent design

Is an open problem !

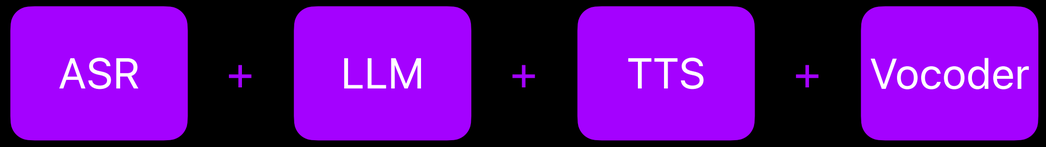
Cascaded ASR + LLM + TTS + Vocoder

	Low latency	No error propagation	Interpretability	Language understanding	Emergent capabilities
Cascaded					

Conversational AI agent design

Is an open problem !

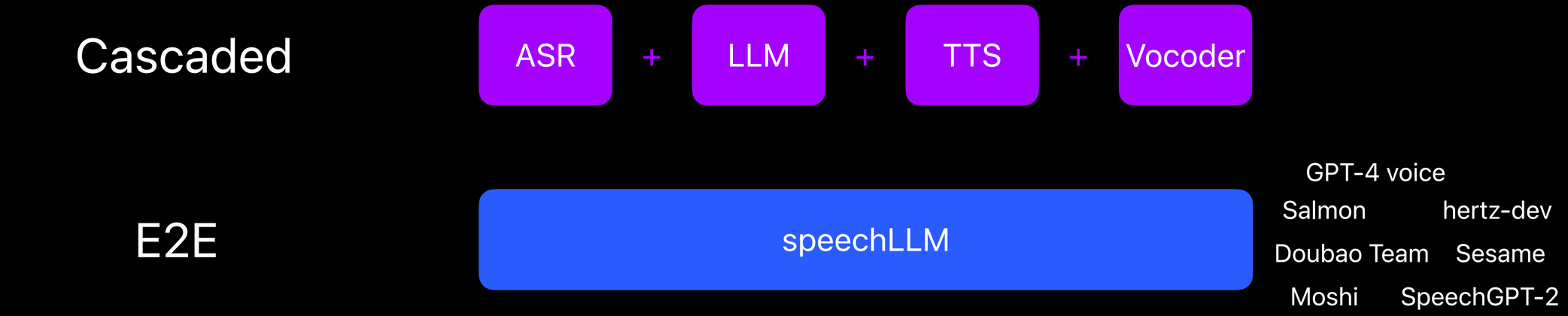
Cascaded



	Low latency	No error propagation	Interpretability	Language understanding	Emergent capabilities
Cascaded	✗	✗	👍	👍	✗

Conversational AI agent design

Trending : E2E conversational AI



	Low latency	No error propagation	Interpretability	Language understanding	Emergent capabilities
Cascaded	✗	✗	👍	👍	✗
E2E					

Conversational AI agent design

Recent publications reveal signs that cascaded systems outperform E2E for many tasks

	Low latency	No error propagation	Interpretability	Language understanding	Emergent capabilities
Cascaded	✗	✗	👍	👍	✗
E2E					

Nguyen, T.A., Muller, B., Yu, B., Costa-jussa, M.R., Elbayad, M., Popuri, S., Duquenne, P.A., Algayres, R., Mavlyutov, R., Gat, I. and Synnaeve, G., Spirit-lm: Interleaved spoken and written language model, https://arxiv. org/abs/2402.05755, 2024

Y. Chen, X. Yue, C. Zhang, X. Gao, R. T. Tan, and H. Li, "Voicebench: Benchmarking llm-based voice assistants," arXiv preprint arXiv:2410.17196, 2024

J. Xu, Z. Guo, J. He, H. Hu, T. He, S. Bai, K. Chen, J. Wang, Y. Fan, K. Dang et al., "Qwen2. 5-omni technical report," arXiv preprint arXiv:2503.20215, 2025

S. Sakshi, U. Tyagi, S. Kumar, A. Seth, R. Selvakumar, O. Nieto, R. Duraiswami, S. Ghosh, and D. Manocha, "Mmau: A massive multi-task audio understanding and reasoning benchmark," in ICLR, 2025

G Heigold, E Variani, T Bagby, C Allauzen, J Ma, S Kumar, M Riley "Massive Sound Embedding Benchmark (MSEB)", in NeurIPS 2025

Chien-yu Huang et.al. Dynamic-superb phase-2: A collaboratively expanding benchmark for measuring the capabilities of spoken language models with 180 tasks, in ICLR, 2025

Conversational AI agent design

Recent publications reveal signs that cascaded systems outperform E2E for many tasks

	Low latency	No error propagation	Interpretability	Language understanding	Emergent capabilities
Cascaded	✗	✗	👍	👍	✗
E2E				✗	

Nguyen, T.A., Muller, B., Yu, B., Costa-jussa, M.R., Elbayad, M., Popuri, S., Duquenne, P.A., Algayres, R., Mavlyutov, R., Gat, I. and Synnaeve, G., Spirit-lm: Interleaved spoken and written language model, https://arxiv. org/abs/2402.05755, 2024

Y. Chen, X. Yue, C. Zhang, X. Gao, R. T. Tan, and H. Li, "Voicebench: Benchmarking llm-based voice assistants," arXiv preprint arXiv:2410.17196, 2024

J. Xu, Z. Guo, J. He, H. Hu, T. He, S. Bai, K. Chen, J. Wang, Y. Fan, K. Dang et al., "Qwen2. 5-omni technical report," arXiv preprint arXiv:2503.20215, 2025

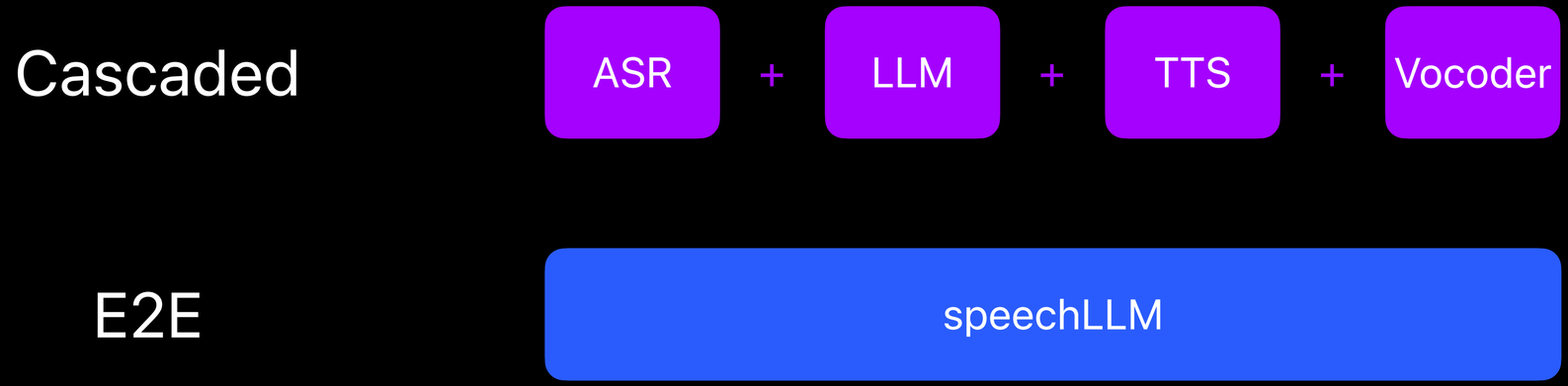
S. Sakshi, U. Tyagi, S. Kumar, A. Seth, R. Selvakumar, O. Nieto, R. Duraiswami, S. Ghosh, and D. Manocha, "Mmau: A massive multi-task audio understanding and reasoning benchmark," in ICLR, 2025

G Heigold, E Variani, T Bagby, C Allauzen, J Ma, S Kumar, M Riley "Massive Sound Embedding Benchmark (MSEB)", in NeurIPS 2025

Chien-yu Huang et.al. Dynamic-superb phase-2: A collaboratively expanding benchmark for measuring the capabilities of spoken language models with 180 tasks, in ICLR, 2025

Conversational AI agent design

Recent publications reveal signs that cascaded systems outperform E2E for many tasks



	Low latency	No error propagation	Interpretability	Language understanding	Emergent capabilities
Cascaded	✗	✗	👍	👍	✗
E2E			✗	✗	

Nguyen, T.A., Muller, B., Yu, B., Costa-jussa, M.R., Elbayad, M., Popuri, S., Duquenne, P.A., Algayres, R., Mavlyutov, R., Gat, I. and Synnaeve, G., Spirit-lm: Interleaved spoken and written language model, https://arxiv. org/abs/2402.05755, 2024

Y. Chen, X. Yue, C. Zhang, X. Gao, R. T. Tan, and H. Li, "Voicebench: Benchmarking llm-based voice assistants," arXiv preprint arXiv:2410.17196, 2024

J. Xu, Z. Guo, J. He, H. Hu, T. He, S. Bai, K. Chen, J. Wang, Y. Fan, K. Dang et al., "Qwen2. 5-omni technical report," arXiv preprint arXiv:2503.20215, 2025

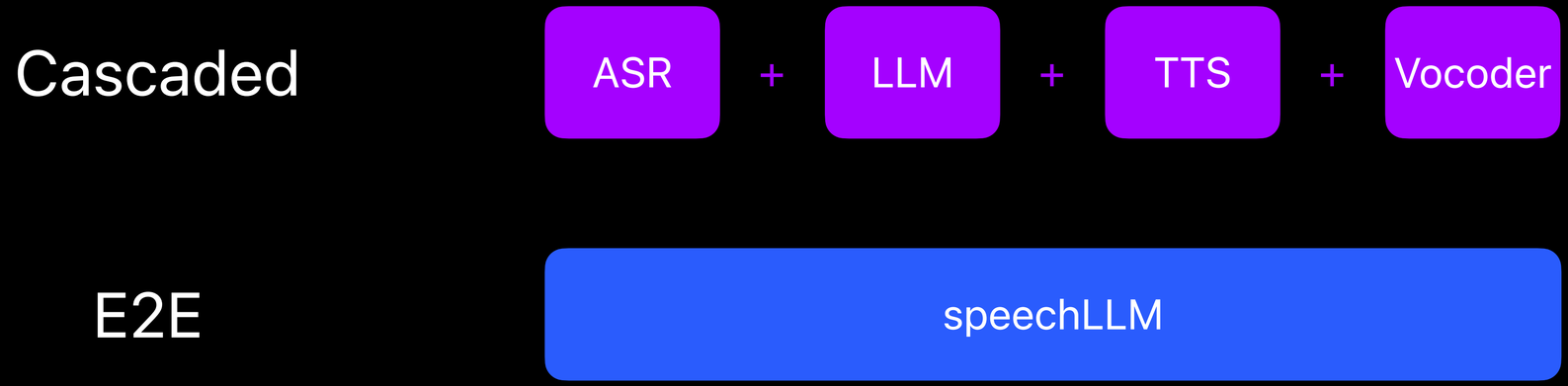
S. Sakshi, U. Tyagi, S. Kumar, A. Seth, R. Selvakumar, O. Nieto, R. Duraiswami, S. Ghosh, and D. Manocha, "Mmau: A massive multi-task audio understanding and reasoning benchmark," in ICLR, 2025

G Heigold, E Variani, T Bagby, C Allauzen, J Ma, S Kumar, M Riley "Massive Sound Embedding Benchmark (MSEB)", in NeurIPS 2025

Chien-yu Huang et.al. Dynamic-superb phase-2: A collaboratively expanding benchmark for measuring the capabilities of spoken language models with 180 tasks, in ICLR, 2025

Conversational AI agent design

Recent publications reveal signs that cascaded systems outperform E2E for many tasks

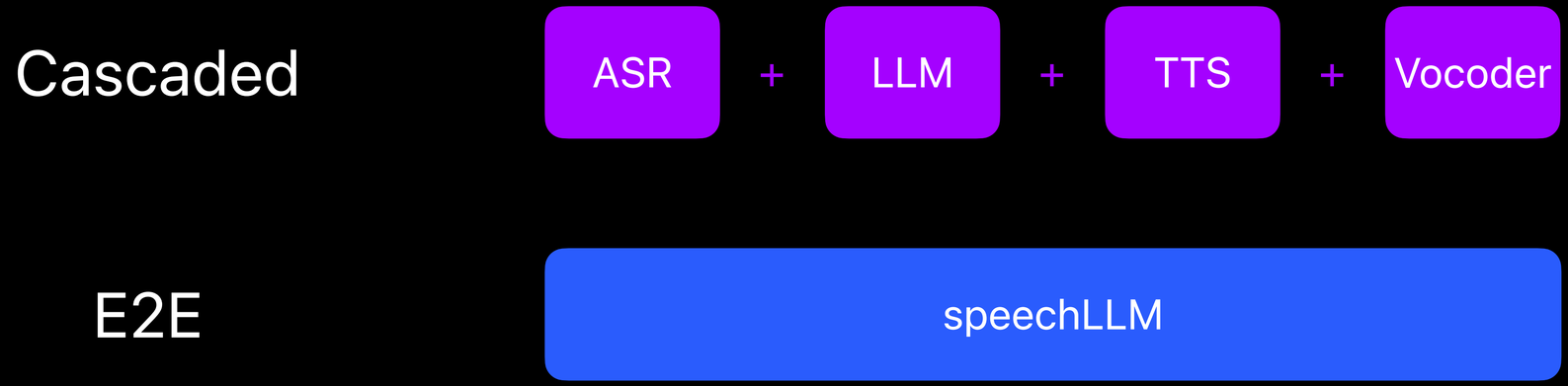


	Low latency	No error propagation	Interpretability	Language understanding	Emergent capabilities
Cascaded	✗	✗	👍	👍	✗
E2E	?	?	✗	✗	

Nguyen, T.A., Muller, B., Yu, B., Costa-jussa, M.R., Elbayad, M., Popuri, S., Duquenne, P.A., Algayres, R., Mavlyutov, R., Gat, I. and Synnaeve, G., Spirit-lm: Interleaved spoken and written language model, https://arxiv. org/abs/2402.05755, 2024
Y. Chen, X. Yue, C. Zhang, X. Gao, R. T. Tan, and H. Li, "Voicebench: Benchmarking llm-based voice assistants," arXiv preprint arXiv:2410.17196, 2024
J. Xu, Z. Guo, J. He, H. Hu, T. He, S. Bai, K. Chen, J. Wang, Y. Fan, K. Dang et al., "Qwen2. 5-omni technical report," arXiv preprint arXiv:2503.20215, 2025
S. Sakshi, U. Tyagi, S. Kumar, A. Seth, R. Selvakumar, O. Nieto, R. Duraiswami, S. Ghosh, and D. Manocha, "Mmau: A massive multi-task audio understanding and reasoning benchmark," in ICLR, 2025
G Heigold, E Variani, T Bagby, C Allauzen, J Ma, S Kumar, M Riley "Massive Sound Embedding Benchmark (MSEB)", in NeurIPS 2025
Chien-yu Huang et.al. Dynamic-superb phase-2: A collaboratively expanding benchmark for measuring the capabilities of spoken language models with 180 tasks, in ICLR, 2025

Conversational AI agent design

Recent publications reveal signs that cascaded systems outperform E2E for many tasks



	Low latency	No error propagation	Interpretability	Language understanding	Emergent capabilities
Cascaded	✗	✗	👍	👍	✗
E2E	?	?	✗	✗	👍

Nguyen, T.A., Muller, B., Yu, B., Costa-jussa, M.R., Elbayad, M., Popuri, S., Duquenne, P.A., Algayres, R., Mavlyutov, R., Gat, I. and Synnaeve, G., Spirit-lm: Interleaved spoken and written language model, https://arxiv. org/abs/2402.05755, 2024

Y. Chen, X. Yue, C. Zhang, X. Gao, R. T. Tan, and H. Li, "Voicebench: Benchmarking llm-based voice assistants," arXiv preprint arXiv:2410.17196, 2024

J. Xu, Z. Guo, J. He, H. Hu, T. He, S. Bai, K. Chen, J. Wang, Y. Fan, K. Dang et al., "Qwen2. 5-omni technical report," arXiv preprint arXiv:2503.20215, 2025

S. Sakshi, U. Tyagi, S. Kumar, A. Seth, R. Selvakumar, O. Nieto, R. Duraiswami, S. Ghosh, and D. Manocha, "Mmau: A massive multi-task audio understanding and reasoning benchmark," in ICLR, 2025

G Heigold, E Variani, T Bagby, C Allauzen, J Ma, S Kumar, M Riley "Massive Sound Embedding Benchmark (MSEB)", in NeurIPS 2025

Chien-yu Huang et.al. Dynamic-superb phase-2: A collaboratively expanding benchmark for measuring the capabilities of spoken language models with 180 tasks, in ICLR, 2025

Conversational AI agent design

Recent publications reveal signs that cascaded systems outperform E2E for many tasks

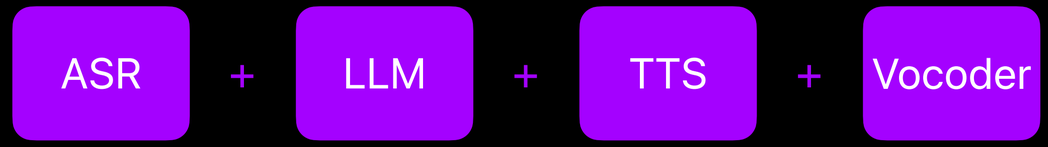
	Low latency	No error propagation	Interpretability	Language understanding	Emergent capabilities
Cascaded	✗	✗	👍	👍	✗
E2E	?	?	✗	✗	👍

Nguyen, T.A., Muller, B., Yu, B., Costa-jussa, M.R., Elbayad, M., Popuri, S., Duquenne, P.A., Algayres, R., Mavlyutov, R., Gat, I. and Synnaeve, G., Spirit-lm: Interleaved spoken and written language model, https://arxiv. org/abs/2402.05755, 2024
Y. Chen, X. Yue, C. Zhang, X. Gao, R. T. Tan, and H. Li, "Voicebench: Benchmarking llm-based voice assistants," arXiv preprint arXiv:2410.17196, 2024
J. Xu, Z. Guo, J. He, H. Hu, T. He, S. Bai, K. Chen, J. Wang, Y. Fan, K. Dang et al., "Qwen2. 5-omni technical report," arXiv preprint arXiv:2503.20215, 2025
S. Sakshi, U. Tyagi, S. Kumar, A. Seth, R. Selvakumar, O. Nieto, R. Duraiswami, S. Ghosh, and D. Manocha, "Mmau: A massive multi-task audio understanding and reasoning benchmark," in ICLR, 2025
G Heigold, E Variani, T Bagby, C Allauzen, J Ma, S Kumar, M Riley "Massive Sound Embedding Benchmark (MSEB)", in NeurIPS 2025
Chien-yu Huang et.al. Dynamic-superb phase-2: A collaboratively expanding benchmark for measuring the capabilities of spoken language models with 180 tasks, in ICLR, 2025

Conversational AI agent : ChipChat

Build a strong cascaded model with low latency !

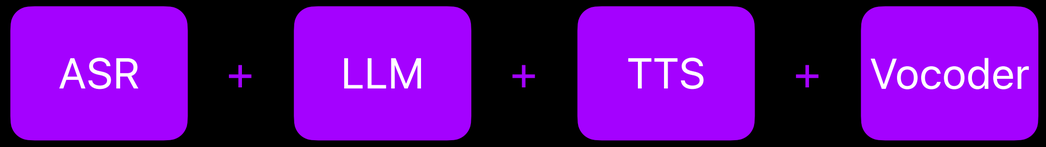
Cascaded



Conversational AI agent : ChipChat

Build a strong cascaded model with low latency !

Cascaded

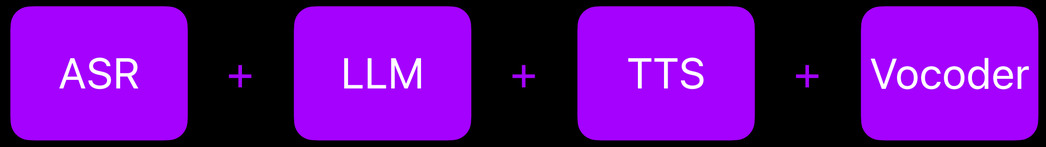


Low latency

Conversational AI agent : ChipChat

Build a strong cascaded model with low latency !

Cascaded



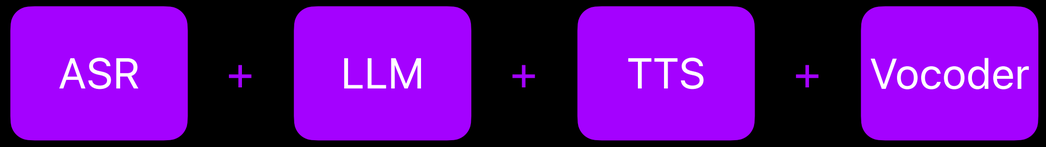
Low latency

User-centric

Conversational AI agent : ChipChat

Build a strong cascaded model with low latency !

Cascaded



Low latency

User-centric

Interruption capability

Conversational AI agent : ChipChat

Build a strong cascaded model with low latency !

Cascaded

ASR + LLM + TTS + Vocoder

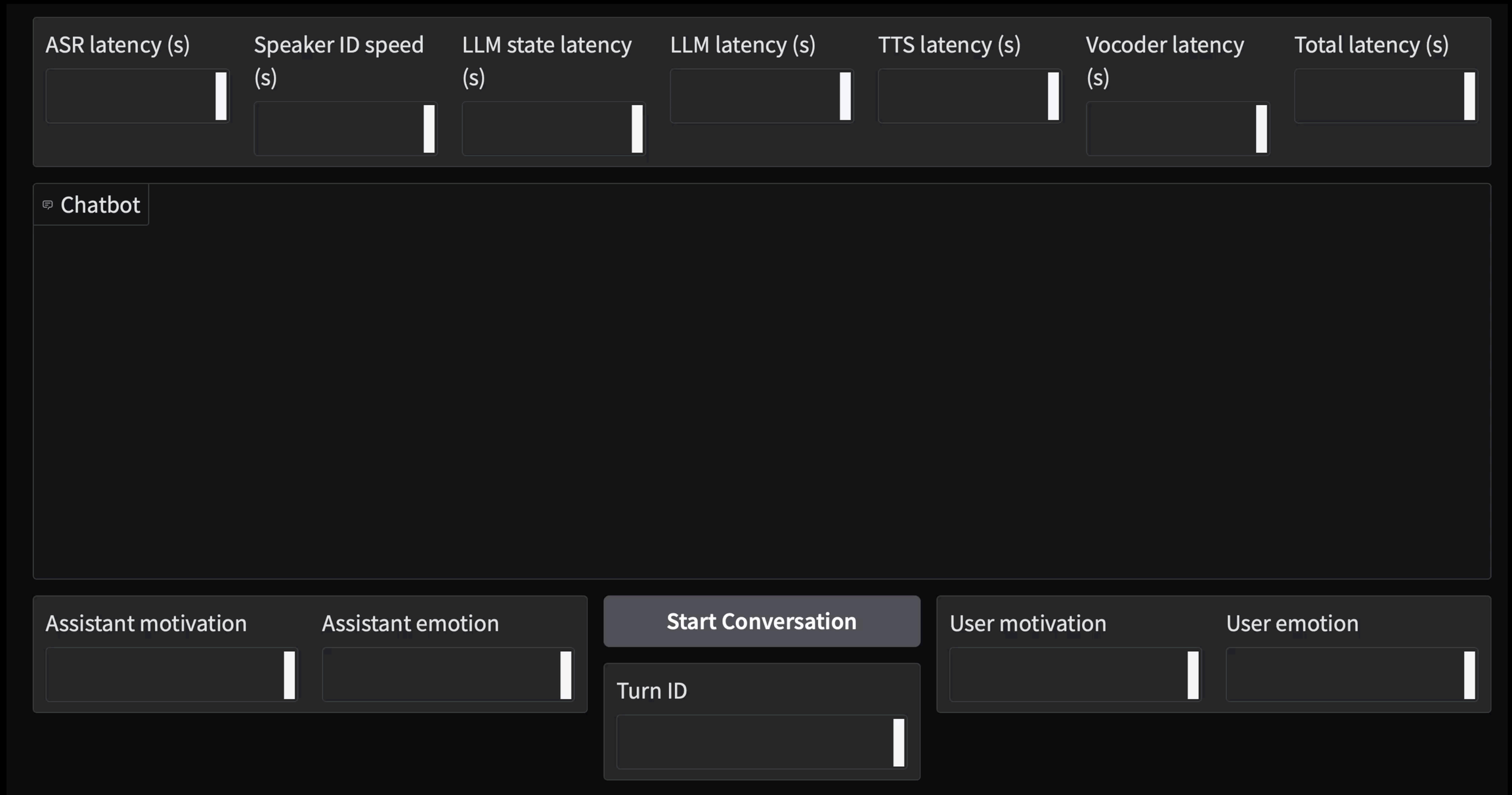
Low latency

User-centric

Interruption capability

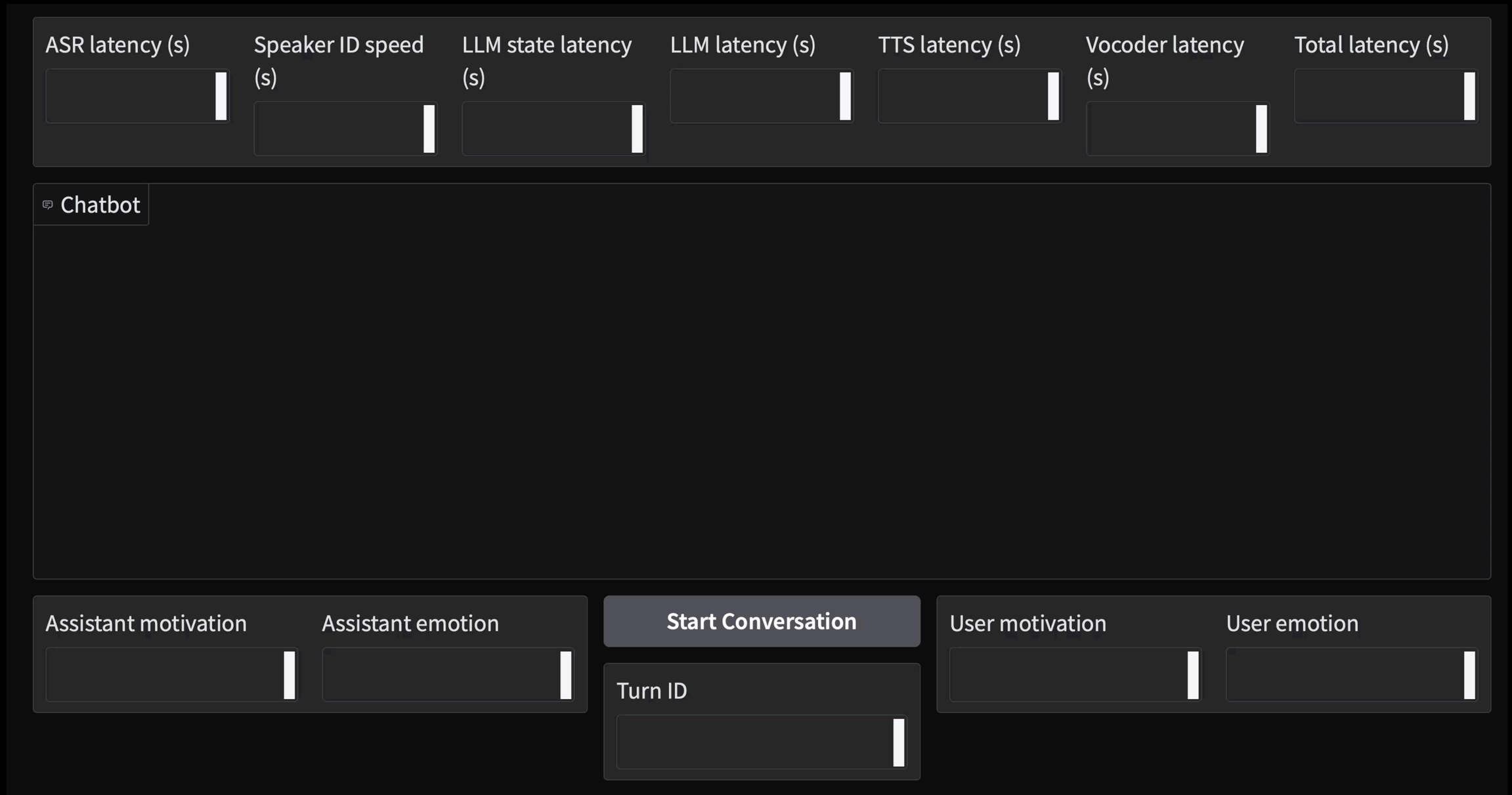
ChipChat : low-latency cascaded conversational agent

Running on Mac Studio M2 with **MLX**



ChipChat : low-latency cascaded conversational agent

Running on Mac Studio M2 with **MLX**



Inside ChipChat

Inside ChipChat :

speaker recognition

Zak Aldeneh

Speaker-IPL: Unsupervised Learning of Speaker Characteristics with i-Vector based Pseudo-Labels

ICASSP 2025

Takeaway : fully unsupervised model that achieves SOTA speaker recognition on diverse data

Inside ChipChat :

speaker recognition

Zak Aldeneh

Speaker-IPL: Unsupervised Learning of Speaker Characteristics with i-Vector based Pseudo-Labels

ICASSP 2025

speech recognition

Zijin Gu

Omni-Router: Sharing Routing Decisions in Sparse Mixture-of-Experts for Speech Recognition

ASRU 2025

Takeaway : simple MoE for ASR that outperforms Whisper-small on conversational data with 2x less active params

Inside ChipChat :

speaker recognition

Zak Aldeneh

Speaker-IPL: Unsupervised Learning of Speaker Characteristics with i-Vector based Pseudo-Labels

ICASSP 2025

speech recognition

Zijin Gu

Omni-Router: Sharing Routing Decisions in Sparse Mixture-of-Experts for Speech Recognition

ASRU 2025

language modeling

Yizhe Zhang

Sage: Steering Dialog Generation with Future-Aware State-Action Augmentation

EMNLP Workshop 2025

Takeaway : emotion-oriented LLM capabilities through conditioning on state-action estimation

Inside ChipChat :

speaker recognition

Zak Aldeneh

Speaker-IPL: Unsupervised Learning of Speaker Characteristics with i-Vector based Pseudo-Labels

ICASSP 2025

speech recognition

Zijin Gu

Omni-Router: Sharing Routing Decisions in Sparse Mixture-of-Experts for Speech Recognition

ASRU 2025

language modeling

Yizhe Zhang

Sage: Steering Dialog Generation with Future-Aware State-Action Augmentation

EMNLP Workshop 2025

speech generation

Richard Bai

SpeakStream: Streaming Text-to-Speech with Interleaved Data

Preprint

Takeaway : first streaming TTS with SOTA latency <50ms on M4

Inside ChipChat :

speaker recognition

Zak Aldeneh

Speaker-IPL: Unsupervised Learning of Speaker Characteristics with i-Vector based Pseudo-Labels

ICASSP 2025

speech recognition

Zijin Gu

Omni-Router: Sharing Routing Decisions in Sparse Mixture-of-Experts for Speech Recognition

ASRU 2025

language modeling

Yizhe Zhang

Sage: Steering Dialog Generation with Future-Aware State-Action Augmentation

EMNLP Workshop 2025

speech generation

Richard Bai

SpeakStream: Streaming Text-to-Speech with Interleaved Data

Preprint

speech vocoding

Tatiana Likhomanenko

VocStream

Part of SpeakStream

Takeaway : first streaming Vocoder with SOTA latency <15ms on M4

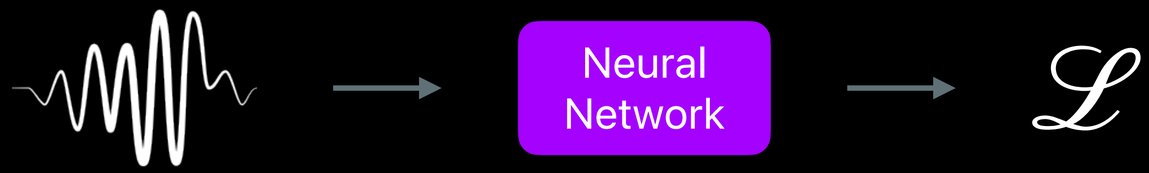
Speaker-IPL

Aldeneh, Z. and et al. *Speaker-IPL: Unsupervised Learning of Speaker Characteristics with i-Vector based Pseudo-Labels*
https://arxiv.org/pdf/2409.10791, ICASSP 2025

Zak
Aldeneh

Speaker representation learning

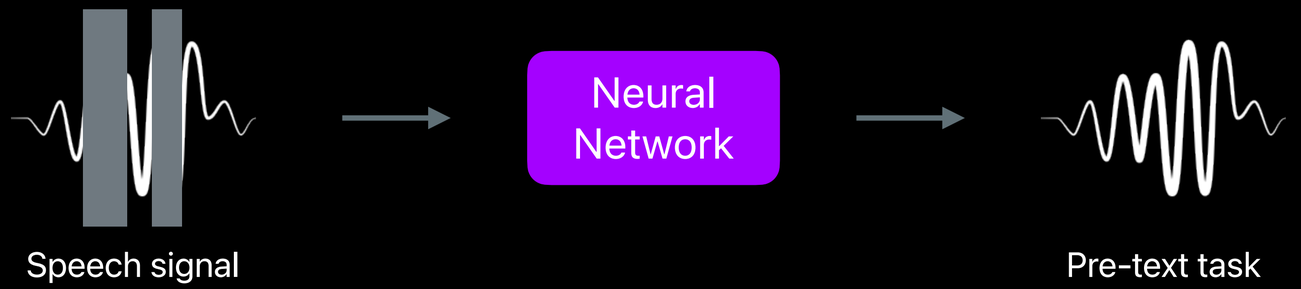
Speaker models can be trained using either supervised or unsupervised learning methods



Speech signal

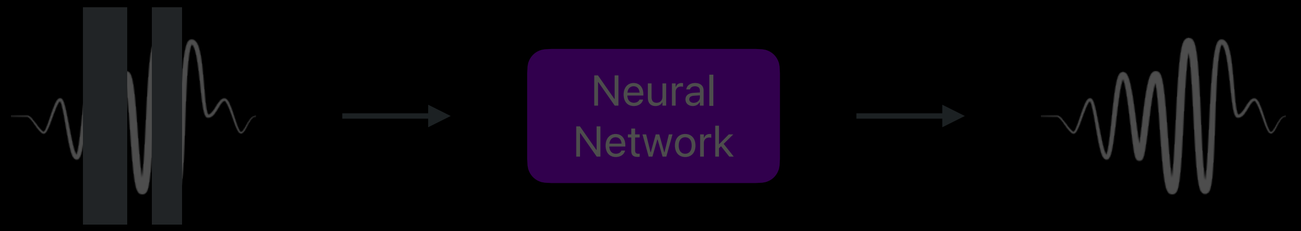
Unsupervised methods

Training phase:

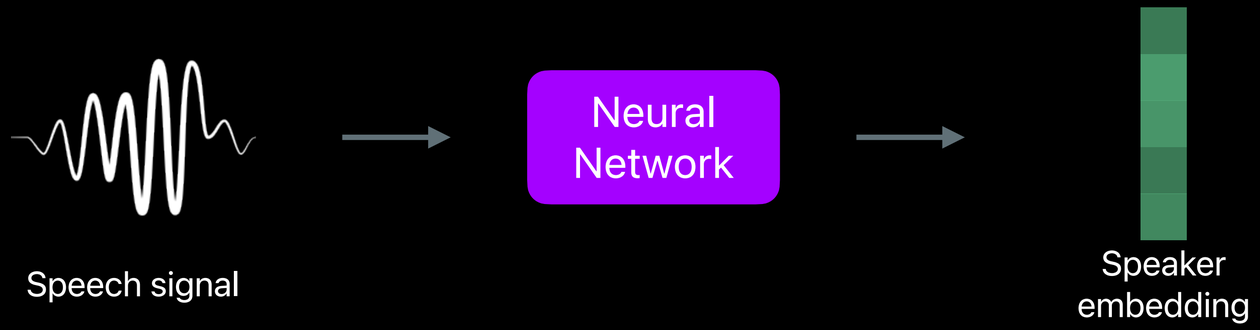


Unsupervised methods

Training phase:

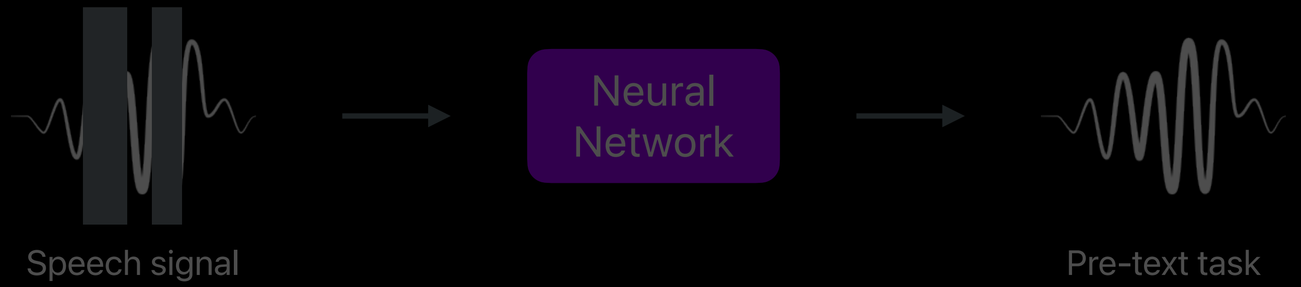


Inference phase:

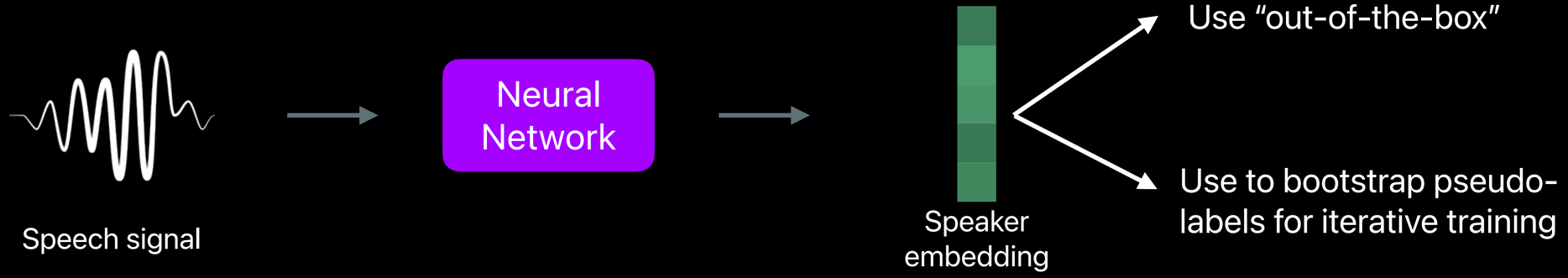


Unsupervised methods

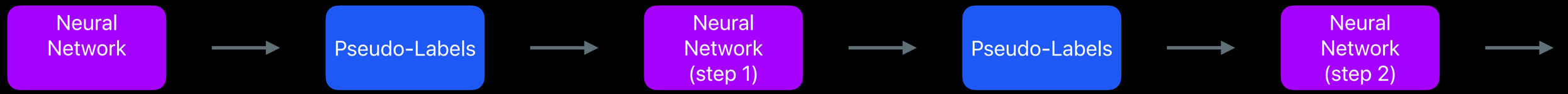
Training phase:



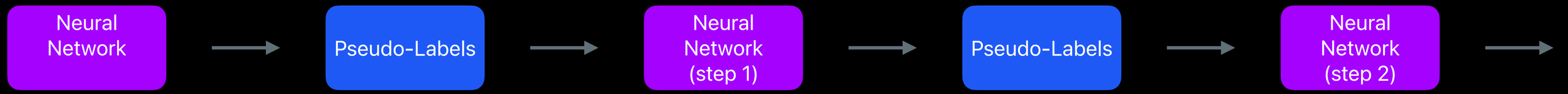
Inference phase:



Iterative training



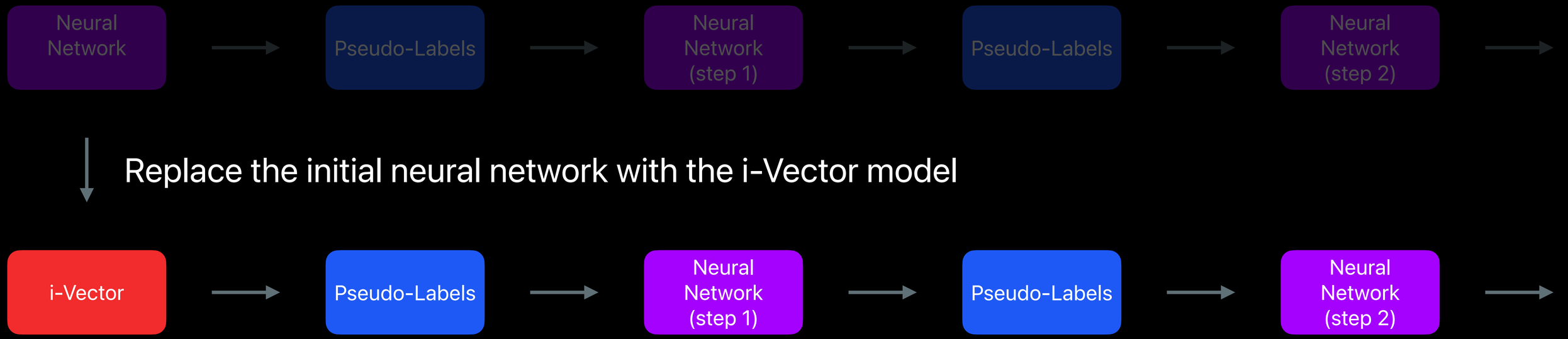
Iterative training



Do we really need a strong model to bootstrap the iterative self-training process ?

Can we instead use the i-vector model to bootstrap the process ?

Iterative training with i-Vector



Setup

Data:

- VoxCeleb2 [Chung et al., Interspeech 2018]

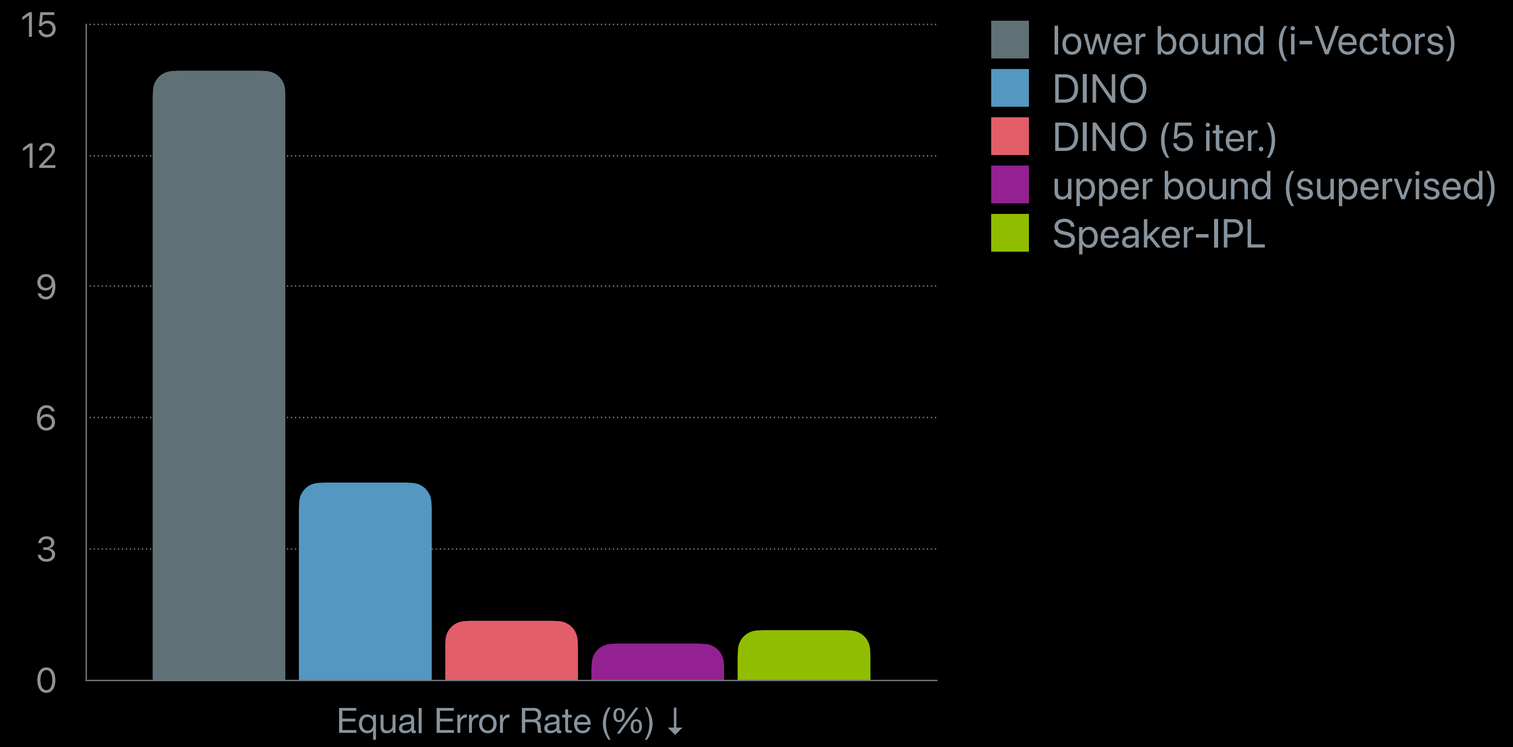
Encoder:

- MFA-Conformer [Zhang et al. Interspeech, 2022], ECAPA-TDNN [Desplanques et al., Interspeech, 2020]

Augmentations:

- Additive noise [Snyder et al., arXiv preprint arXiv:1510.08484, 2015]
- Reverberation [Ko et al., ICASSP 2017]

Results



Number of clusters matters (should be covering training data)

Augmentation improves performance

MFA-Conformer is better than ECAPA-TDNN

Agglomerative clustering improves performance

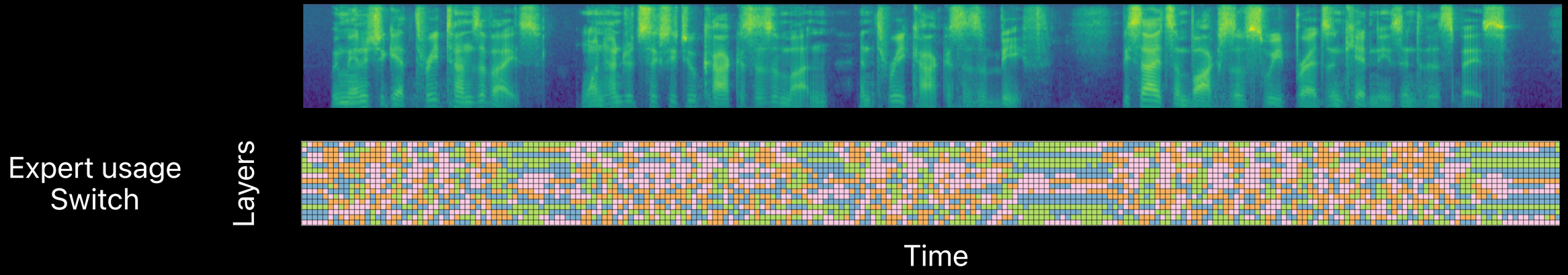
Omni-router ASR

Gu, Z. and et al. *Omni-Router: Sharing Routing Decisions in Sparse Mixture-of-Experts for Speech Recognition*
https://arxiv.org/abs/2507.05724, ASRU 2025

Zijin
Gu

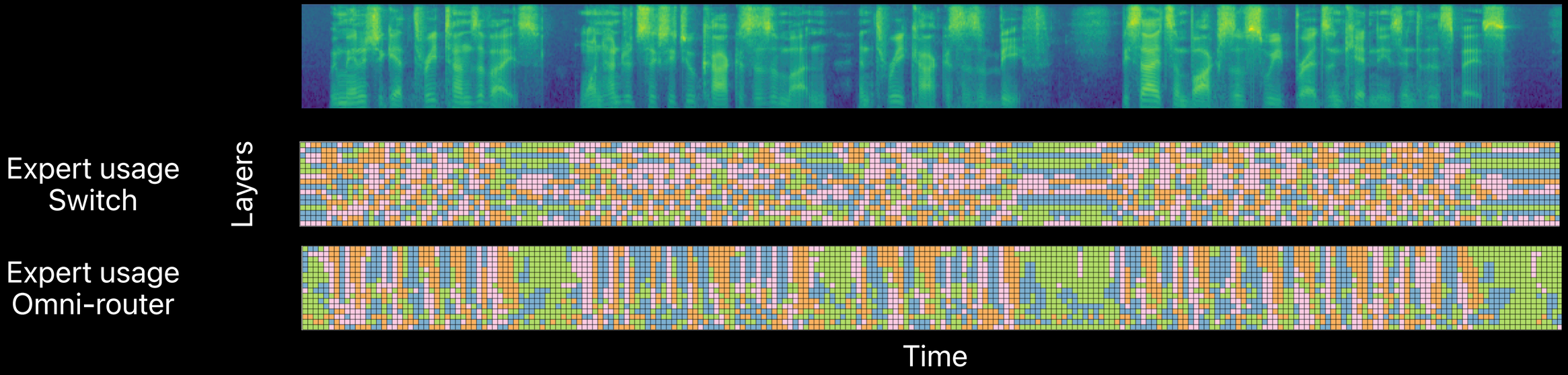
Improving ASR with specialized mixture of experts (MOE)

Different layers seem to route data to experts independently



Omni-router transformer experts

Different layers seem to route data to experts independently



Omni-router encourages :

- explicit coordination among layers
- stronger specialization
  e.g. structured expert usage correlated with Mel-spectrogram

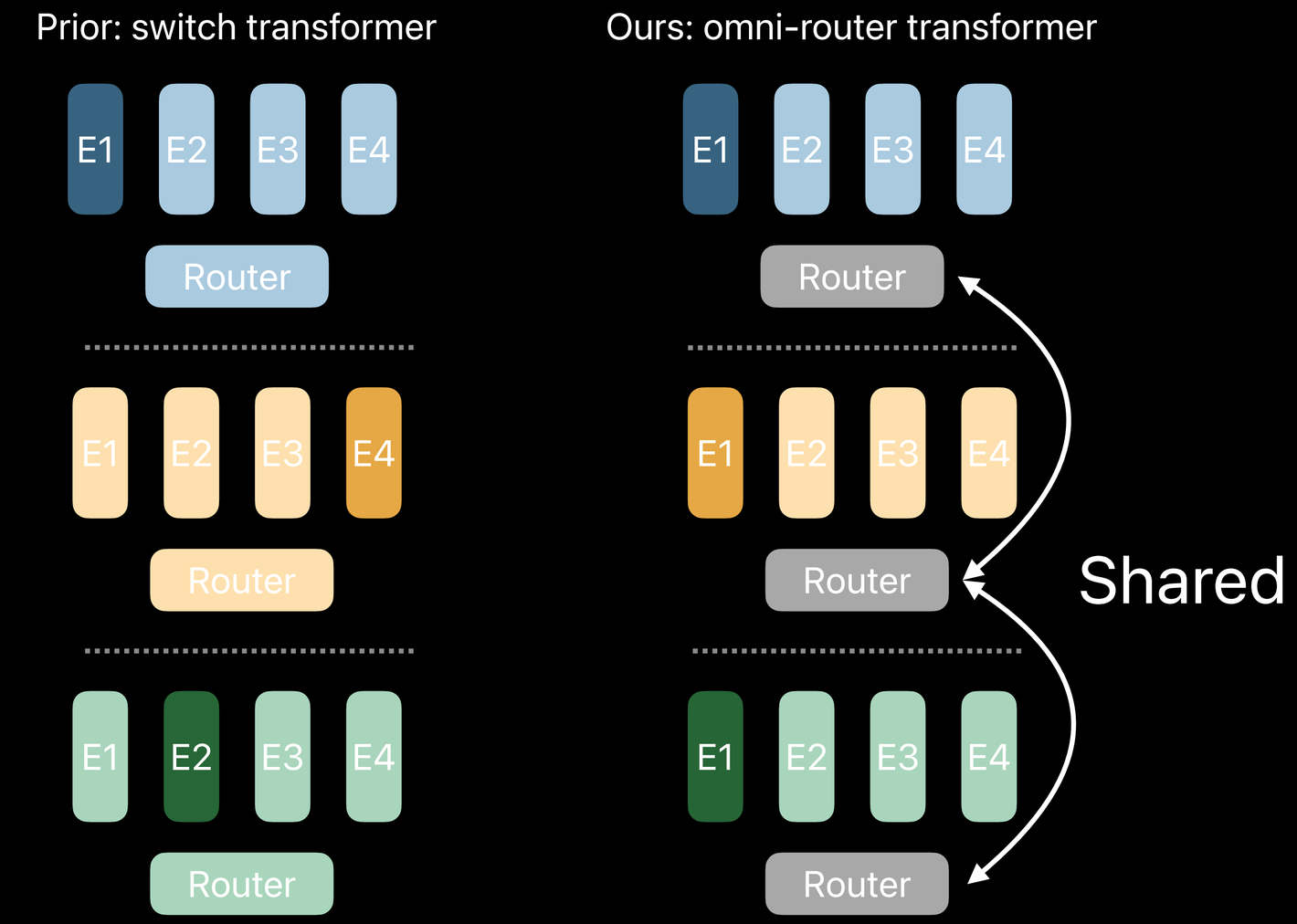
Omni-router : sharing router weights between layers

Improves

- accuracy
- training stability

Simple and straightforward

- no embedding network
- no loss combinations
- no inductive bias



Omni-router : training data

SpeechCrawl dataset: collect large-scale conversational audio data from publicly accessible sources

Generate pseudo-labels (WhisperX pipeline + Whisper’s large-v2 model)

Filter unreliable pseudo-labels

Retain approximately 1M hours of English speech segments for training

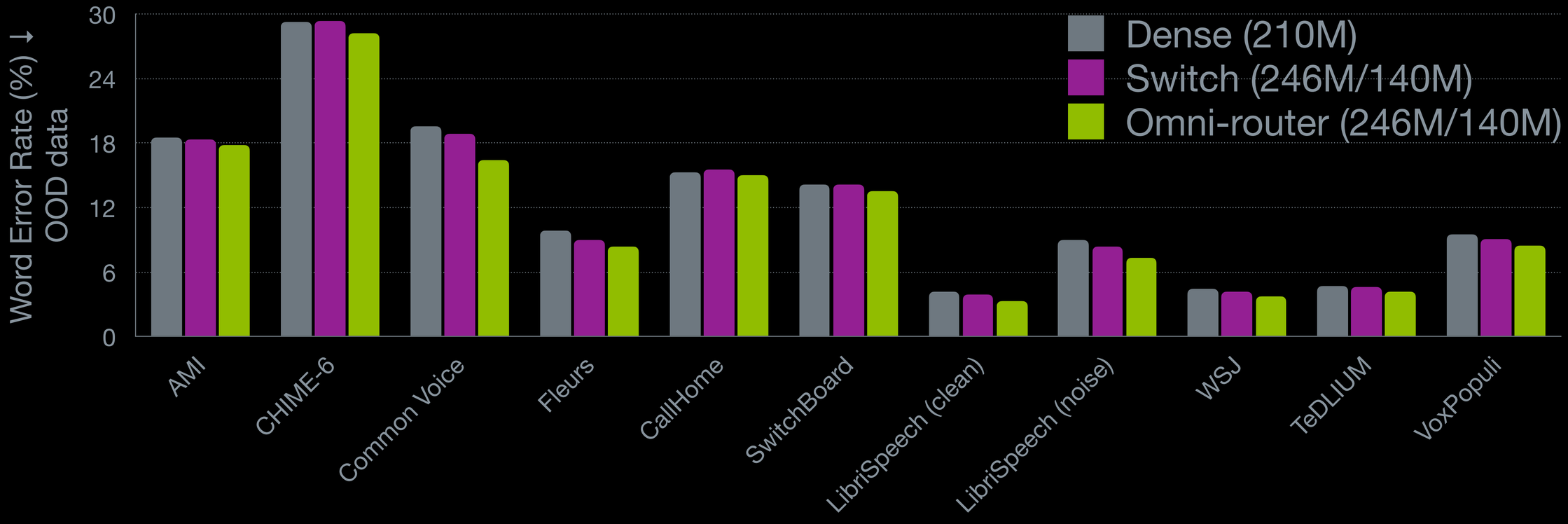
Omni-router : models

Baselines : dense and switch vanilla transformer models

All models are encoder-based trained with CTC loss

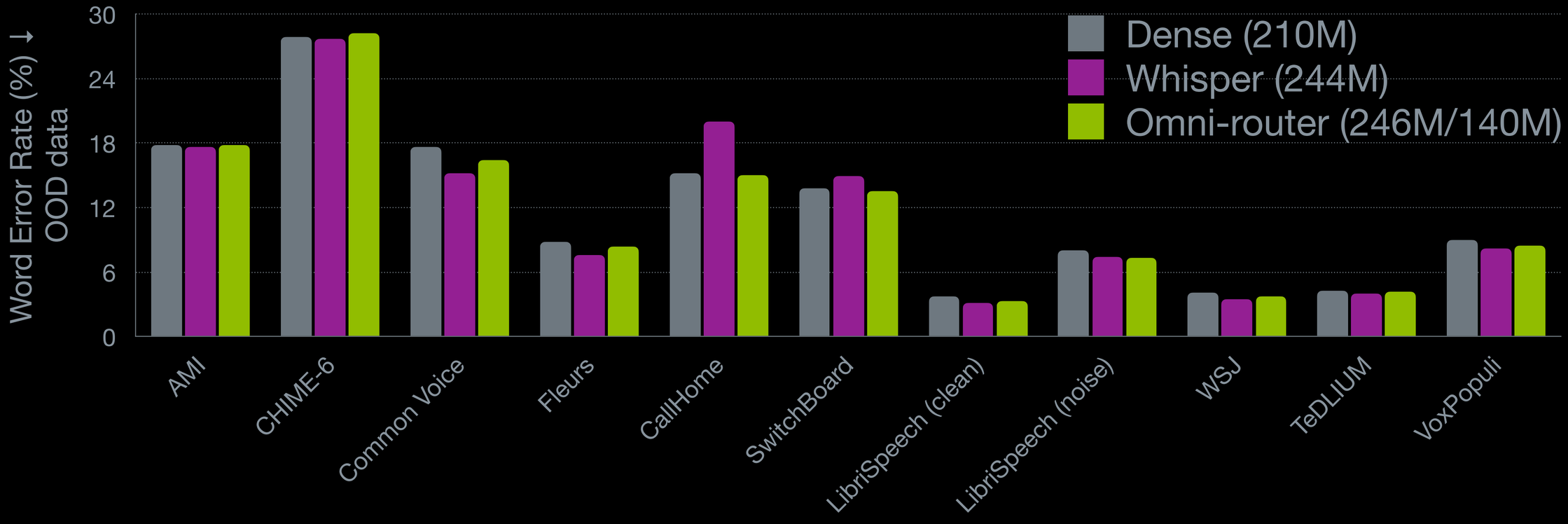
Omni-router : results

Omni-router improves ASR across different OOD data
(total : 246M parameters, active : 140M parameters, 2 experts)



Omni-router : results

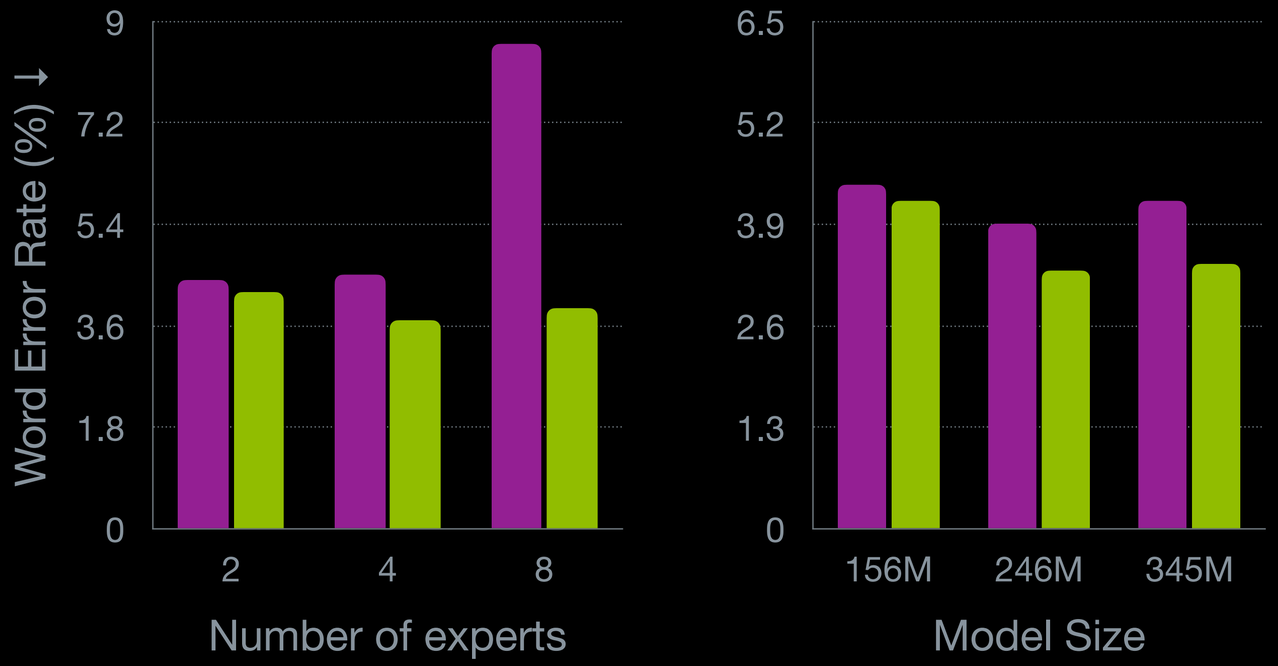
Omni-router with CTC (no LM) is comparable with Whisper of similar size (total : 246M parameters, active : 140M parameters, 2 experts)



Omni-router : results

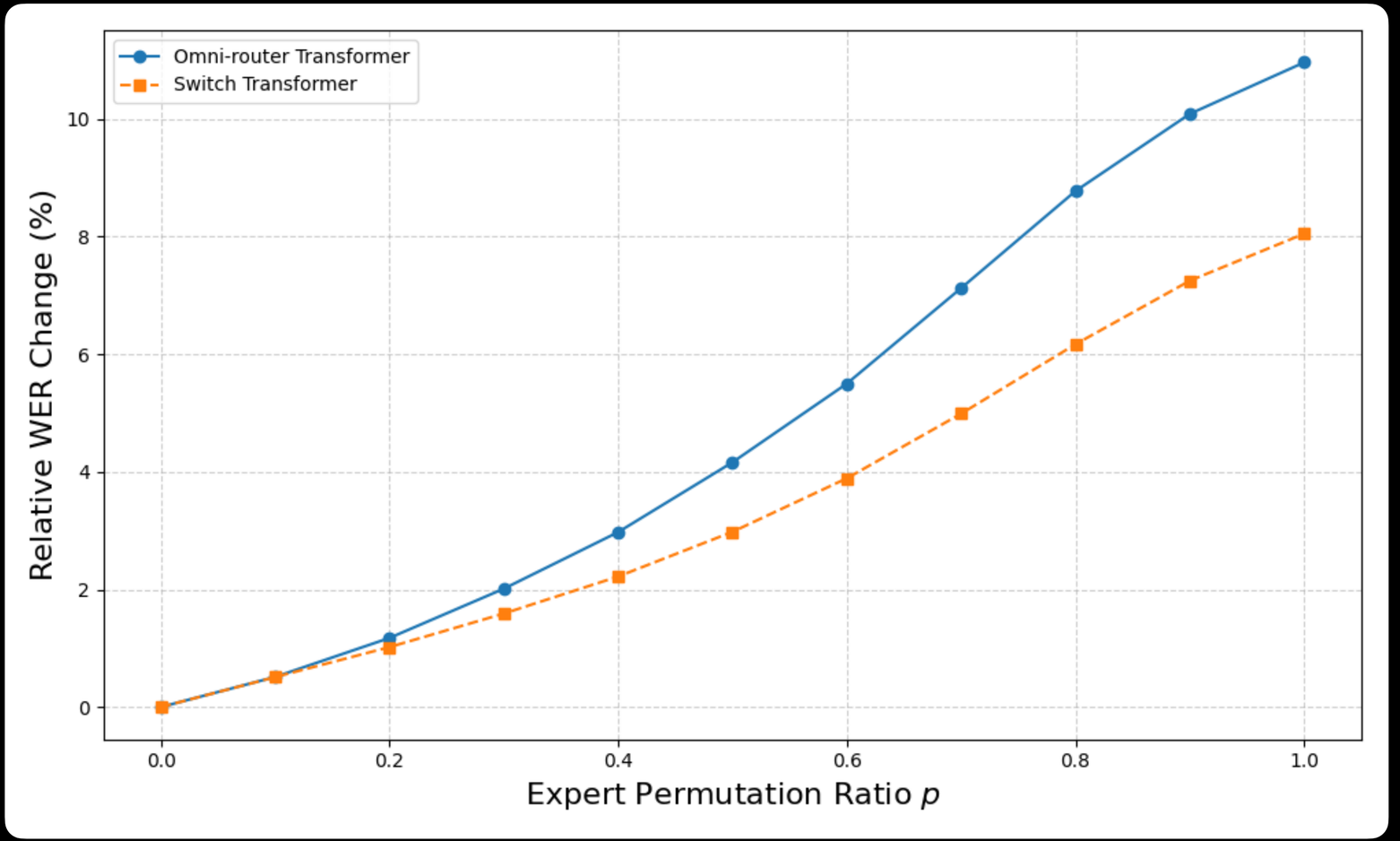
Omni-router improves ASR across

- different number of experts
- different model sizes



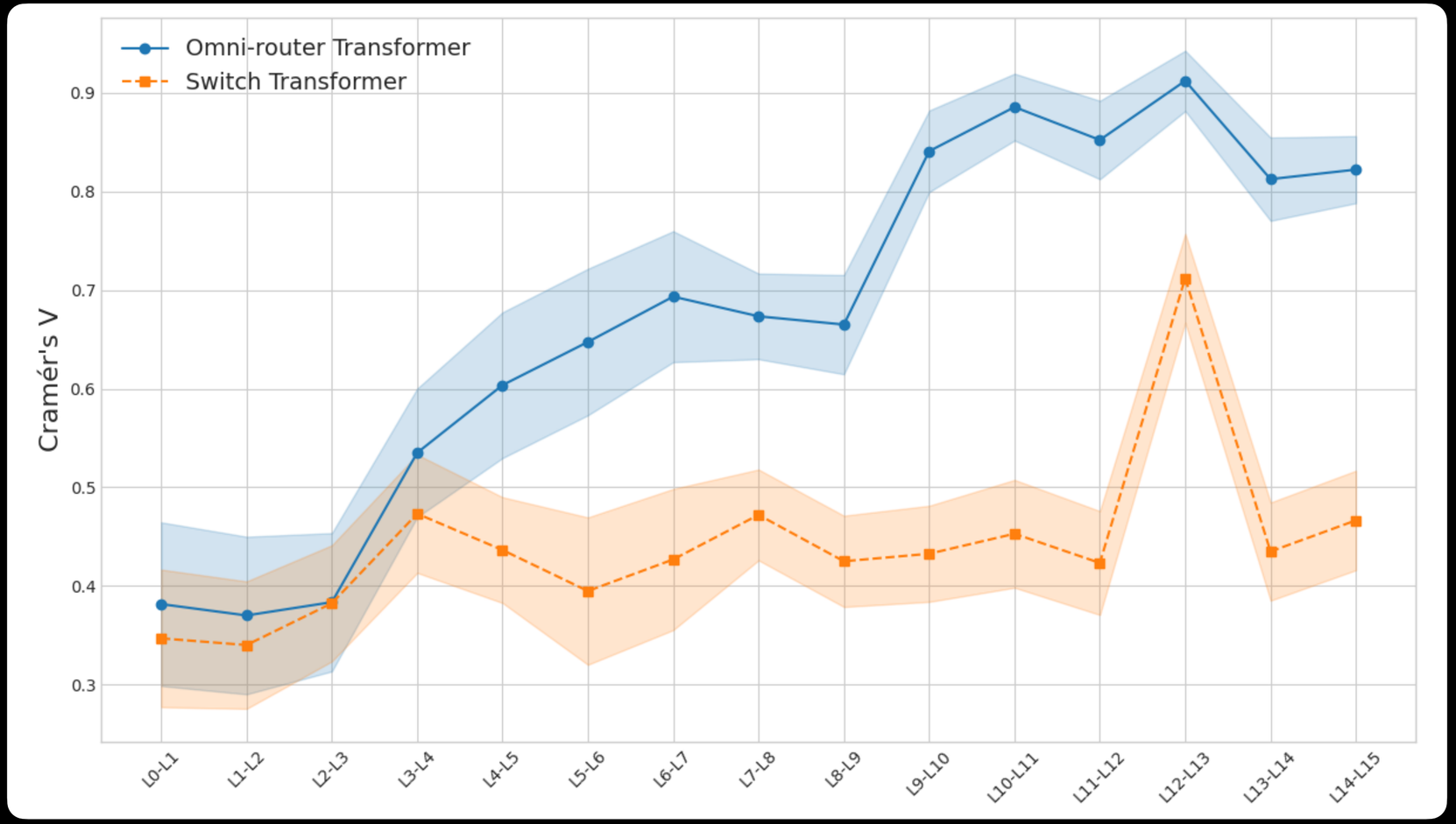
Omni-router : greater expert specialization

Randomly permute expert assignments for each token with varying probability *p* and measure WER degradation



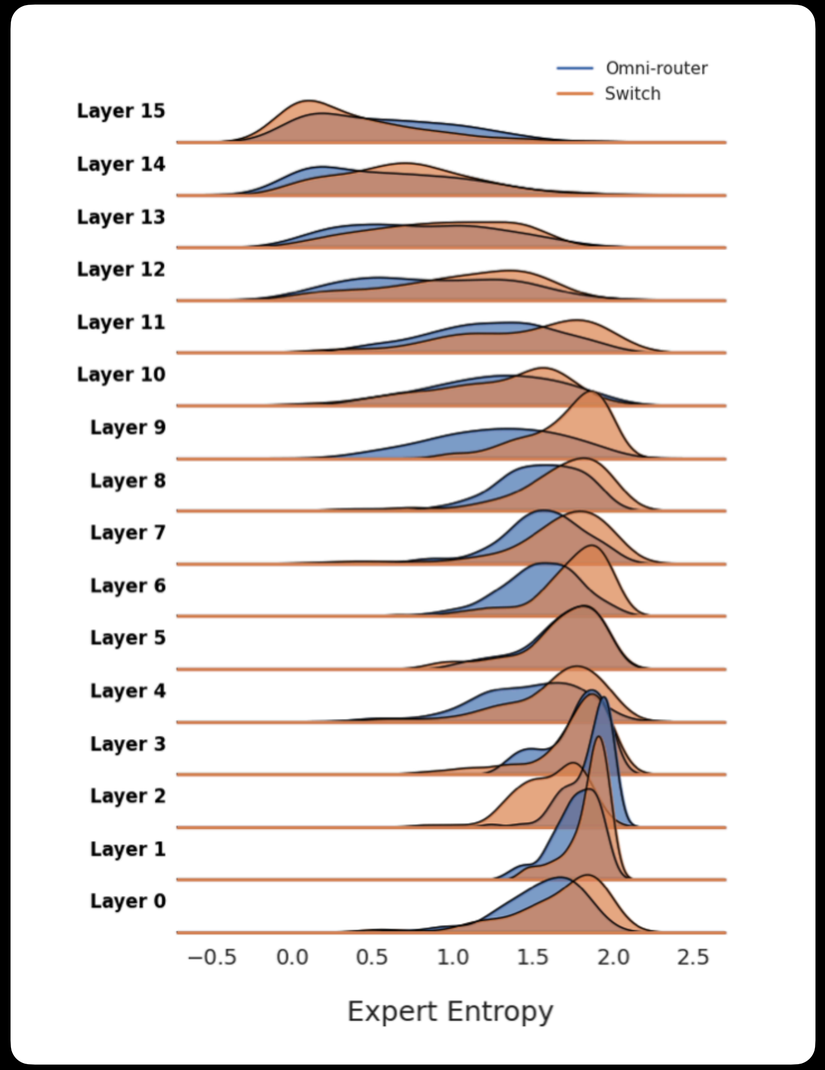
Omni-router : enhanced inter-layer cooperation

Examined the consistency of expert choices between adjacent layers



Omni-router : lower expert entropy

Analyzed the confidence in routing decisions by evaluating expert entropy distributions across layers for the top 100 most frequently occurring tokens



SAGE LLM

Zhang, Y. and Jaitly, N. *Sage: Steering Dialog Generation with Future-Aware State-Action Augmentation*
https://arxiv.org/abs/2503.03040, EMNLP Workshop 2025

Yizhe
Zhang

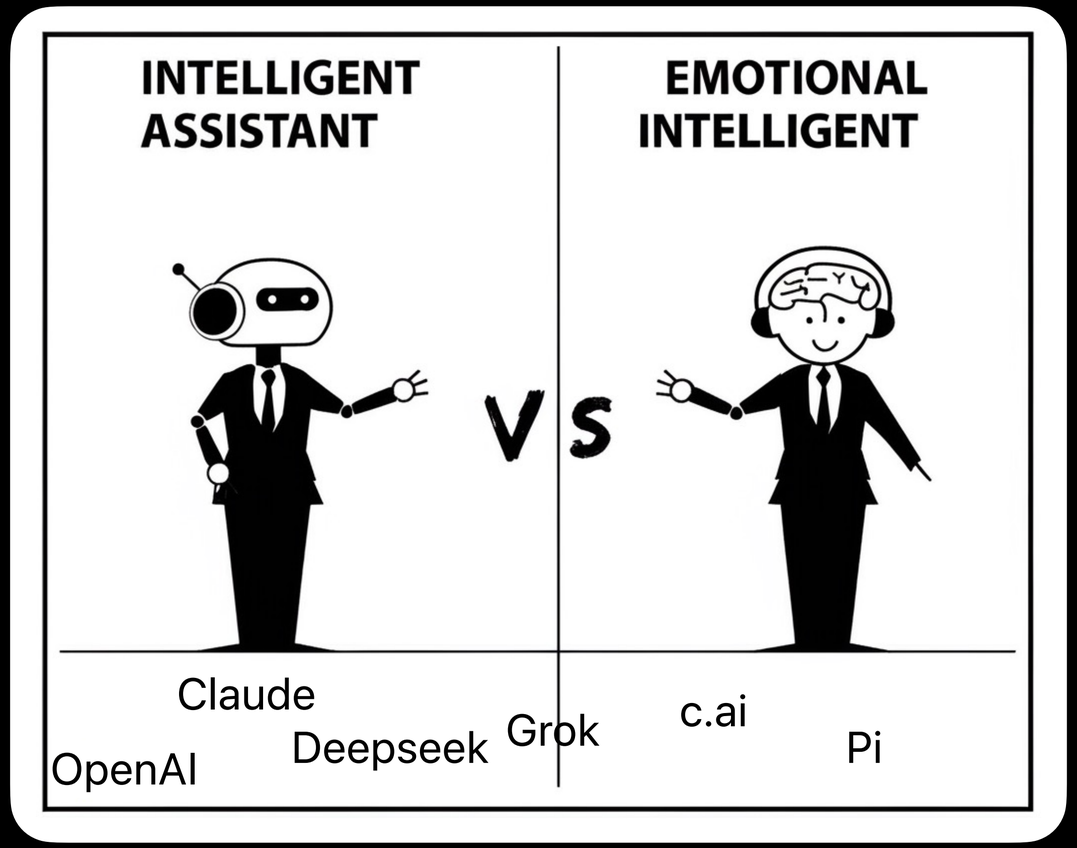
Problem-oriented Agent v.s. Emotion-oriented Agent

Passive

QA

Code

Math



Proactive

Empathy

Trust

Engagement

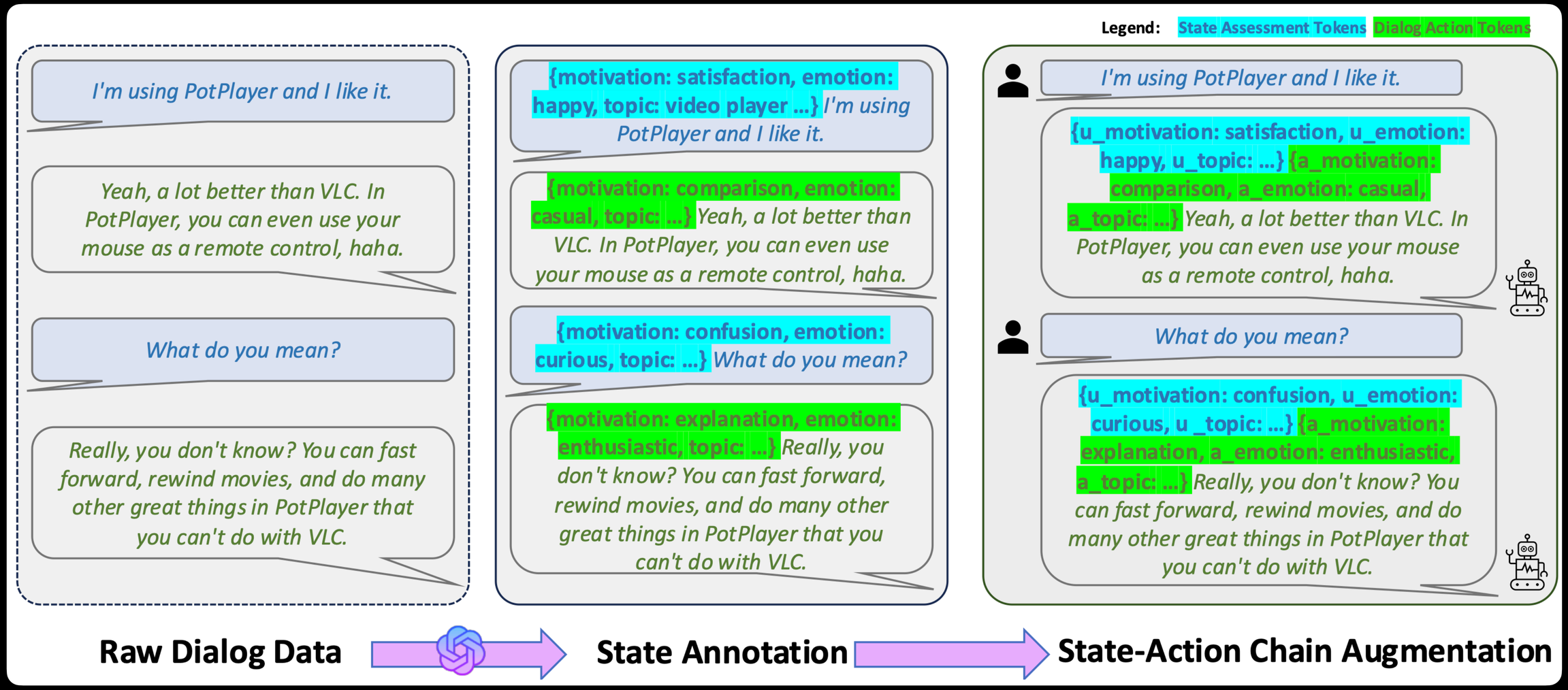
State-action chain augmentation

Introduce latent variables which

- represent longer-term conversational information
- encapsulate emotional and conversational states
- are of discrete nature

State-action chain augmentation

Augment in-house dataset extracted from Reddit with state-action



State-action chain augmentation

Training :

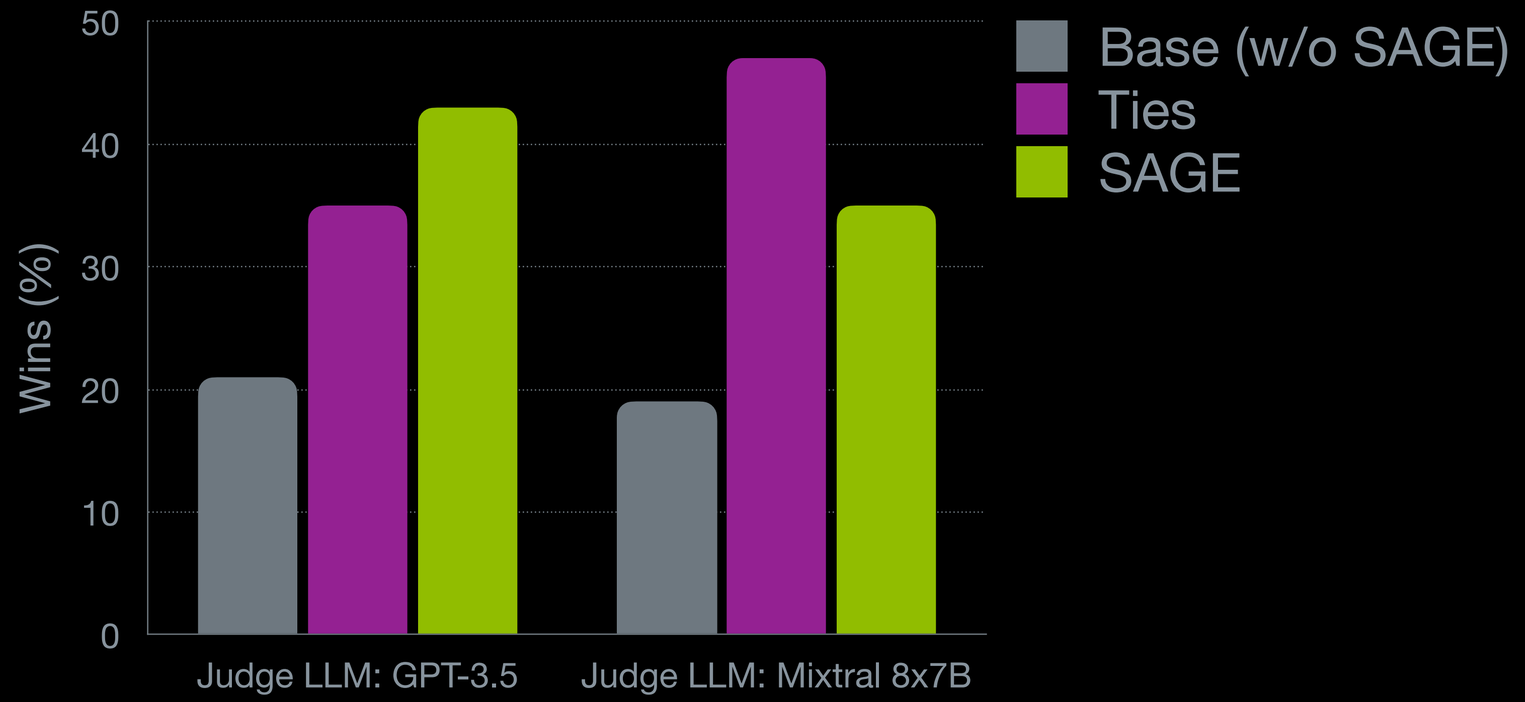
- Fine-tune (LoRA) initial LLM (Mixtral 8x7B) on data with state-action augmentation
- Self-play with initial conversation from EmpatheticDialogs
  - Current model is agent SAGE_1 is user
  - Generate up to 12 turns
  - External selector (Mixtral 8x7B) evaluates and selects best trajectories

Inference :

- generate latent variables
- generate dialog response

Comparing dialog responses with base model

Fine-tune Mixtral 8x7B model using our state-action augmentation



Comparing dialog responses with base model

Fine-tune Mixtral 8x7B model using our state-action augmentation

Only slight degradation on standard LLM benchmarks except GSM8k

TTS + Vocoder

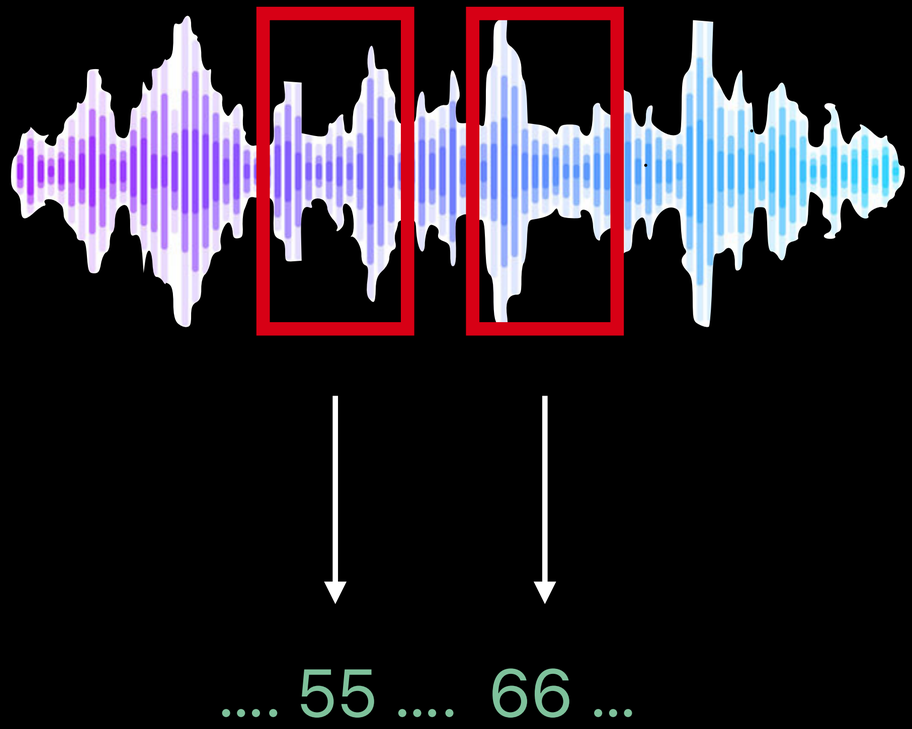
Tokenization of text

Simple mapping

.....
great: <778>
ID: <779>
small: <780>
through: <781>
application: <782>
.....

Tokenization of audio

Convert windows of waveforms into discrete tokens using either contrastive losses or VQ-VAEs



Why is tokenization necessary ?

We are very good at language modeling and want to reuse this technology universally

Seems easier to model tokenized data compared to raw data

- Less important details are probably compressed in data

Main components :

- Data is a sequence of discrete indices
- Chain rule is used to model the probability distribution autoregressively

The case against tokenization

Tokenization compresses data before you know what you want to do with it

The manifold of the compressed codes can be quite curved

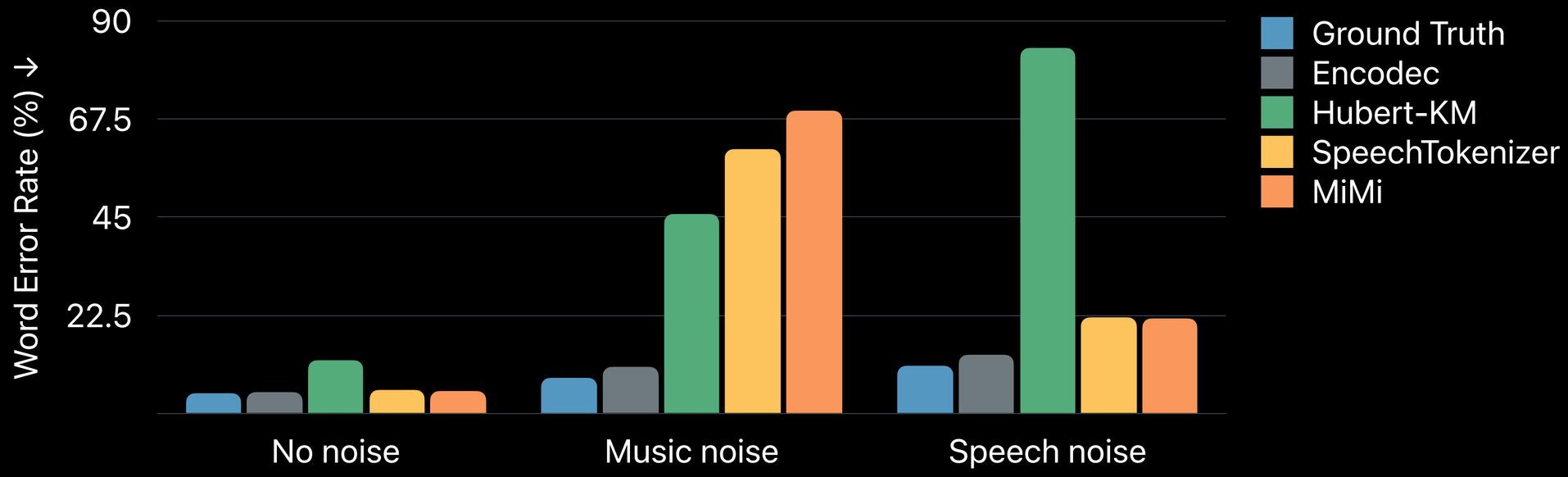
- Downstream models may go off the manifold in weird ways
- Could make things worse in out of domain settings

The case against tokenization

Tokenization compresses data before you know what you want to do with it

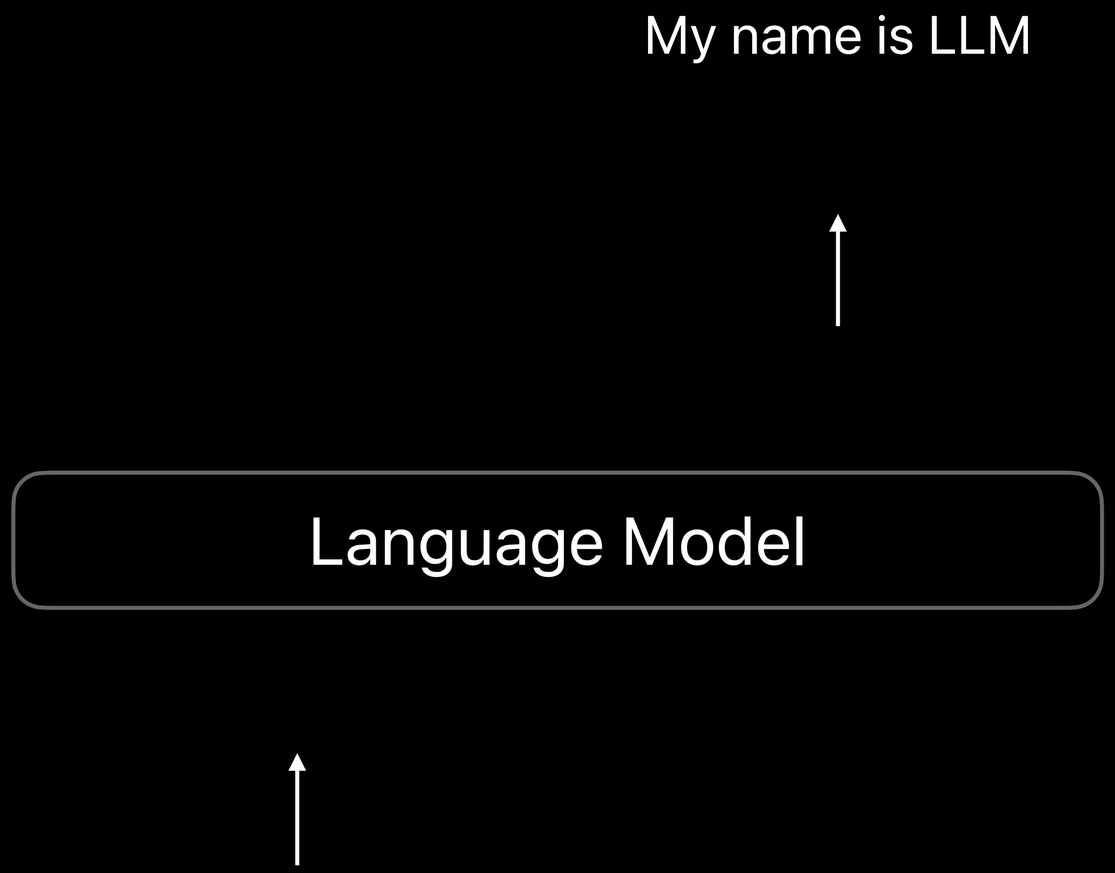
The manifold of the compressed codes can be quite curved

- Downstream models may go off the manifold in weird ways
- Could make things worse in out of domain settings



Why is tokenization necessary ?

Reuse language modeling machinery for autoregressive modeling of speech by tokenization



Why is tokenization necessary ?

Reuse language modeling machinery for autoregressive modeling of speech by tokenization

My name is LLM

Language Model

Speech Language Model

What is your name

Why is tokenization necessary ?

Reuse language modeling machinery for autoregressive modeling of speech by tokenization

what: <13>
is: <17>
my: <21>
name: <3>
llm: <111>
.....

My name is LLM

<21> <3> ..... <111>

Language Model

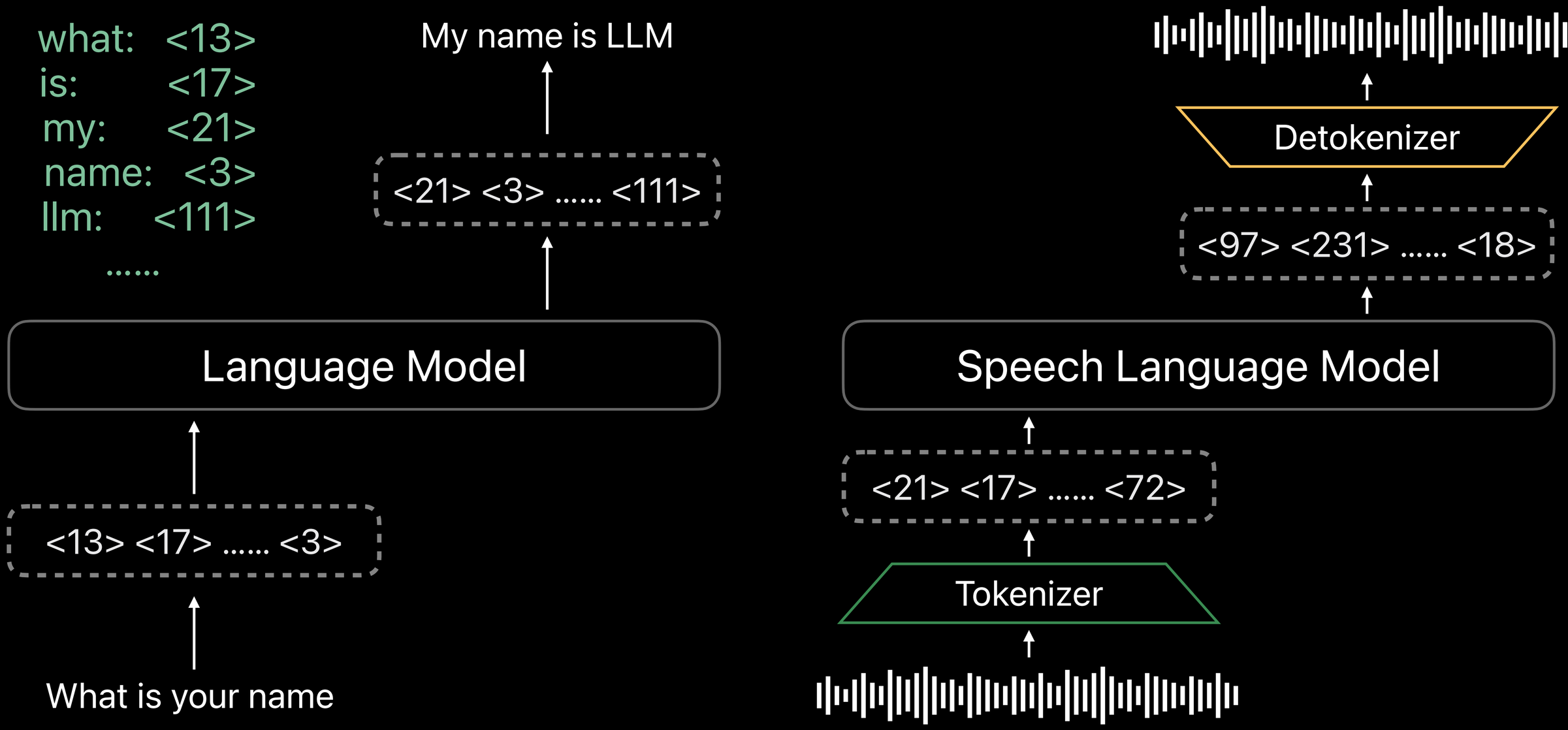
<13> <17> ..... <3>

What is your name

Speech Language Model

Why is tokenization necessary ?

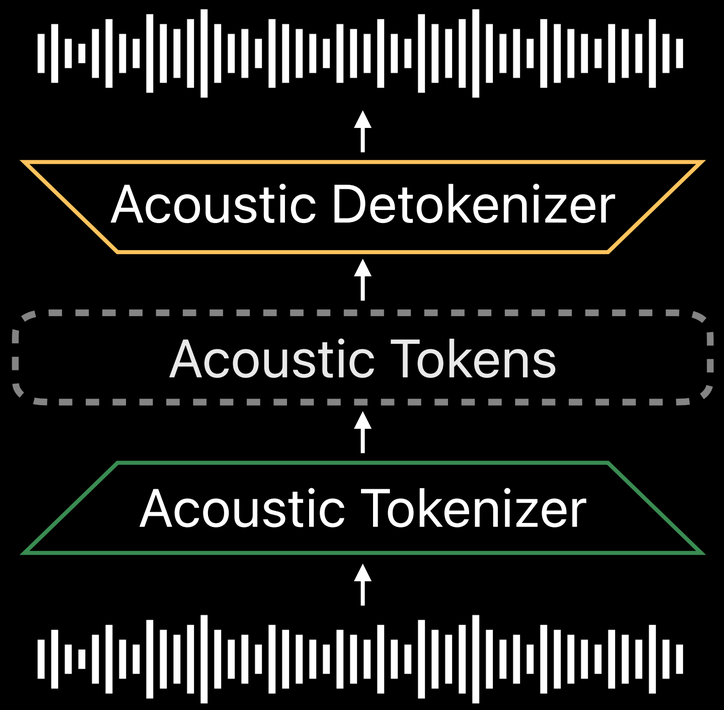
Reuse language modeling machinery for autoregressive modeling of speech by tokenization



Popular generative models

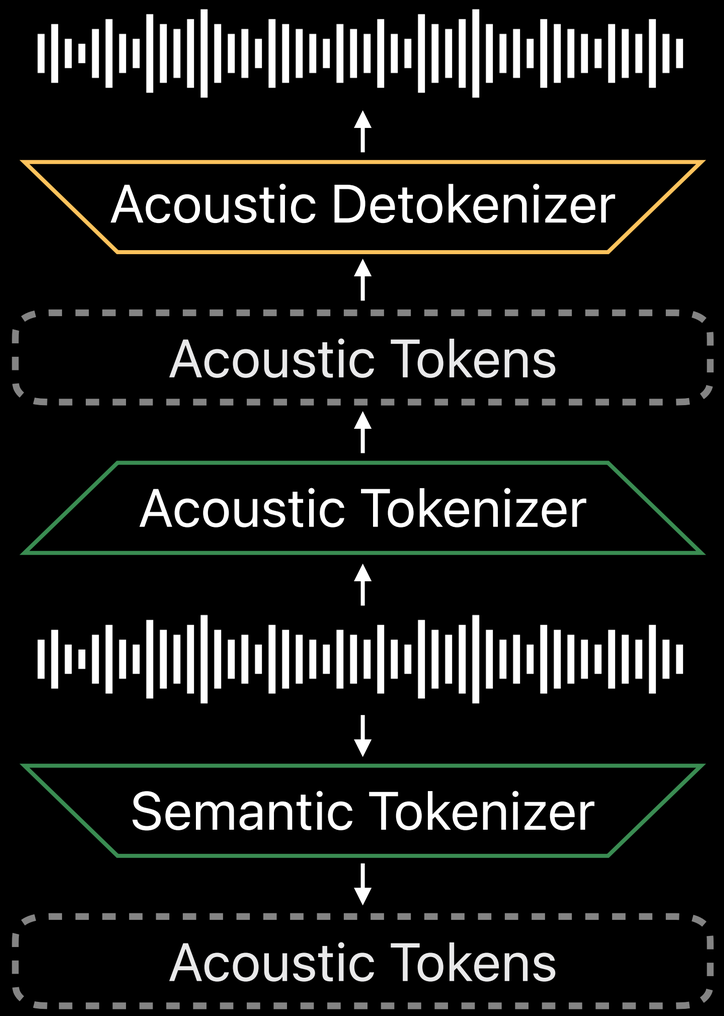
Borsos, Zalán et al. “AudioLM: A Language Modeling Approach to Audio Generation.” IEEE/ACM Transactions on Audio, Speech, and Language Processing 31 (2022): 2523-2533.

Popular generative models



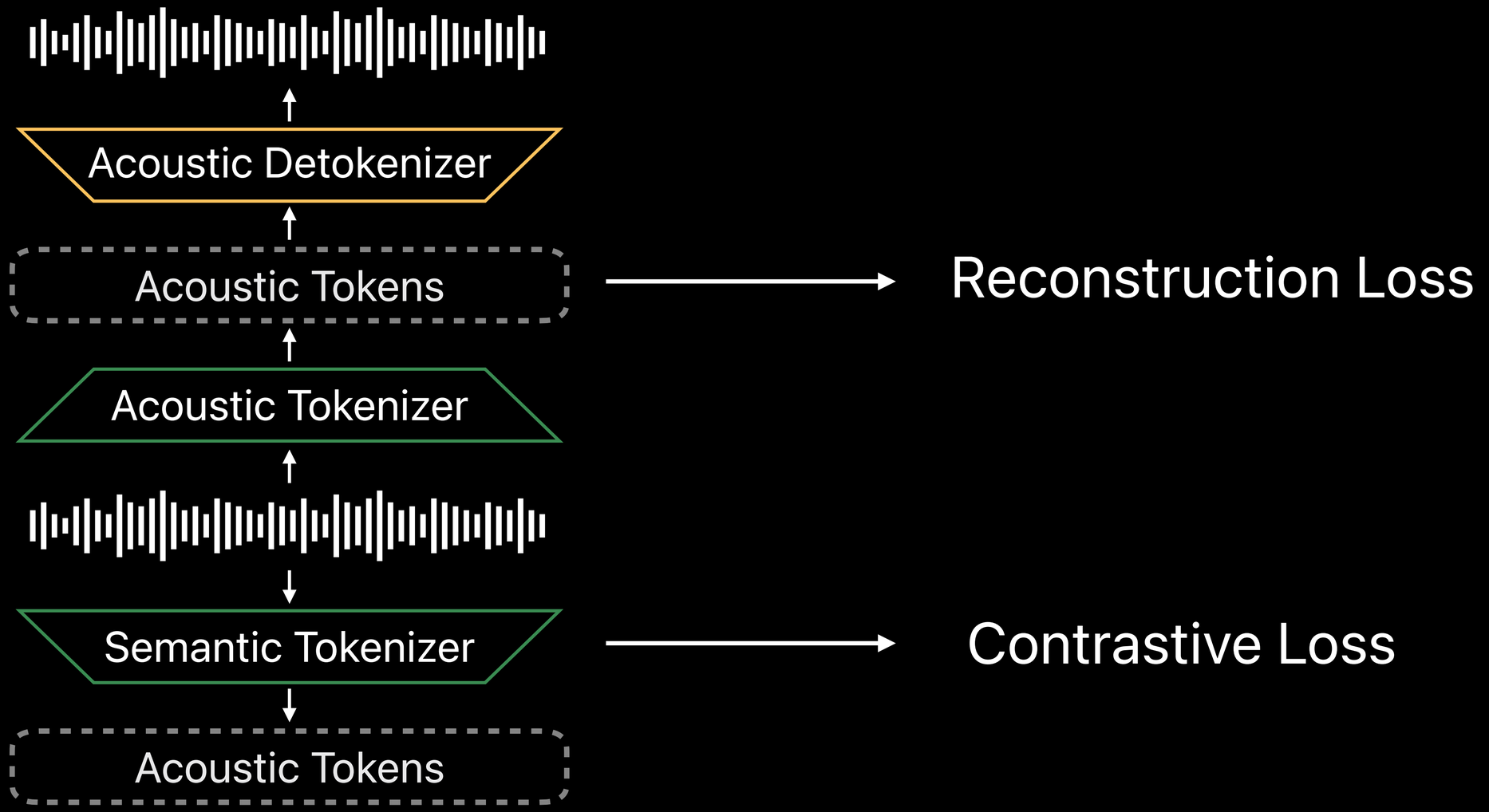
Borsos, Zalán et al. “AudioLM: A Language Modeling Approach to Audio Generation.” IEEE/ACM Transactions on Audio, Speech, and Language Processing 31 (2022): 2523-2533.

Popular generative models



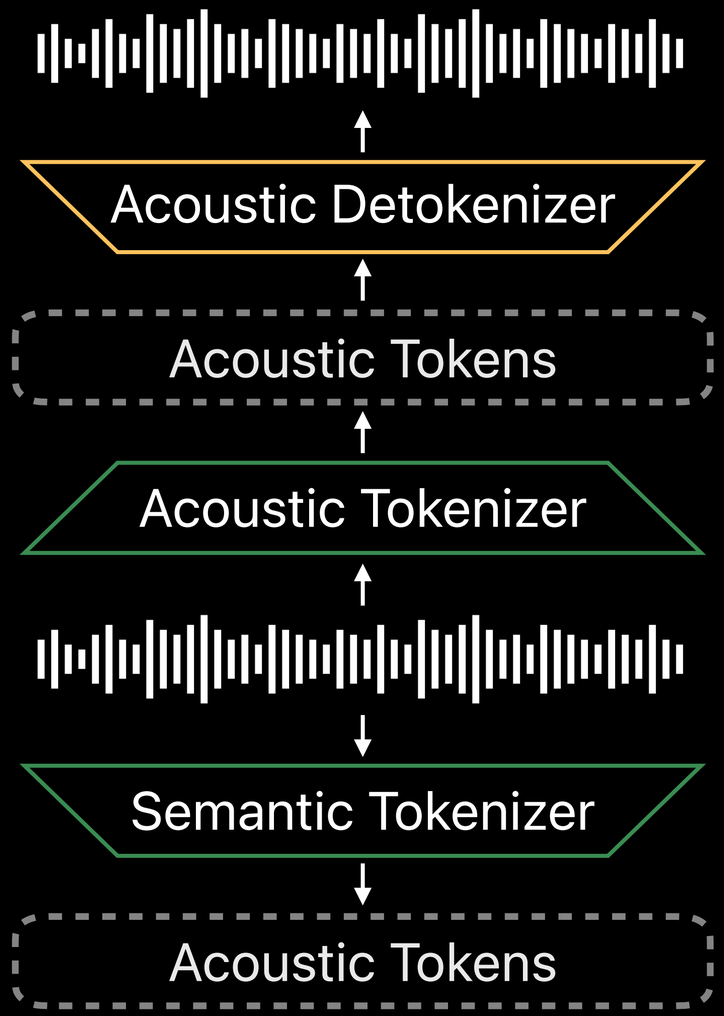
Borsos, Zalán et al. “AudioLM: A Language Modeling Approach to Audio Generation.” IEEE/ACM Transactions on Audio, Speech, and Language Processing 31 (2022): 2523-2533.

Popular generative models



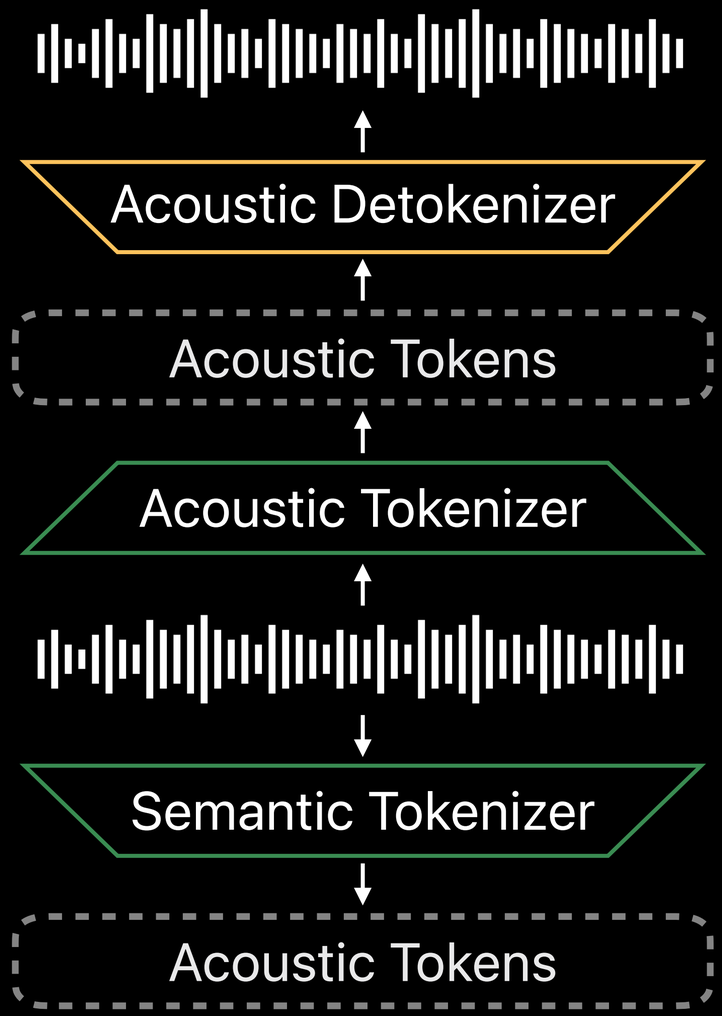
Borsos, Zalán et al. “AudioLM: A Language Modeling Approach to Audio Generation.” IEEE/ACM Transactions on Audio, Speech, and Language Processing 31 (2022): 2523-2533.

Popular generative models



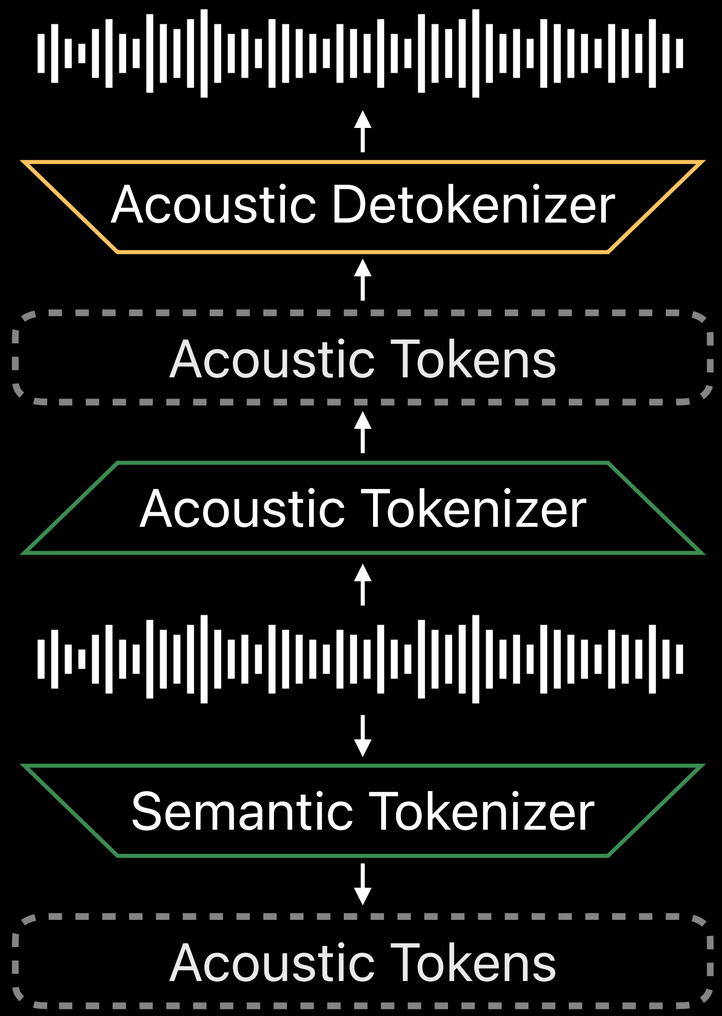
Borsos, Zalán et al. “AudioLM: A Language Modeling Approach to Audio Generation.” IEEE/ACM Transactions on Audio, Speech, and Language Processing 31 (2022): 2523-2533.

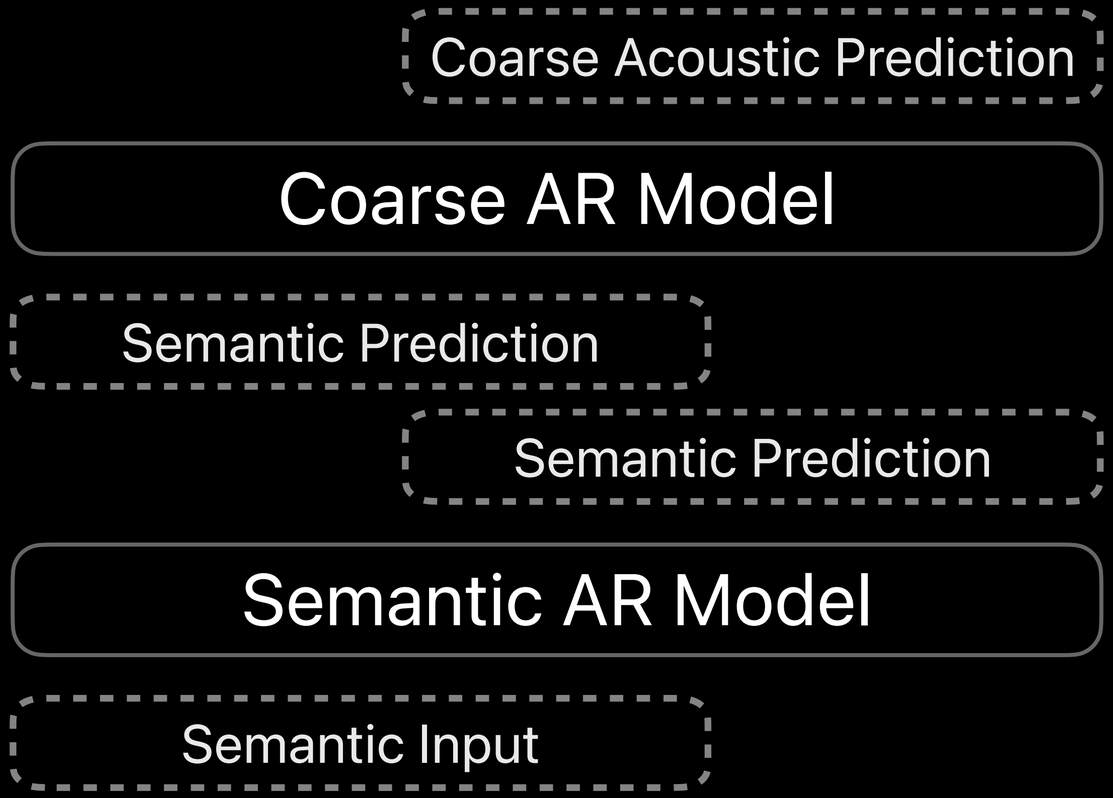
Popular generative models



Borsos, Zalán et al. “AudioLM: A Language Modeling Approach to Audio Generation.” IEEE/ACM Transactions on Audio, Speech, and Language Processing 31 (2022): 2523-2533.

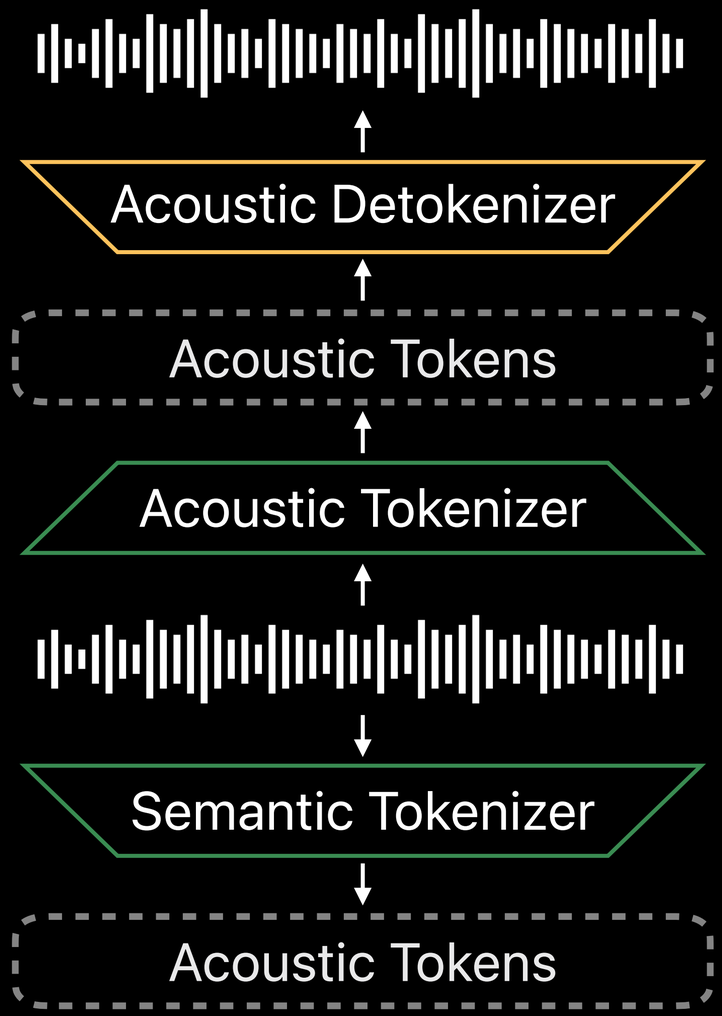
Popular generative models

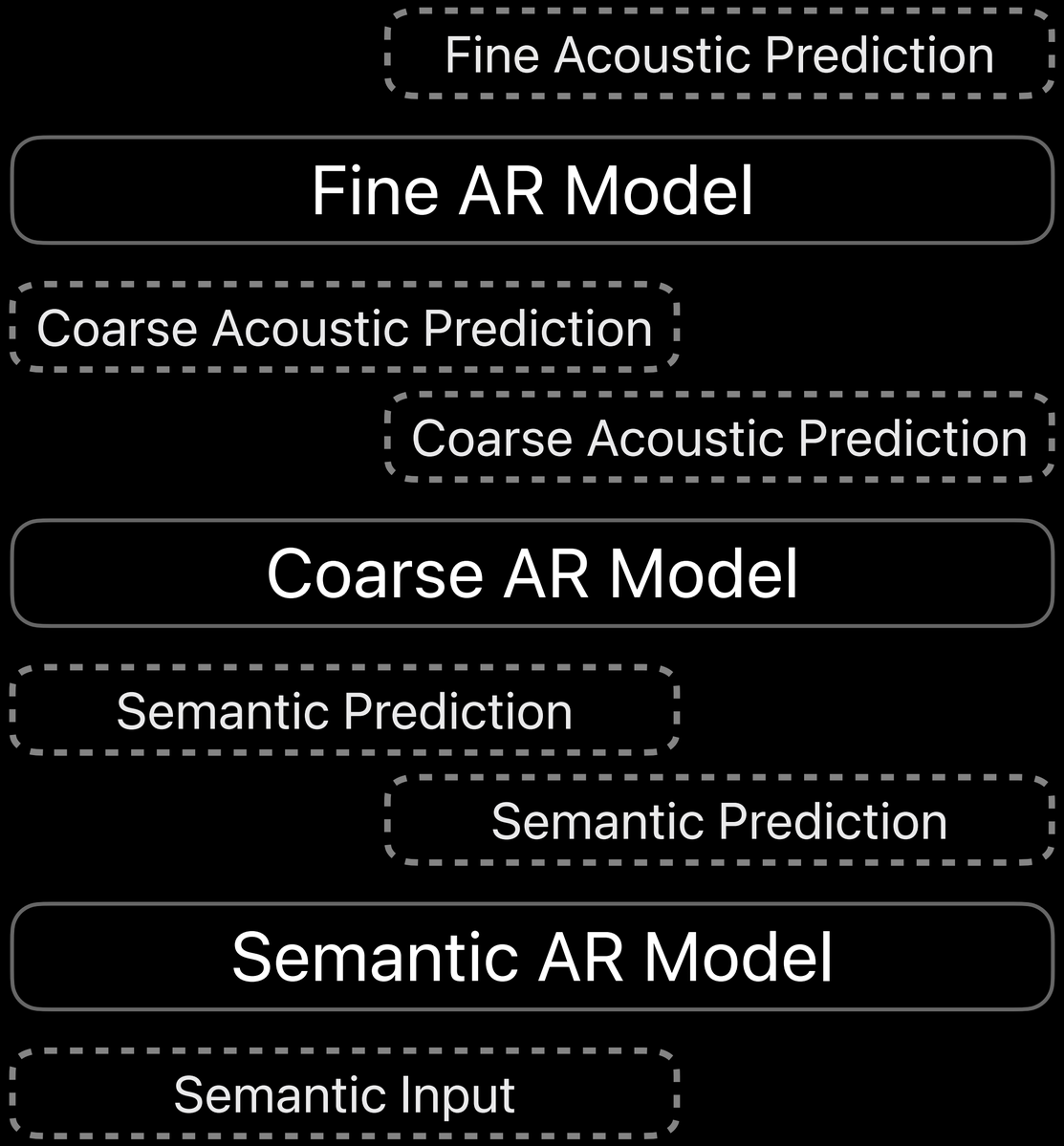




Borsos, Zalán et al. “AudioLM: A Language Modeling Approach to Audio Generation.” IEEE/ACM Transactions on Audio, Speech, and Language Processing 31 (2022): 2523-2533.

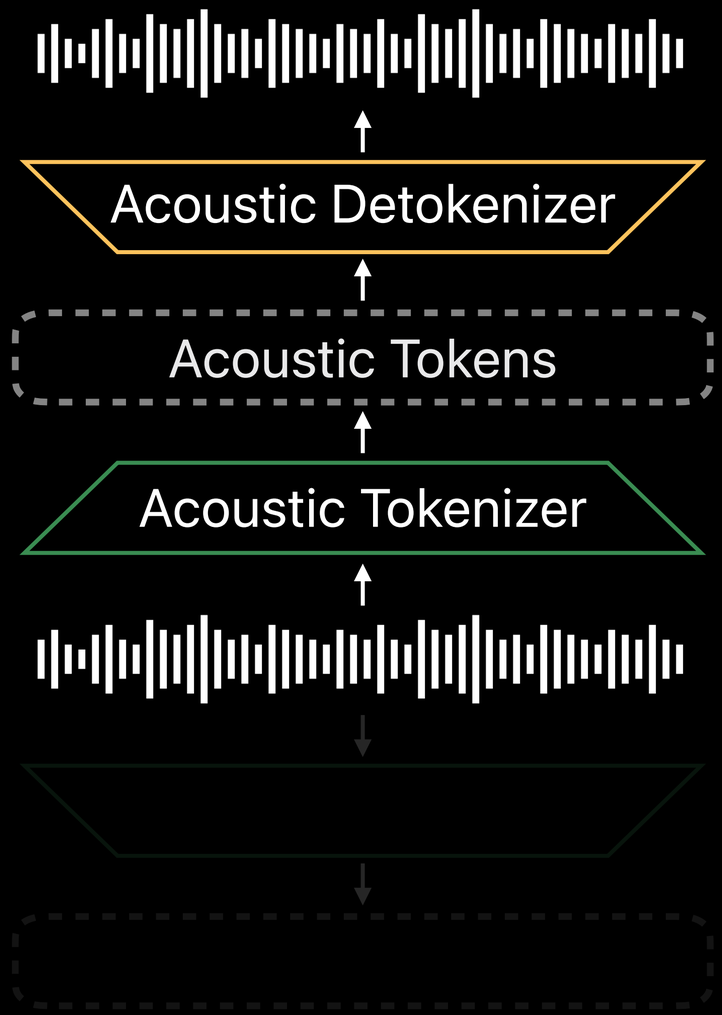
Popular generative models

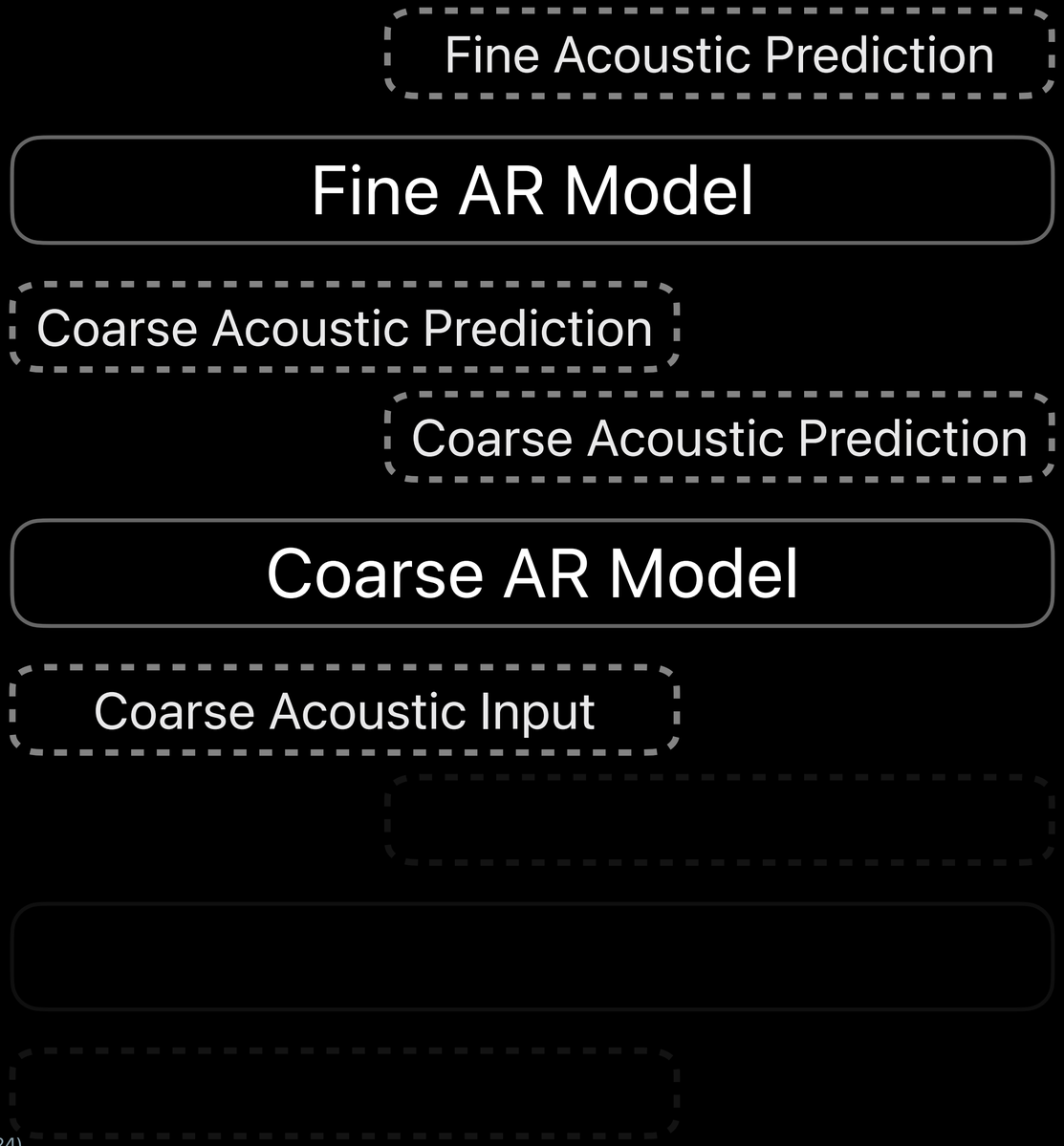




Borsos, Zalán et al. “AudioLM: A Language Modeling Approach to Audio Generation.” IEEE/ACM Transactions on Audio, Speech, and Language Processing 31 (2022): 2523-2533.

Popular generative models





Défossez, Alexandre, et al. "Moshi: a speech-text foundation model for real-time dialogue." *arXiv preprint arXiv:2410.00037* (2024).
Zhang, Xin, et al. "Speechtokenizer: Unified speech tokenizer for speech large language models." arXiv preprint arXiv:2308.16692 (2023).

Each of these models factorize data in different ways

Data

$$x_1^1 \cdots x_n^1 \cdots x_1^T \cdots x_n^T$$

T = sequence length

n = number of channels per time step

Audio LLM / Speech Tokenizer

$$p(x_1^1 \cdots x_n^1 \cdots x_1^T \cdots x_n^T) = \left\{ \prod_{t=1}^{T} p(x_1^t \mid x_1^{<t}) \right\} \left\{ \prod_{i=2}^{n} \prod_{t=1}^{T} p(x_i^t \mid x_{<i}^{1 \cdots T}) \right\}$$

Moshi

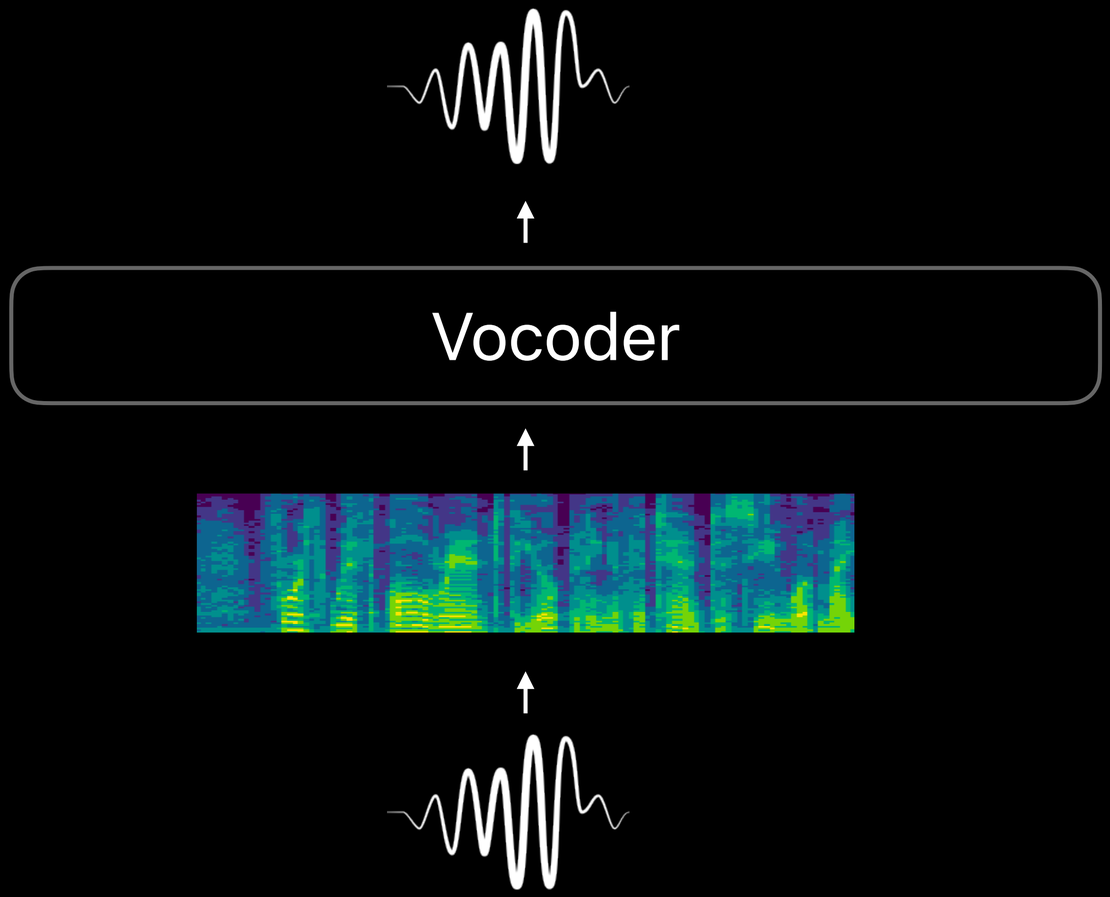
$$p(x_1^1 \cdots x_n^1 \cdots x_1^T \cdots x_n^T) = \prod_{t=1}^{T} \prod_{i=1}^{n} p(x_i^t \mid x_{<i}^t, x_\cdot^{<t})$$

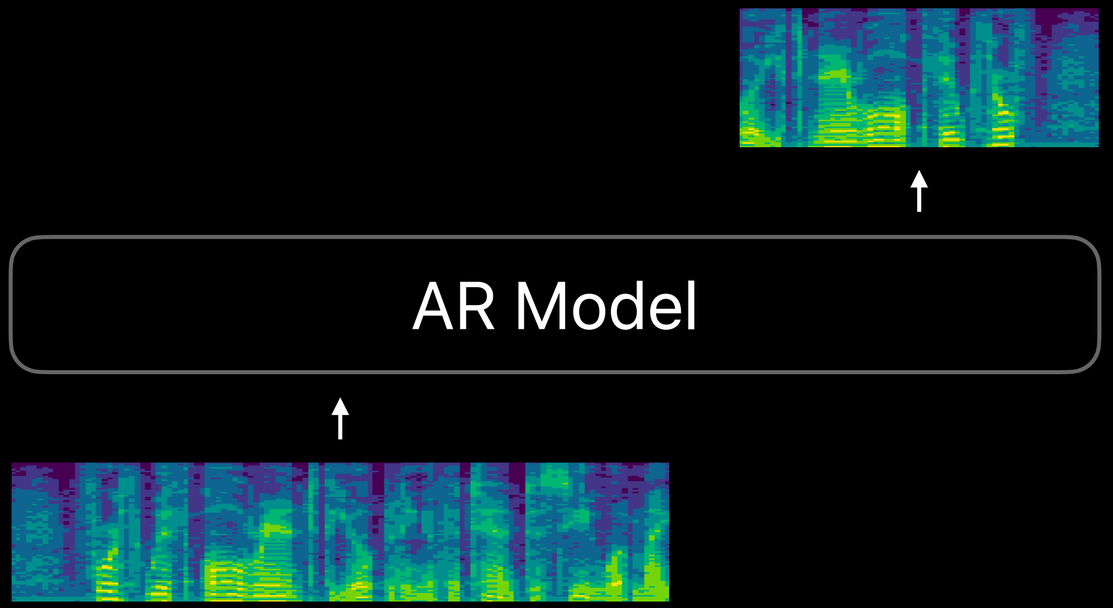
Our solution

Model the data directly

- Mel-spectral data is quite close to raw speech
  - Off the shelf mel-spectra → raw speech vocoders are available
- However continuous nature of mel-spectra might be challenging
  - Meng, L., et.al. Autoregressive speech synthesis without vector quantization. *arXiv preprint arXiv:2407.08551, 2024*
- Discretize mel-spectra for modeling
- Recent work shows it might be possible to model continuous mel-spectra some variational techniques
  - Meng, L., et.al. Autoregressive speech synthesis without vector quantization. *arXiv preprint arXiv:2407.08551, 2024*

Our solution: discretized log mel-filerbanks (dMel)





Predicting channels in parallel

Prior approaches serialize feature dimensions to be able to apply autoregressive loss

We predict each channel independently per step, and each step autoregressively

$$p(x_1^1 \cdots x_n^1 \cdots x_1^T \cdots x_n^T) = \prod_{t=1}^{T} \prod_{i=1}^{n} p(x_i^t \mid x_{1 \ldots n}^{<t})$$

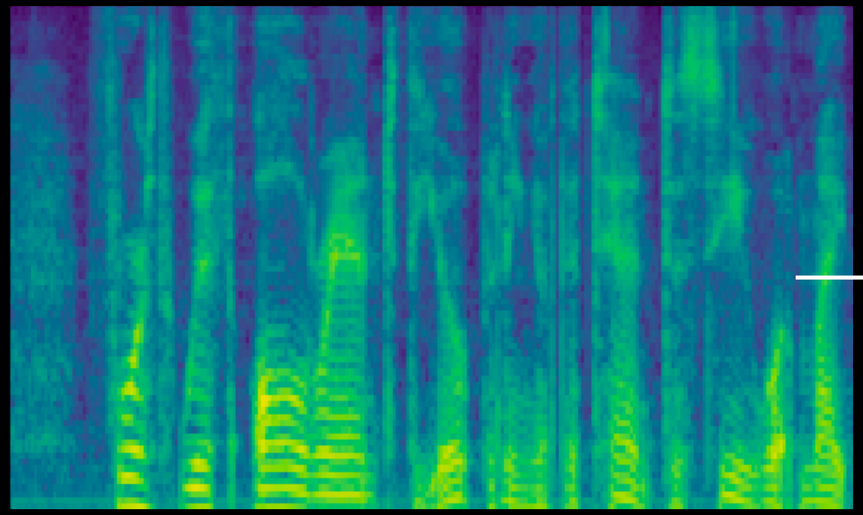
dMel: Trivial Tokenization

Bai, H. and et al. *dMel: Speech Tokenization made Simple*
https://arxiv.org/abs/2407.15835

Richard
Bai

Trivial tokenization (dMel)

Convert each Mel-filter bank continuous value to discrete bin based on range (min, max) and # of bins, N

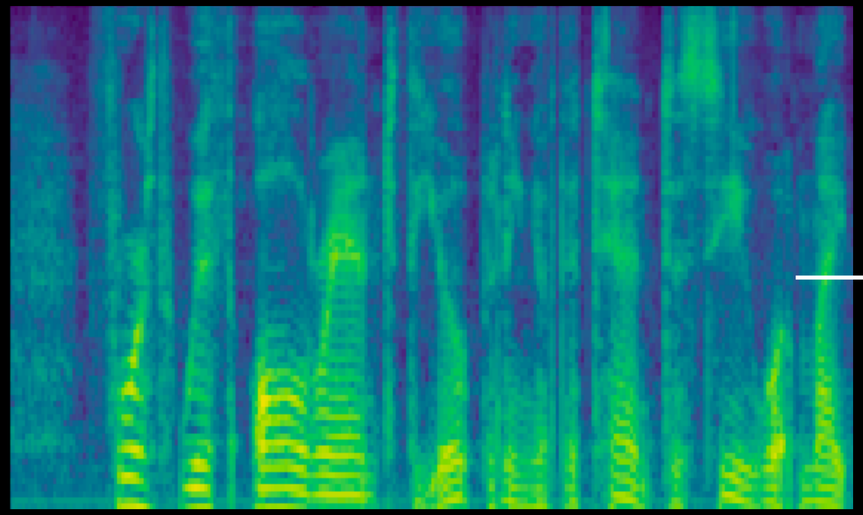


tokenize

$$i = \text{floor}\left[\frac{N \cdot (f - min)}{max - min}\right]$$

Trivial detokenization of dMel

Map each index back to (min, max) range in continuous space

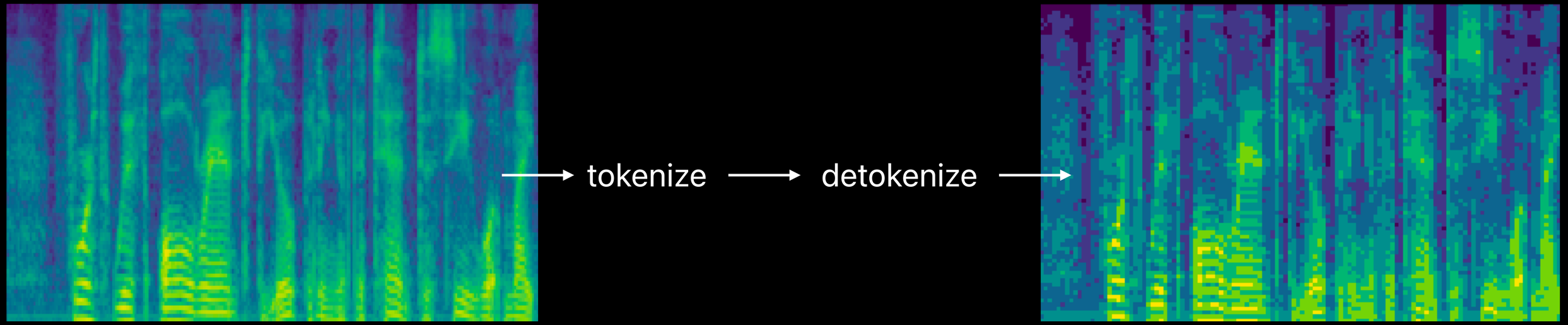


→ tokenize → detokenize

$$f' = \frac{i \cdot (max - min)}{N} + min$$

Round trip of tokenization / detokenization with dMel

Some round trip loss of resolution (like quantization)



Why would you do something so trivial ?

Mel spectral are grounded in the human perceptual bias

- Can be computed from any raw wave (even 8K telephony recordings)
- Do not depend on an encoder

Can be converted back to raw waveforms with any off the shelf Mel-spectral vocoder

- Can we reconstruct back the original speech with good enough fidelity ?

Quantized mel spectrograms preserve content

Measure Word Error Rate on reconstructed speech using a good ASR model

	Encoder Size	Decoder Size	Frame Rate (Hz)	WER
GroundTruth				2.02
HuBERT+KMeans	95M	111M	50	8.71
EnCodec	7M	7M	75	2.03
SpeechTokenizer	65M	34M	50	2.41
Mel + HifiGAN Vocoder	n/a	12M	80	**2.08**
dMel + HifiGAN Vocoder	n/a	12M	80	**2.11**

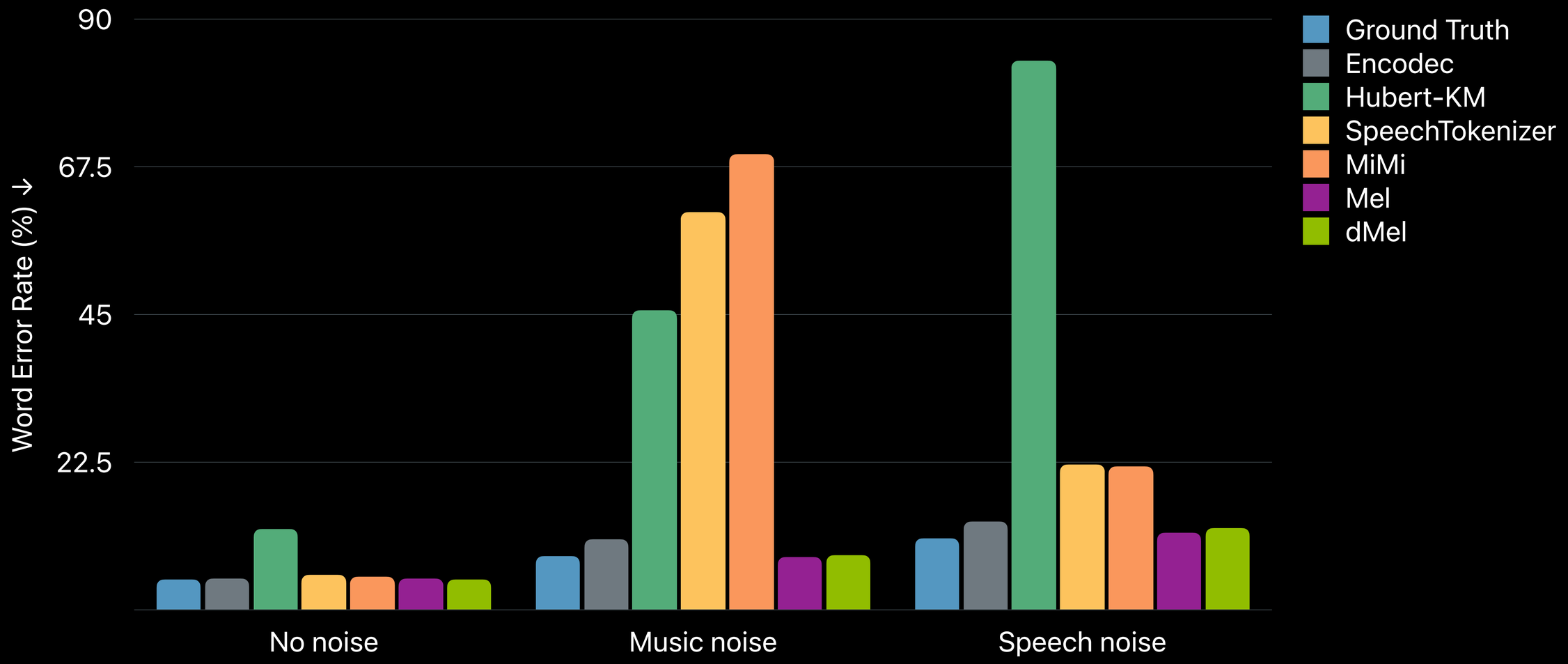
Quantized mel spectrograms preserve style / quality

Mean Opinion Score (MOS) of reconstructed speech assessed by raters

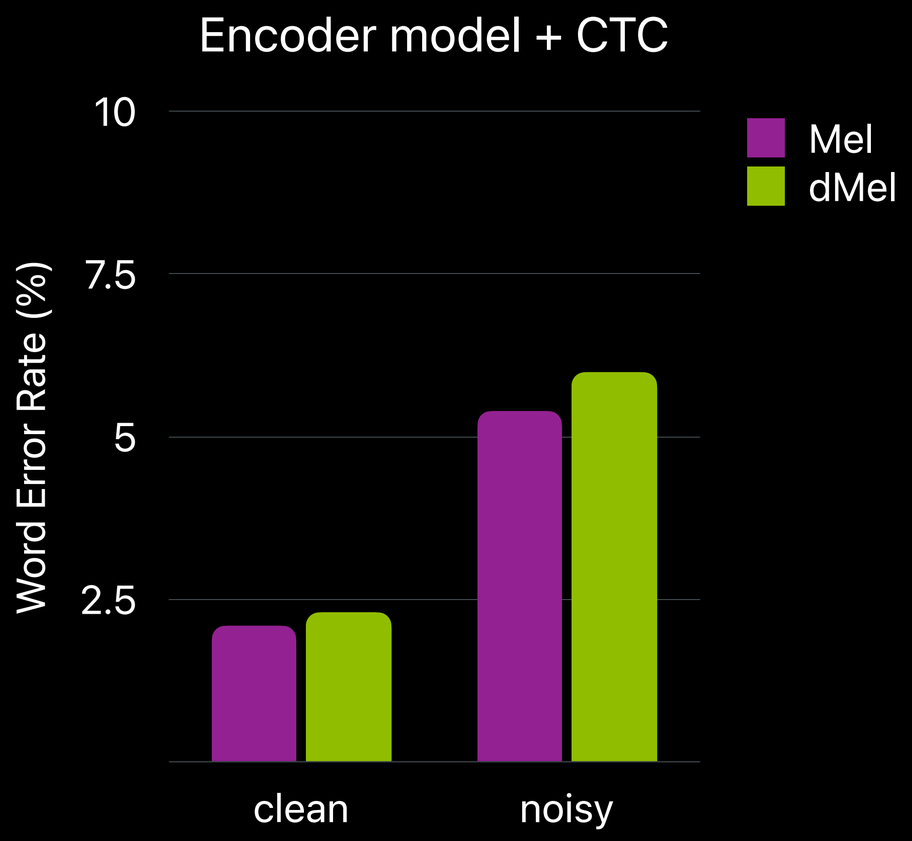
	Encoder Size	Decoder Size	Frame Rate (Hz)	WER	MOS
GroundTruth				2.02	3.91
HuBERT+KMeans	95M	111M	50	8.71	2.78
EnCodec	7M	7M	75	2.03	3.77
SpeechTokenizer	65M	34M	50	2.41	3.89
Mel + HifiGAN Vocoder	n/a	12M	80	**2.08**	**3.85**
dMel + HifiGAN Vocoder	n/a	12M	80	**2.11**	**3.73**

Mel and dMel are robust representations

Measure Word Error Rate on reconstructed speech using a good ASR model

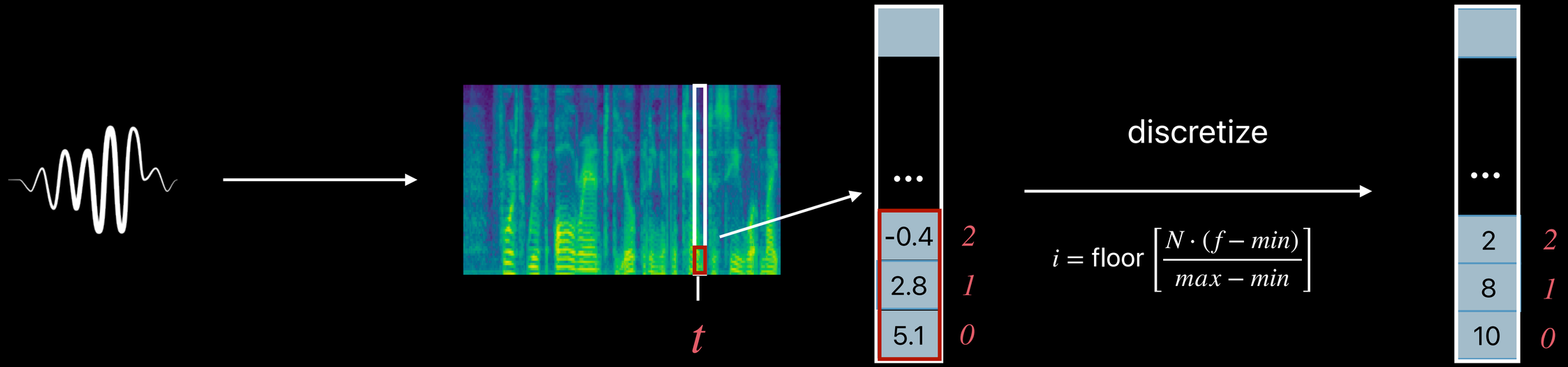


dMel is as good as Mel as inputs to conventional ASR models



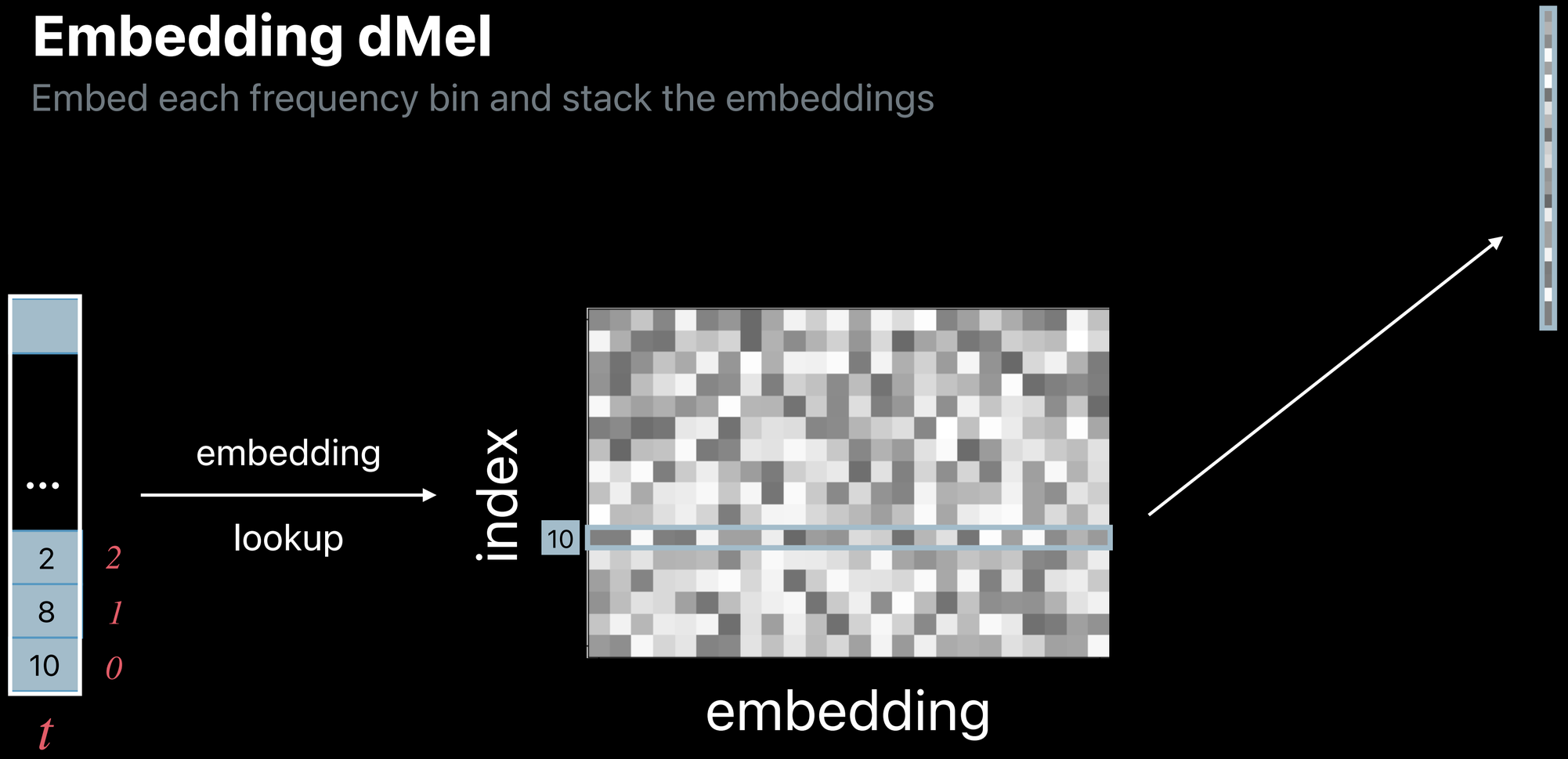
Embedding dMel

Discretize each frequency bin into indices 1-18 using minimum and maximum values



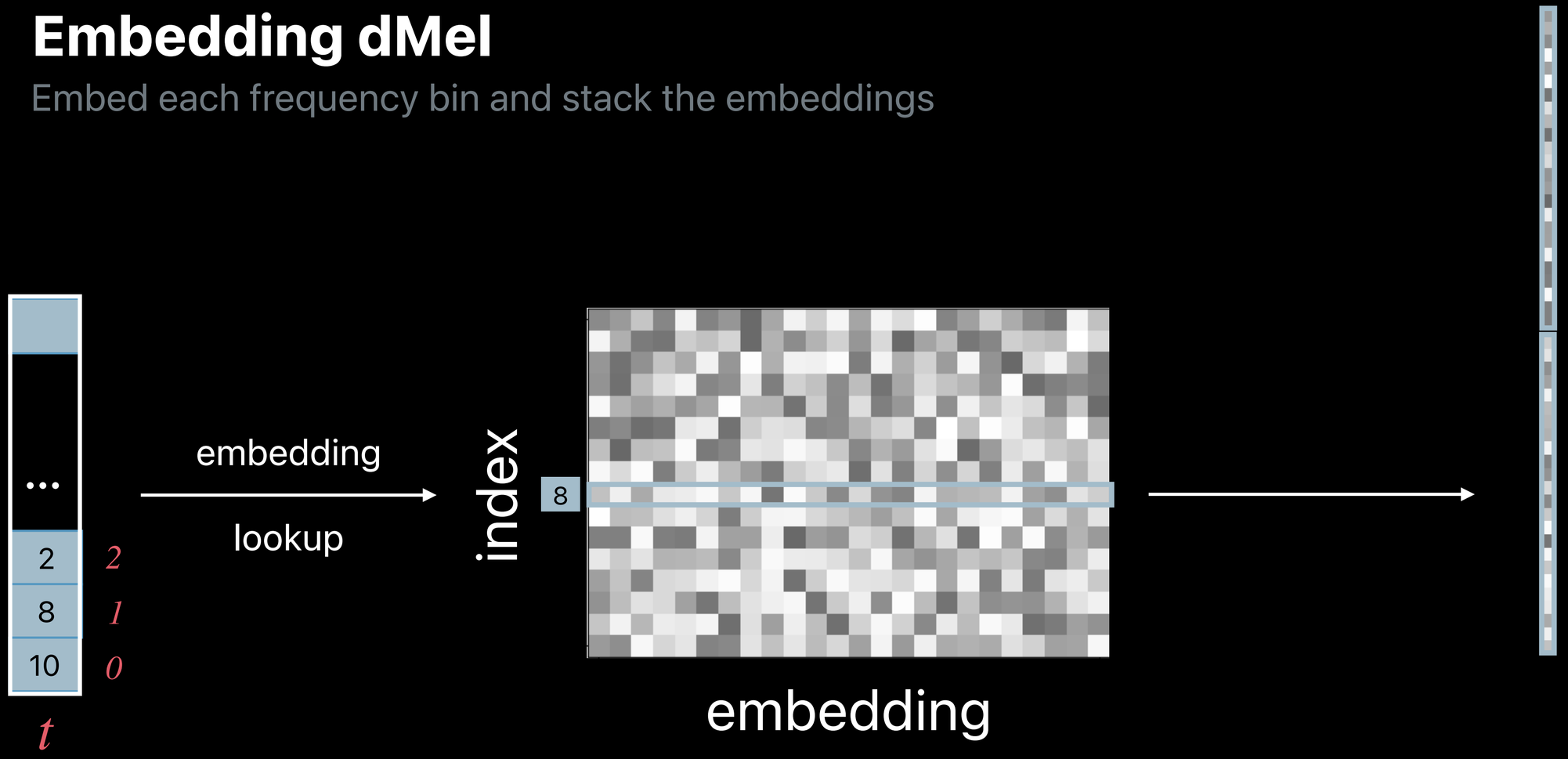
Embedding dMel

Embed each frequency bin and stack the embeddings



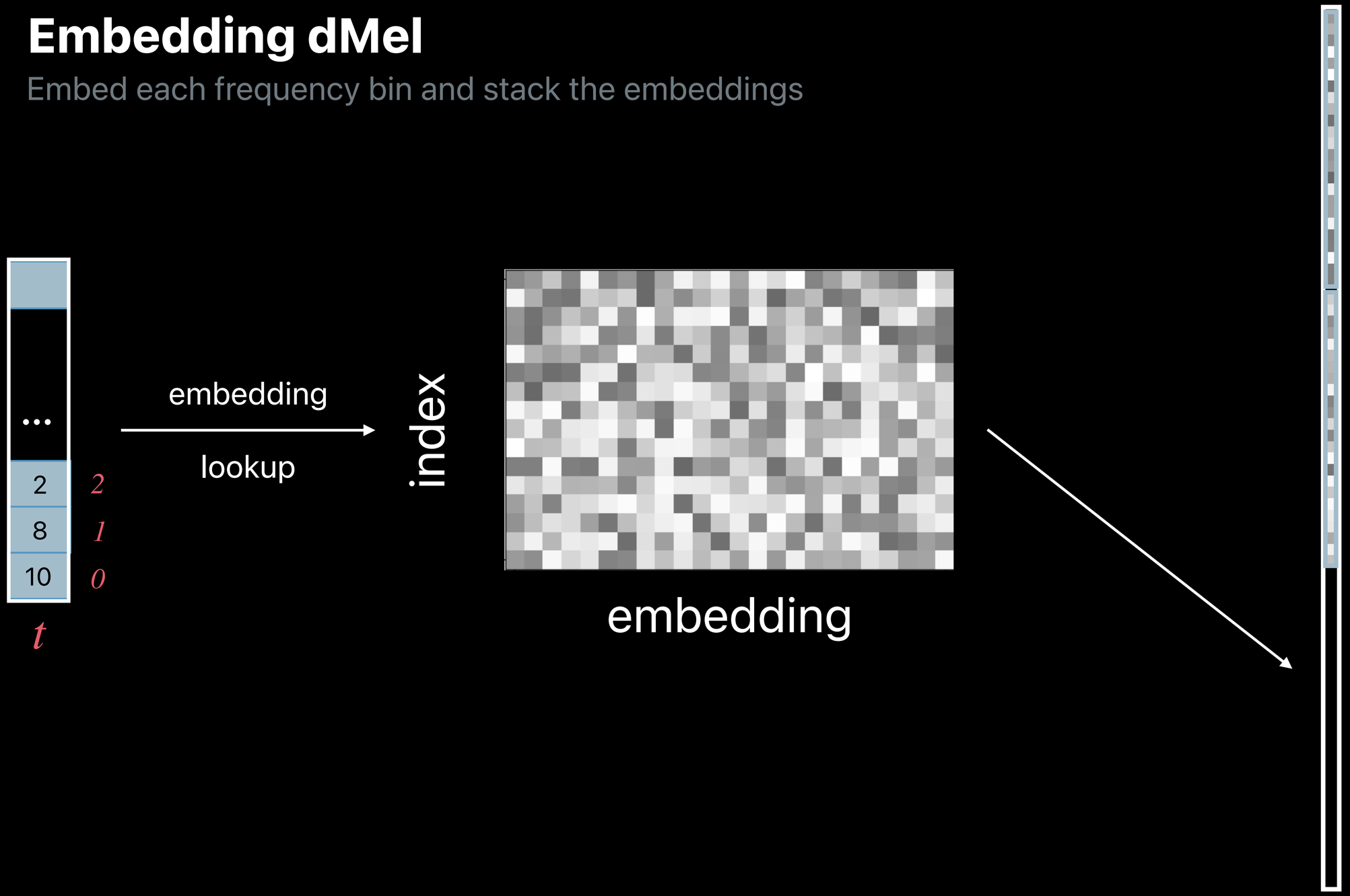
Embedding dMel

Embed each frequency bin and stack the embeddings



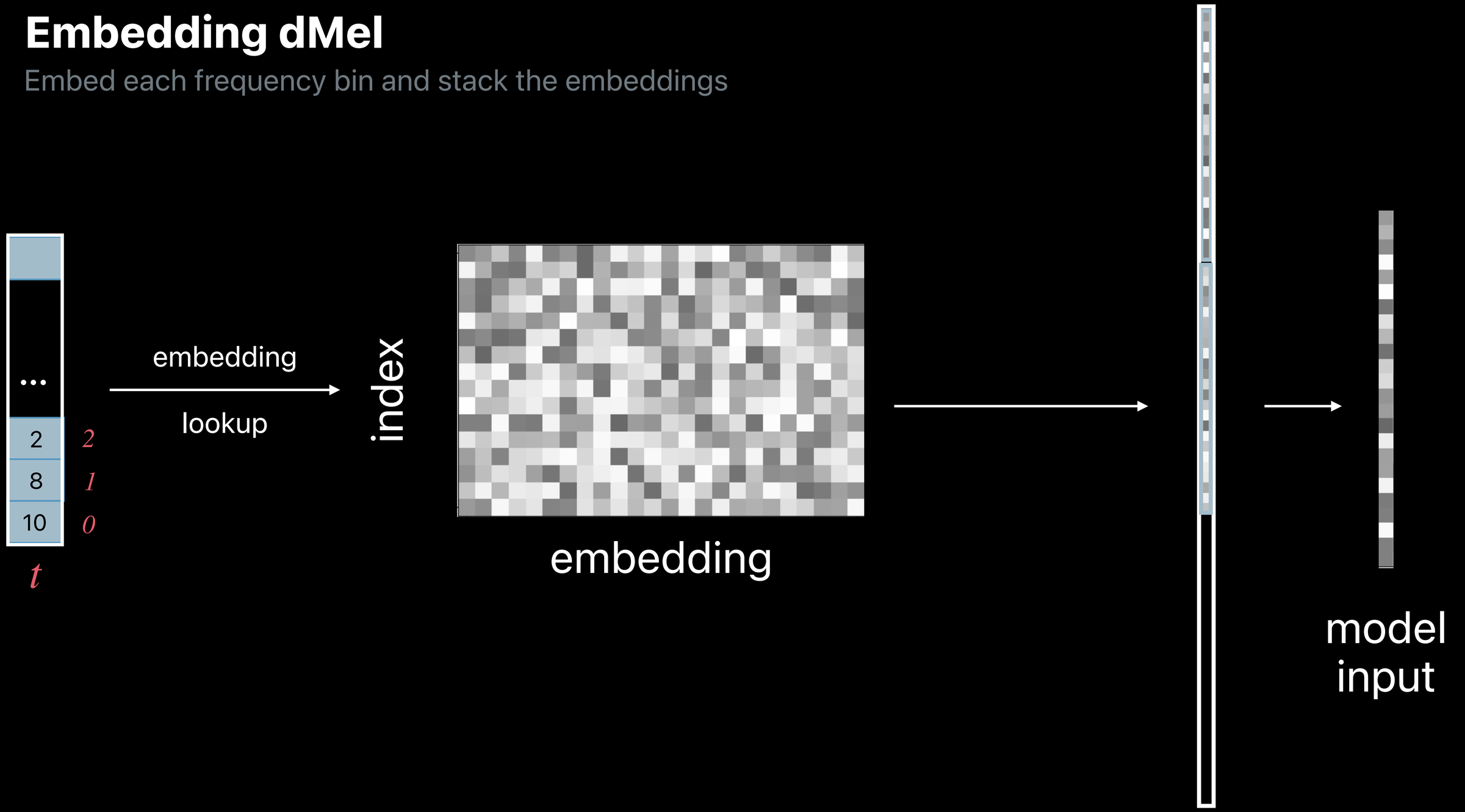
Embedding dMel

Embed each frequency bin and stack the embeddings



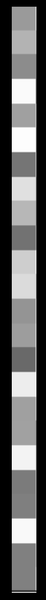
Embedding dMel

Embed each frequency bin and stack the embeddings



Predicting dMel

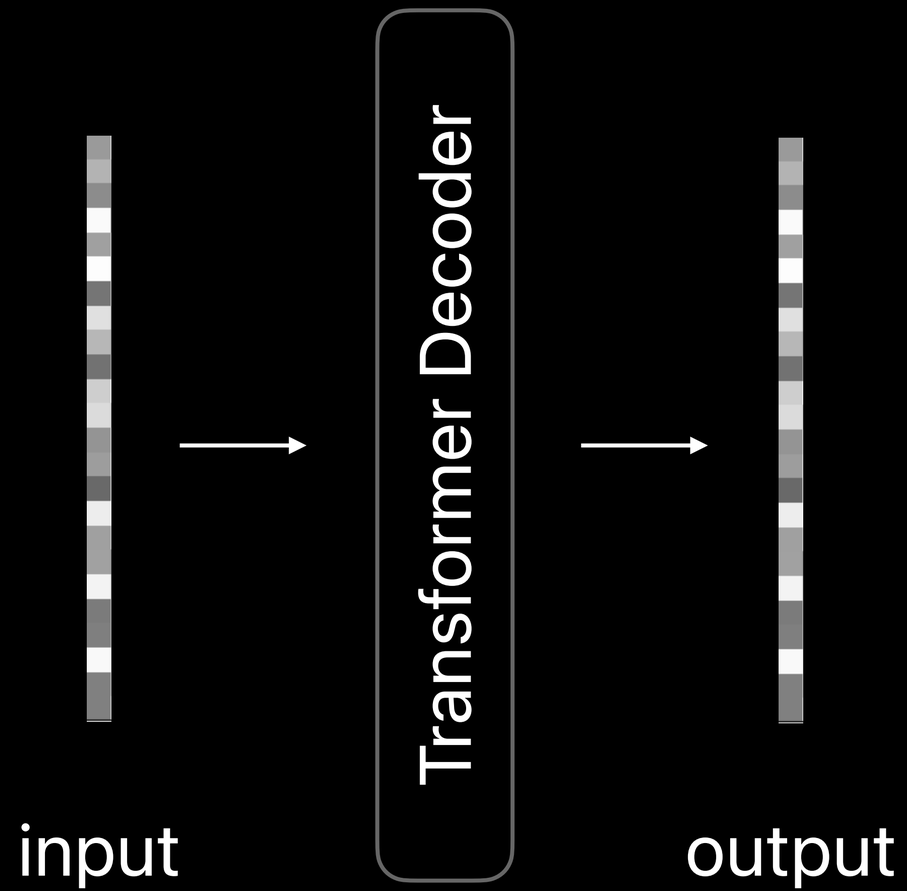
Predict all mel bins for a time step in parallel



input

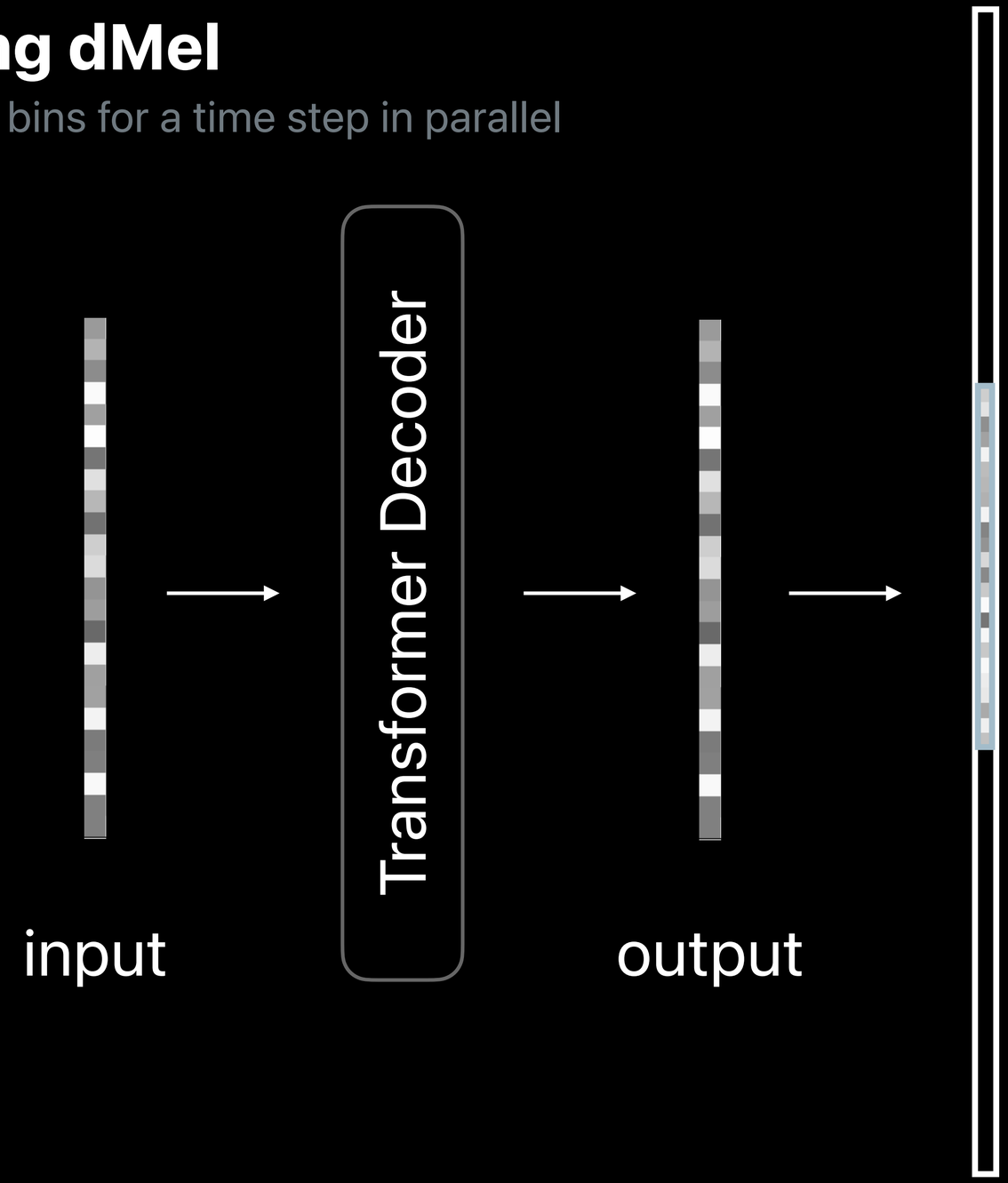
Predicting dMel

Predict all mel bins for a time step in parallel



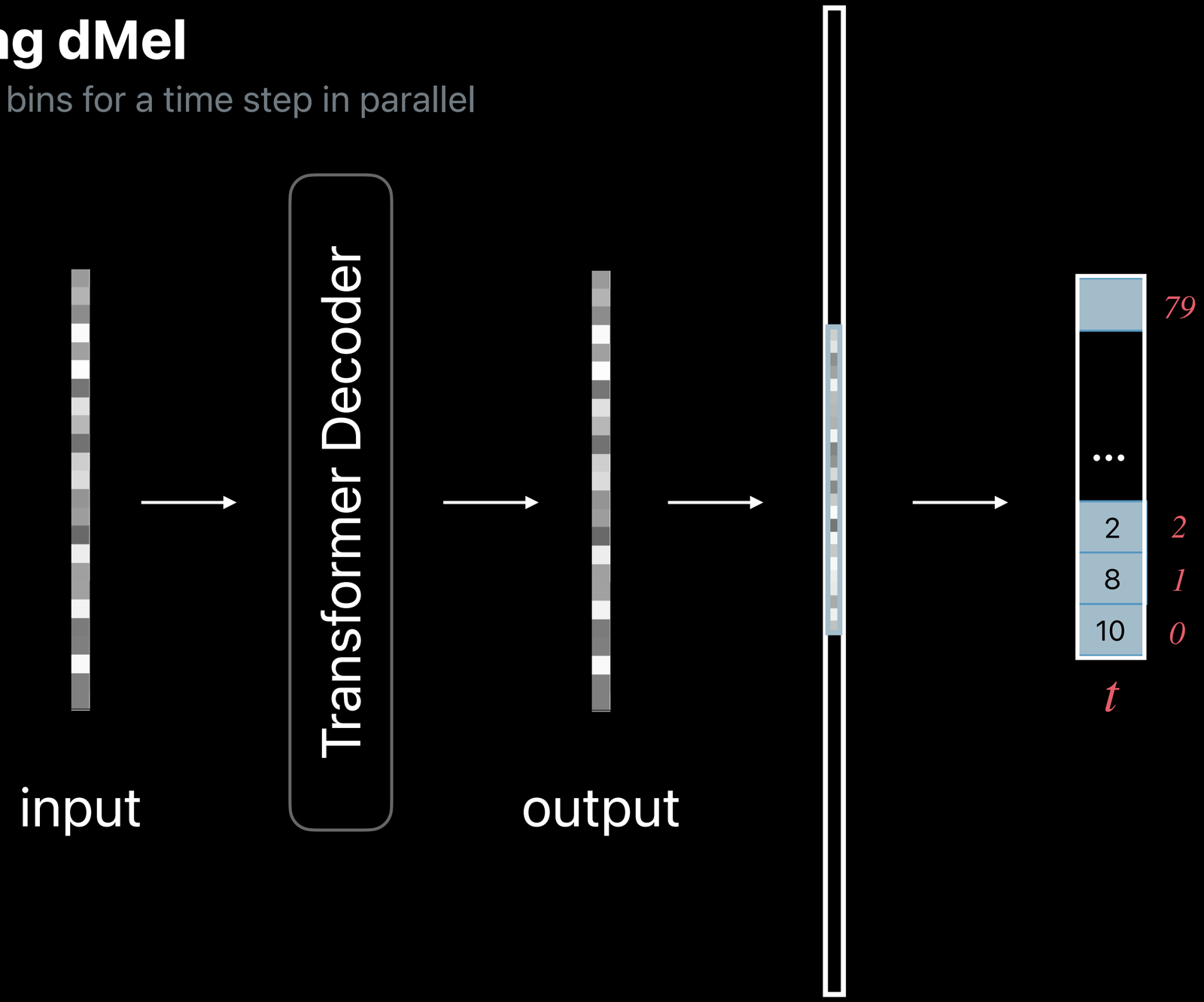
Predicting dMel

Predict all mel bins for a time step in parallel

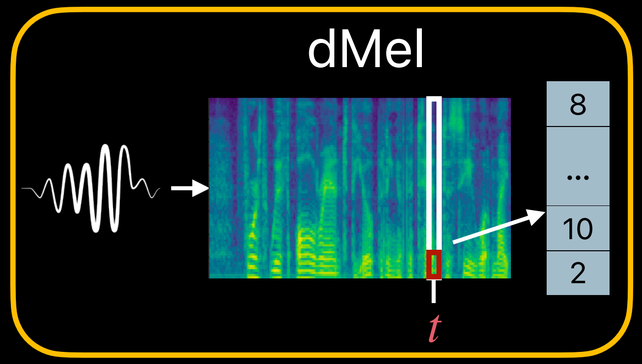


Predicting dMel

Predict all mel bins for a time step in parallel



RichTTS — decoder only text to speech model with dMel



Why should independently predicting channels work ?

Independent channels represent redundant predictions

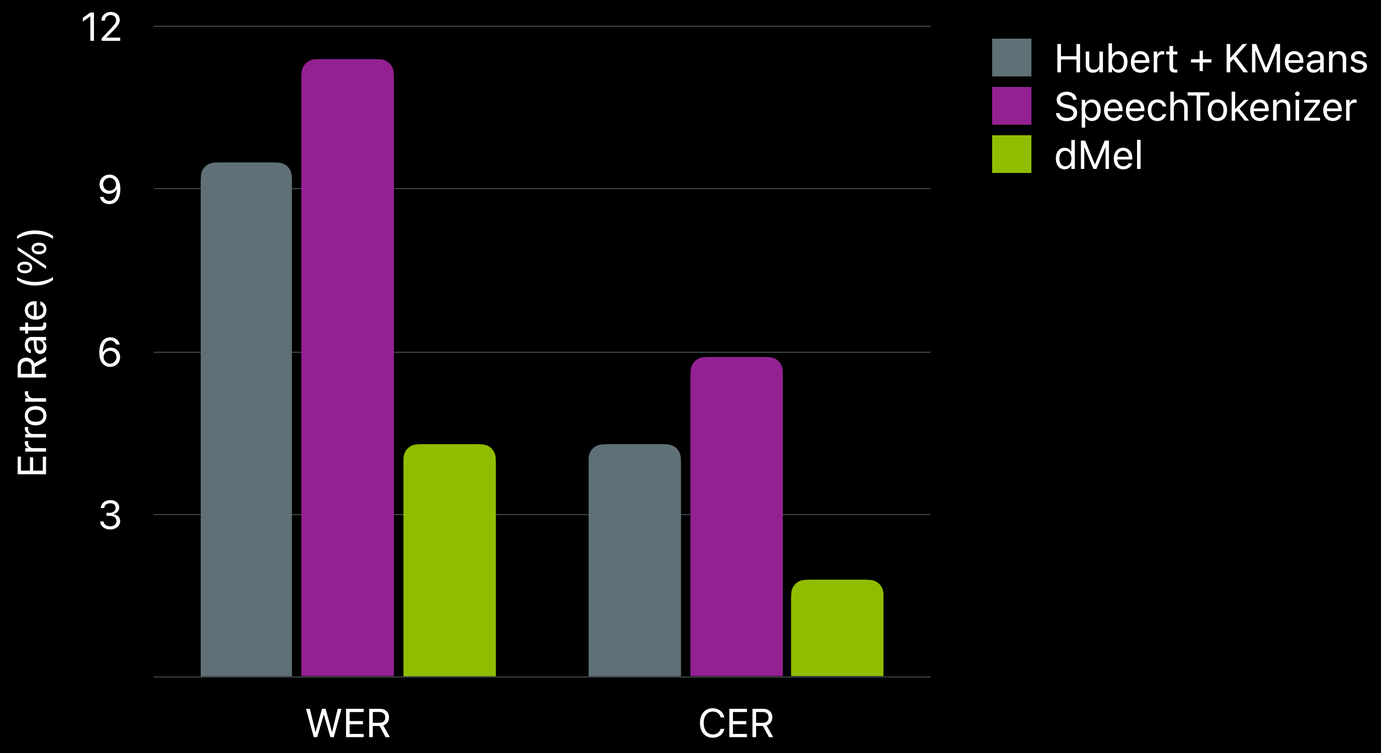
Mistakes in different channels at one time-step can be corrected by the transformer model in the next step by bringing the prediction back to the manifold

Easier when channels are less intertwined, as is in the case with dMel, compared to complicated tokenizations

RichTTS results with different tokenizers

Assess WER of synthetic data with whisper ASR model

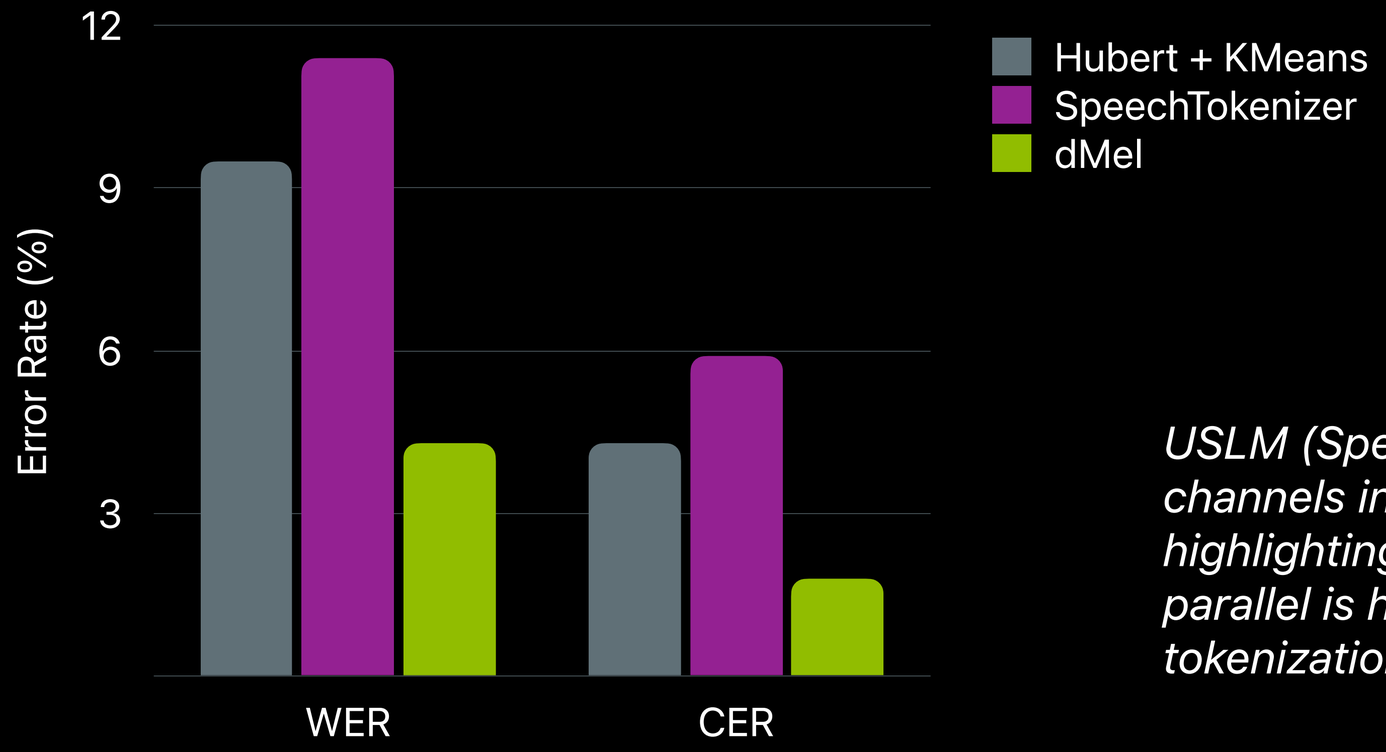
255M parameters; training on LibriSpeech



RichTTS results with different tokenizers

Assess WER of synthetic data with whisper ASR model

255M parameters; training on LibriSpeech



USLM (SpeechTokenizer) does not predict channels in parallel and achieves 6.5 highlighting that predicting channels in parallel is harder for more complicated tokenizations

What about bit-rates of the code ?

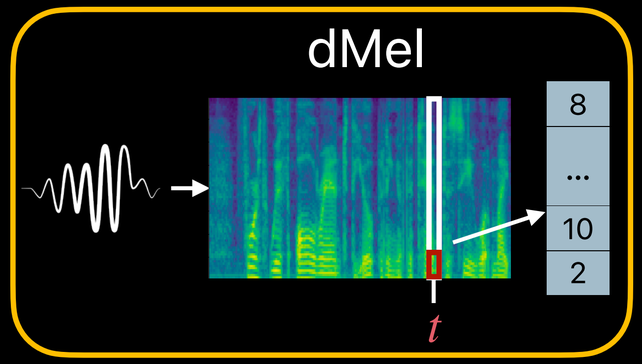
Higher bit-rate because of more codes

But much fewer time steps of modeling because all the channels are embedded together

- Transformer time step modeling is much more intensive than bit-rate of coding when it comes to efficiency

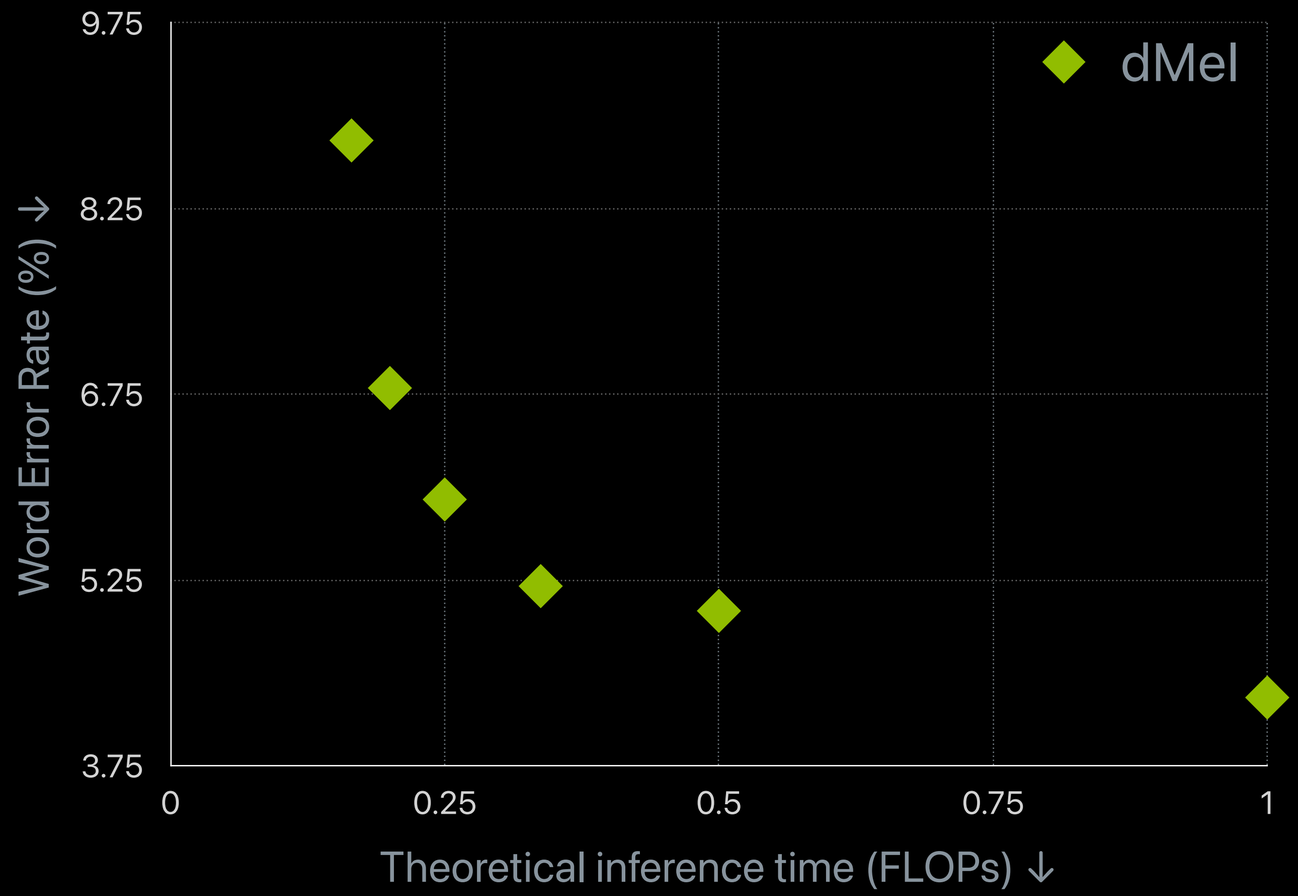
Can we can go much further, and reduce the number of time steps of modeling even more ?

Predict multiple frames in parallel



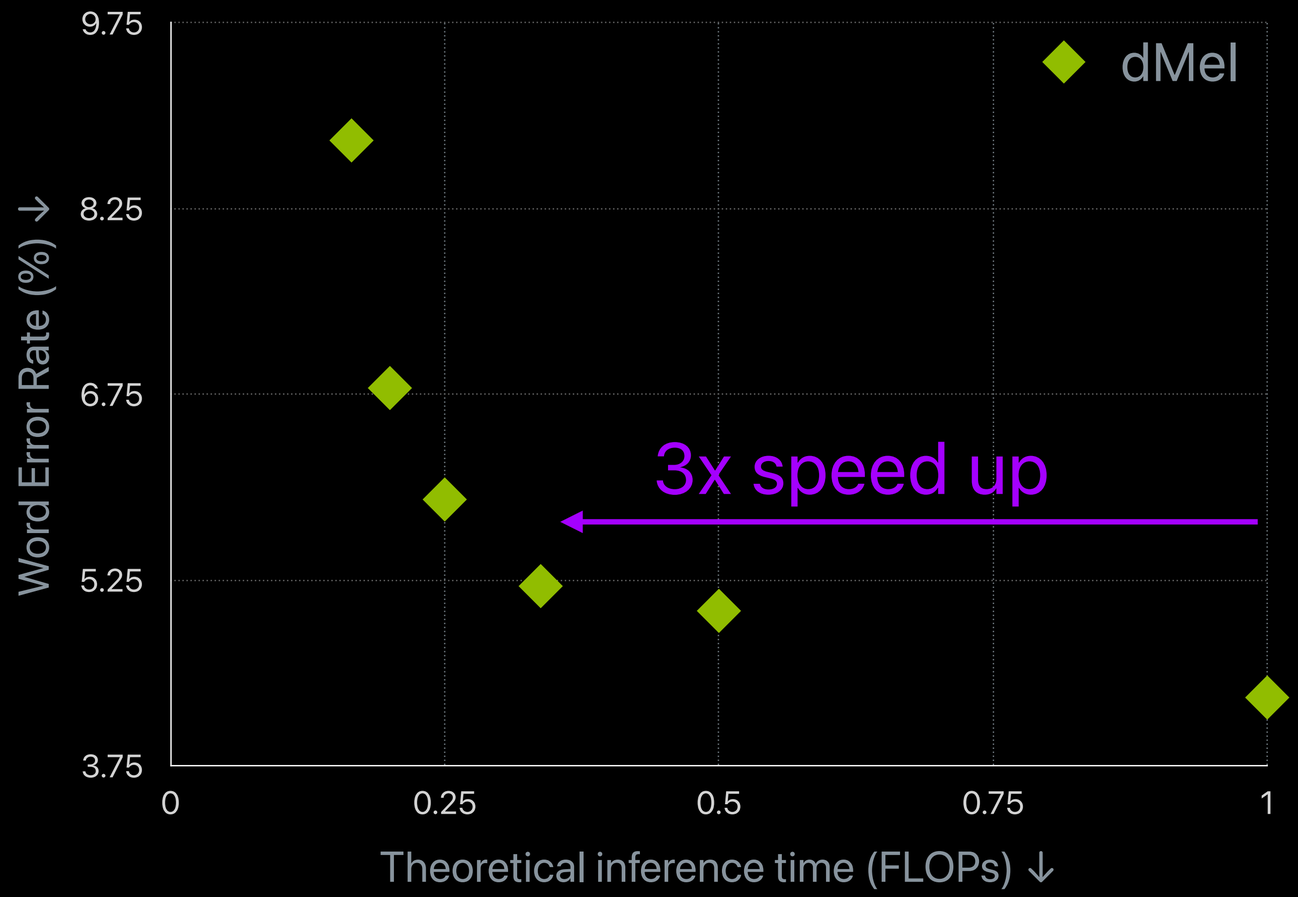
Predict multiple frames in parallel

Assess WER of synthetic data with whisper ASR model

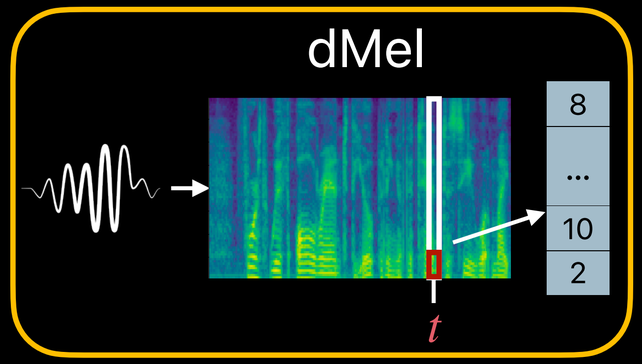


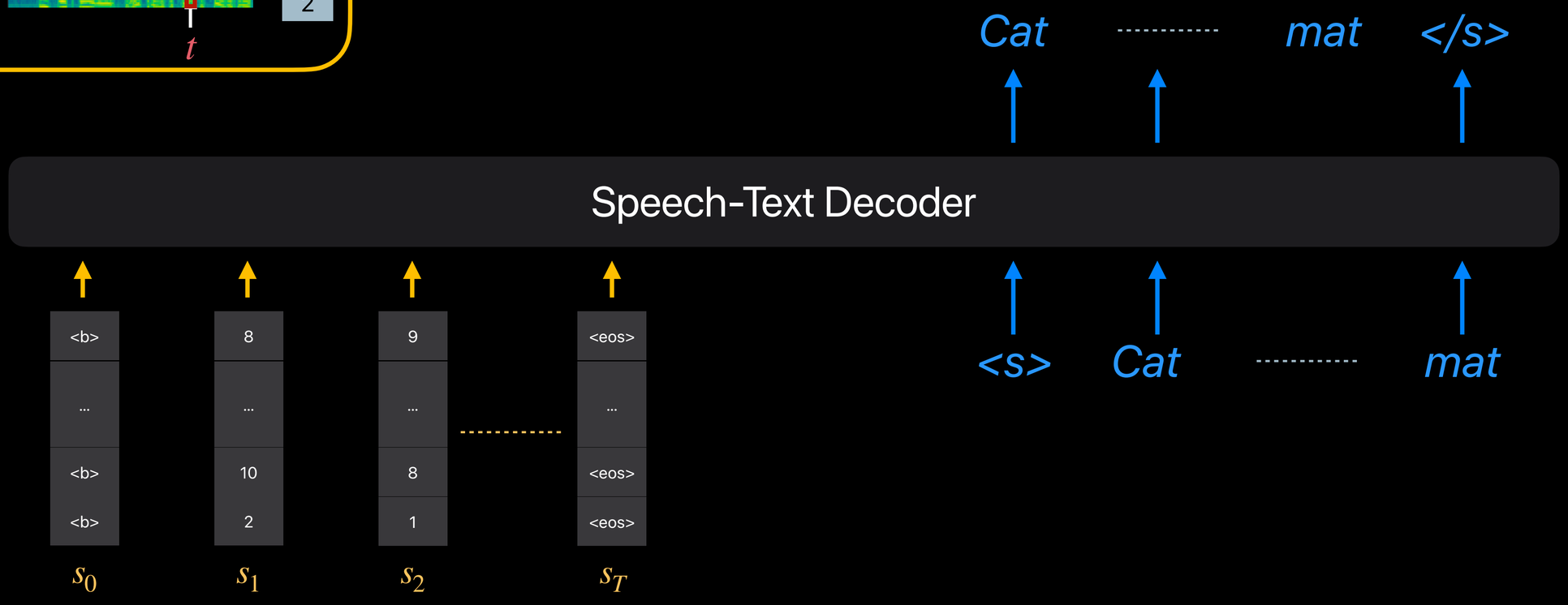
Predict multiple frames in parallel

Assess WER of synthetic data with whisper ASR model



RichASR — decoder only speech to text model with dMel

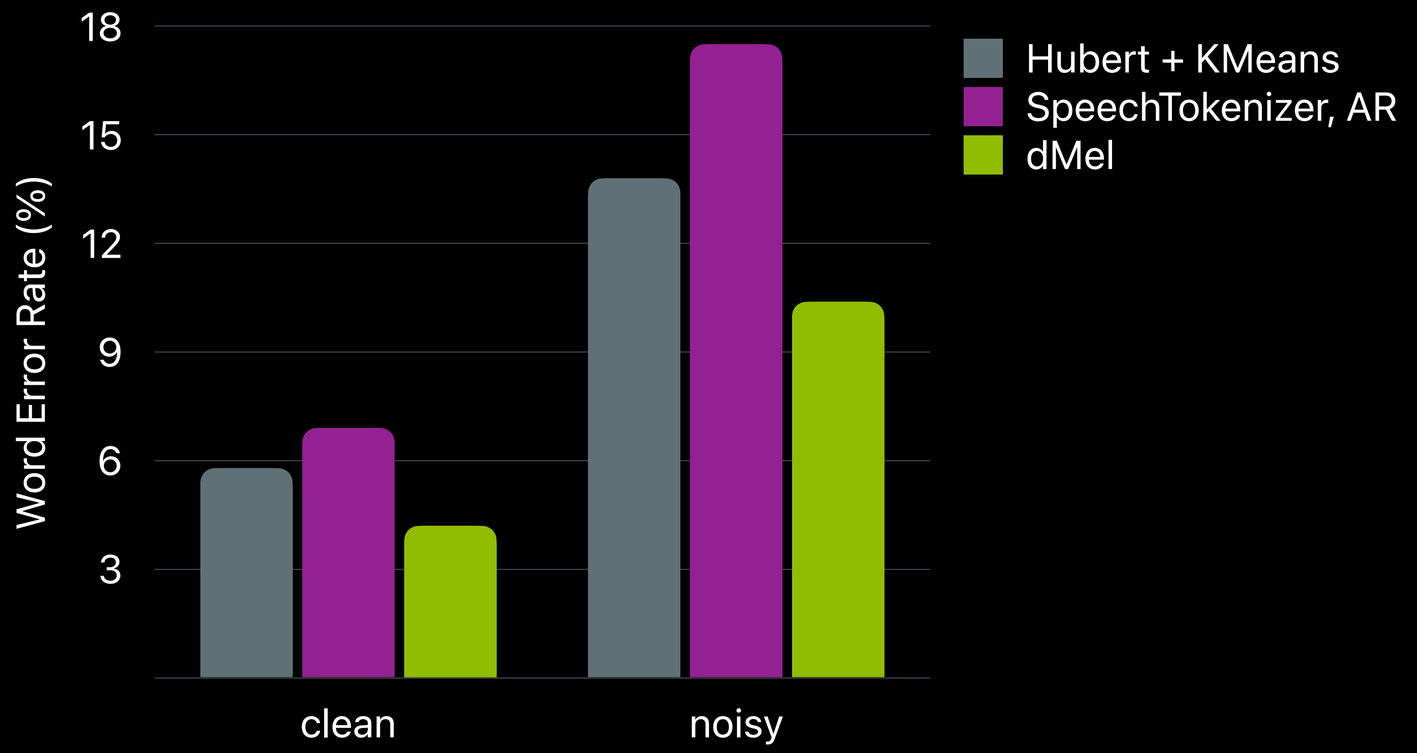




RichASR results with different tokenizers

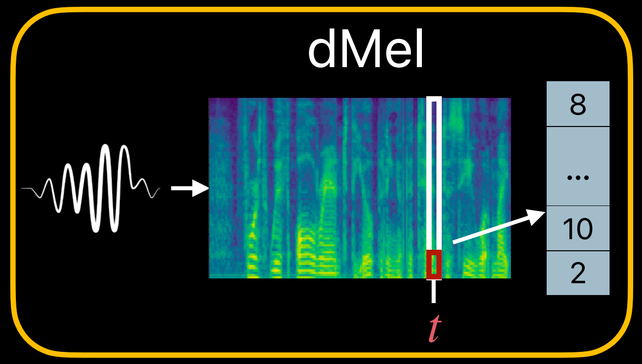
dMel is able to achieve good results, probably because it is able to preserve content

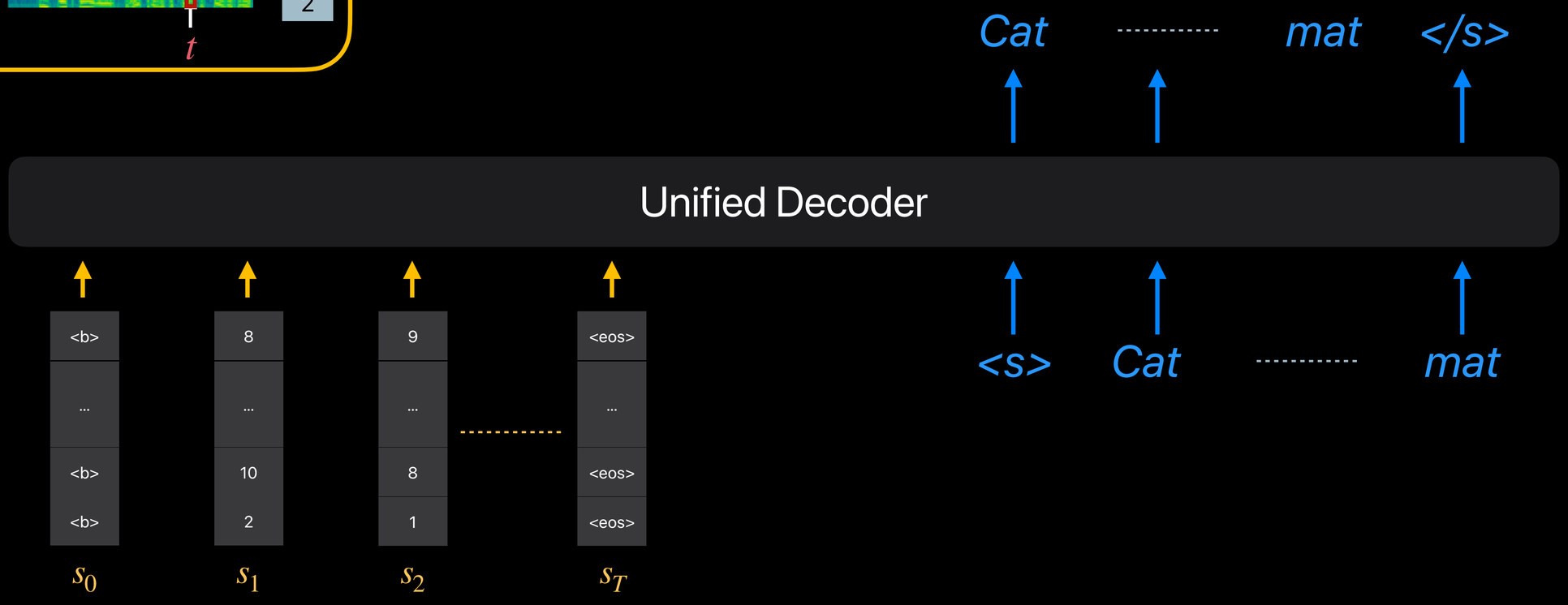
255M parameters; training on LibriSpeech



Joint decoder-only model

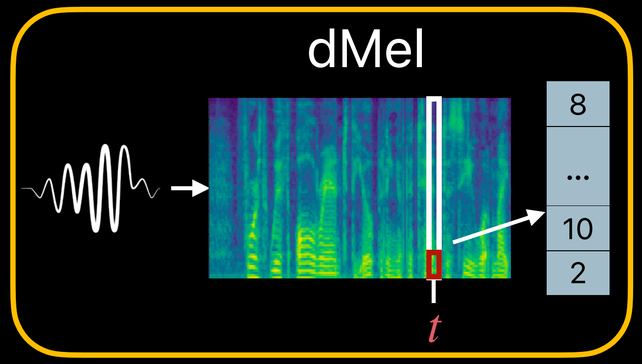
Speech → Text





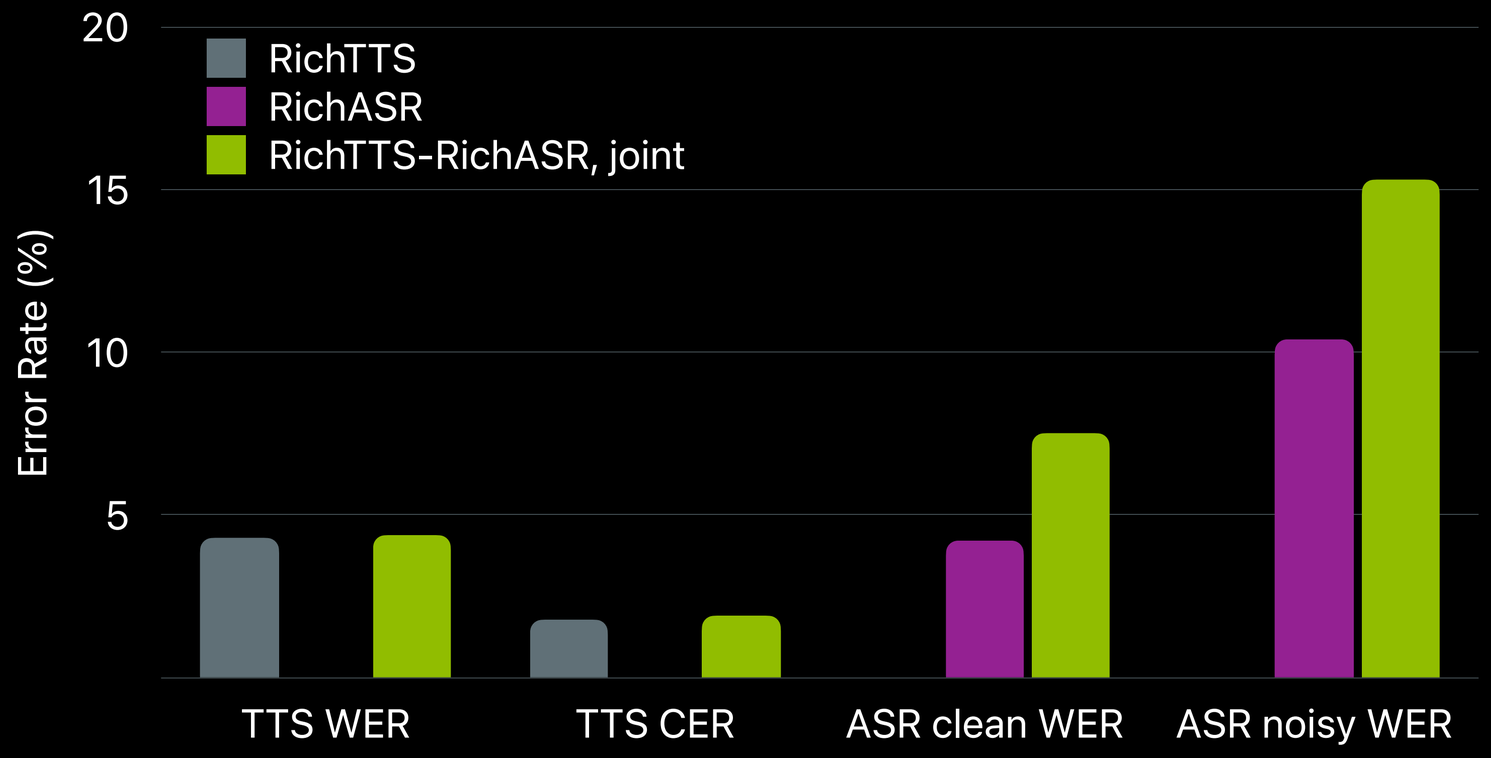
Joint decoder-only model

Text → Speech



Joint decoder-only model for both ASR and TTS

ASR seems to be under-trained



SpeakStream

Bai, H. and et al. *SpeakStream: Streaming Text-to-Speech with Interleaved Data*
https://arxiv.org/abs/2505.19206

Richard
Bai

Streaming TTS

Boom in systems to emulate voice-assistant interactions

- generate LLM responses to user questions
- synthesize the response to speech as soon as possible

Streaming TTS

Boom in systems to emulate voice-assistant interactions

- generate LLM responses to user questions
- synthesize the response to speech as soon as possible

Need for streaming TTS with low latency

- 30ms on M4 ✅

Streaming TTS

Boom in systems to emulate voice-assistant interactions

- generate LLM responses to user questions
- synthesize the response to speech as soon as possible

Need for streaming TTS with low latency

- 30ms on M4 ✅

Need for streaming Vocoder with low latency

- 15ms on M4 ✅

Streaming TTS

Boom in systems to emulate voice-assistant interactions

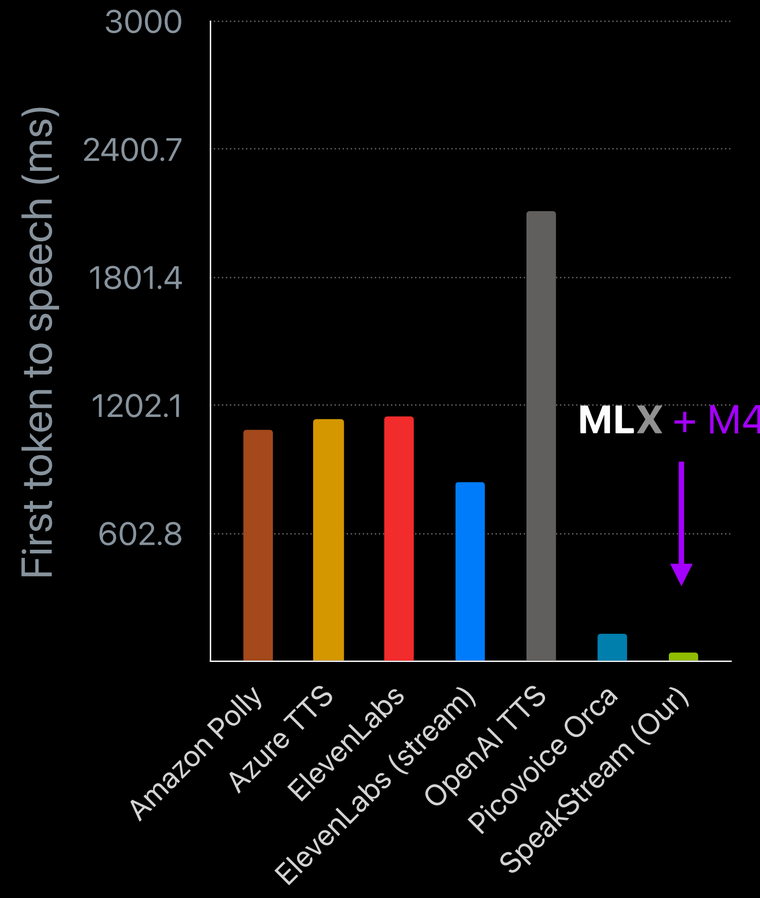
- generate LLM responses to user questions
- synthesize the response to speech as soon as possible

Need for streaming TTS with low latency

- 30ms on M4 ✅

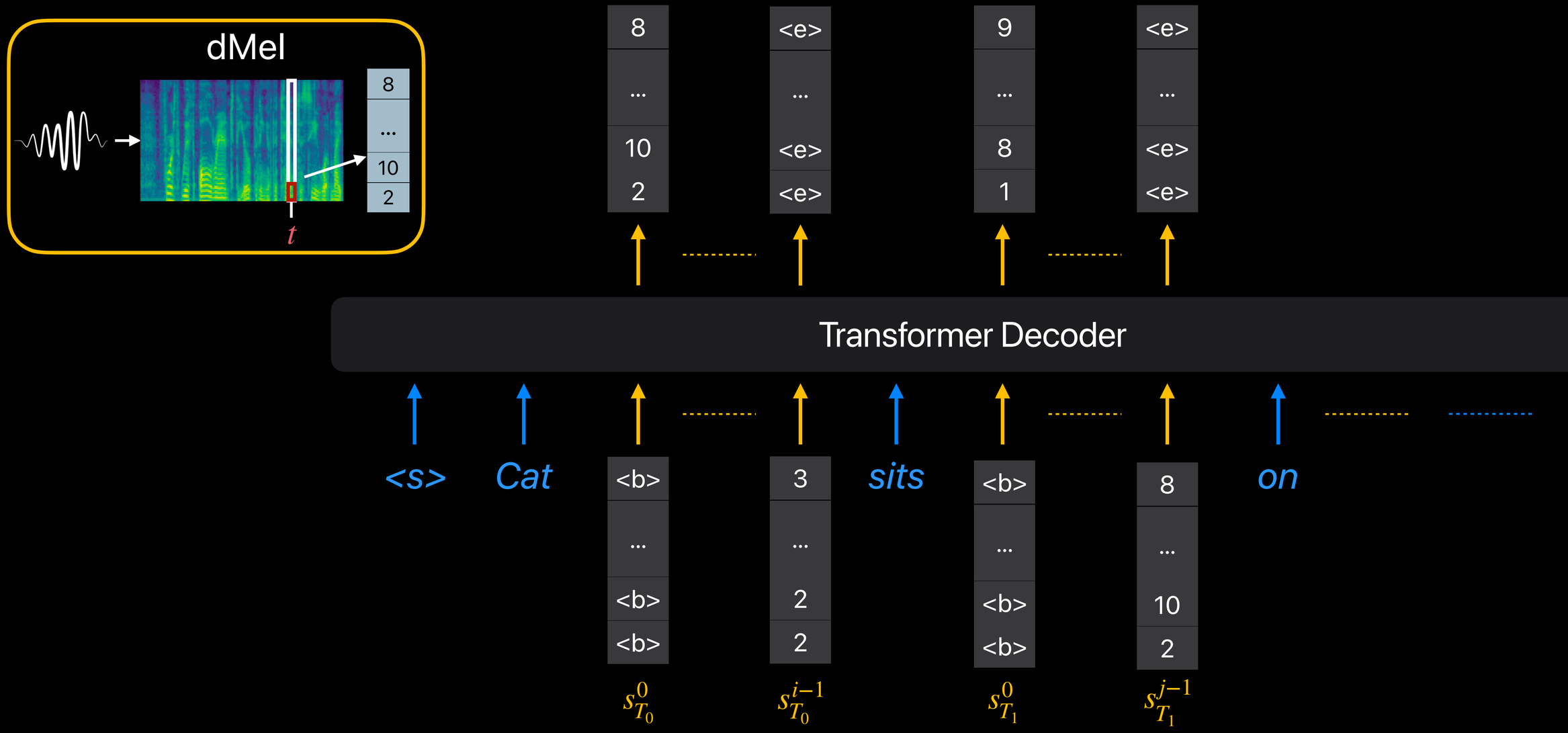
Need for streaming Vocoder with low latency

- 15ms on M4 ✅

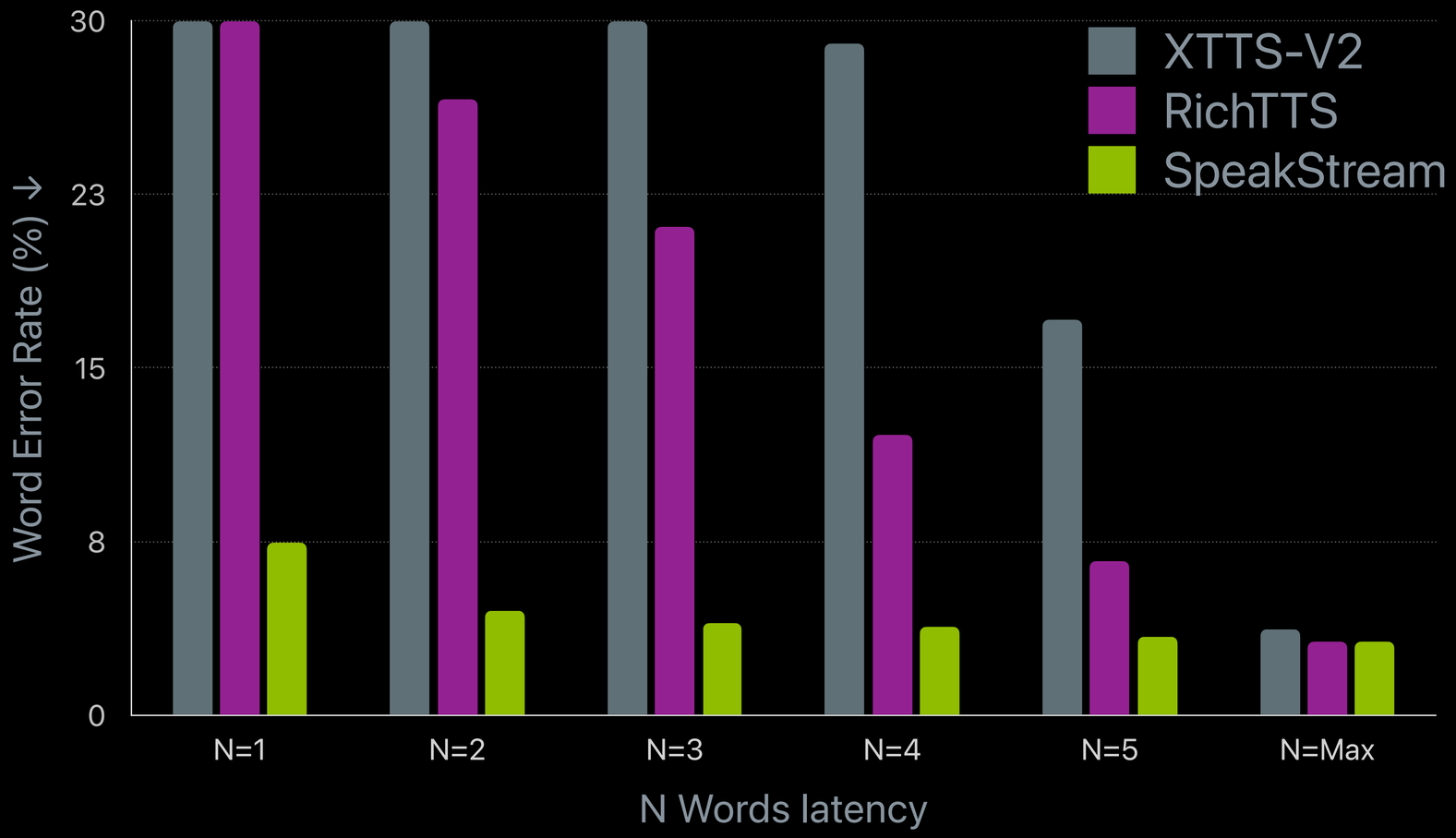


SpeakStream : LLM inspired streaming TTS on dMel

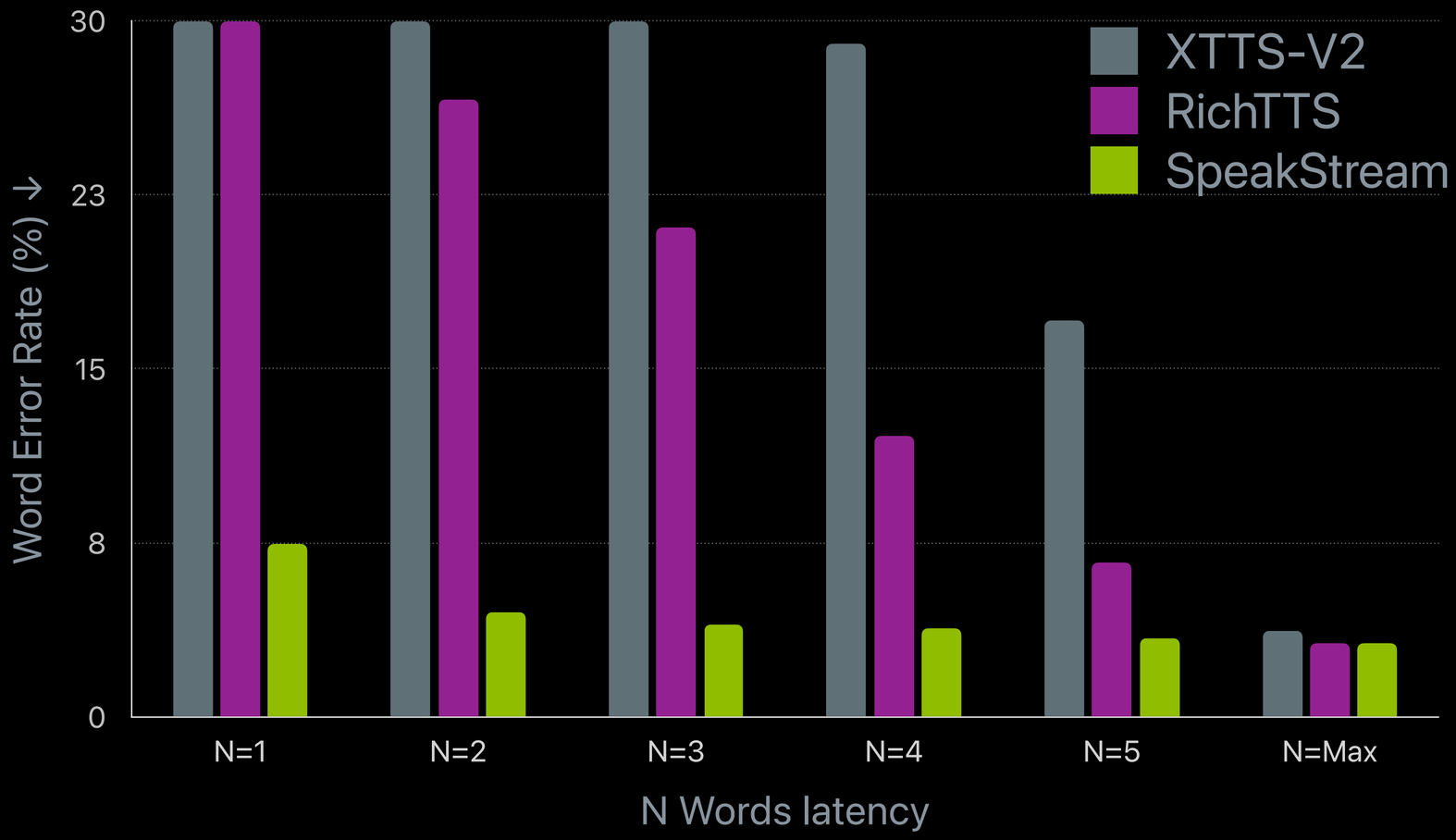
Interleave written *n*-gram with spoken *n*-gram



SpeakStream : LLM inspired streaming TTS on dMel

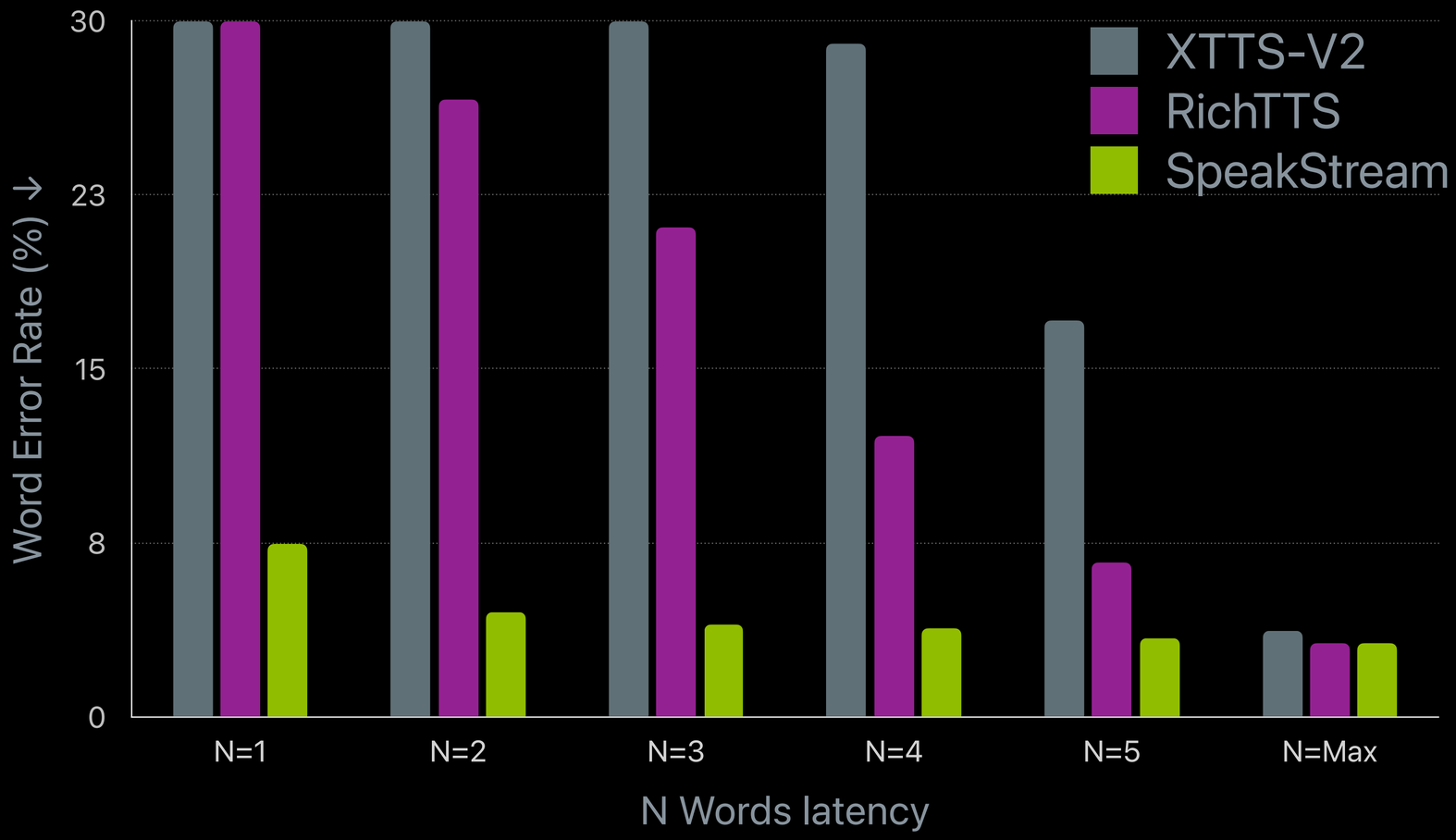


SpeakStream : LLM inspired streaming TTS on dMel



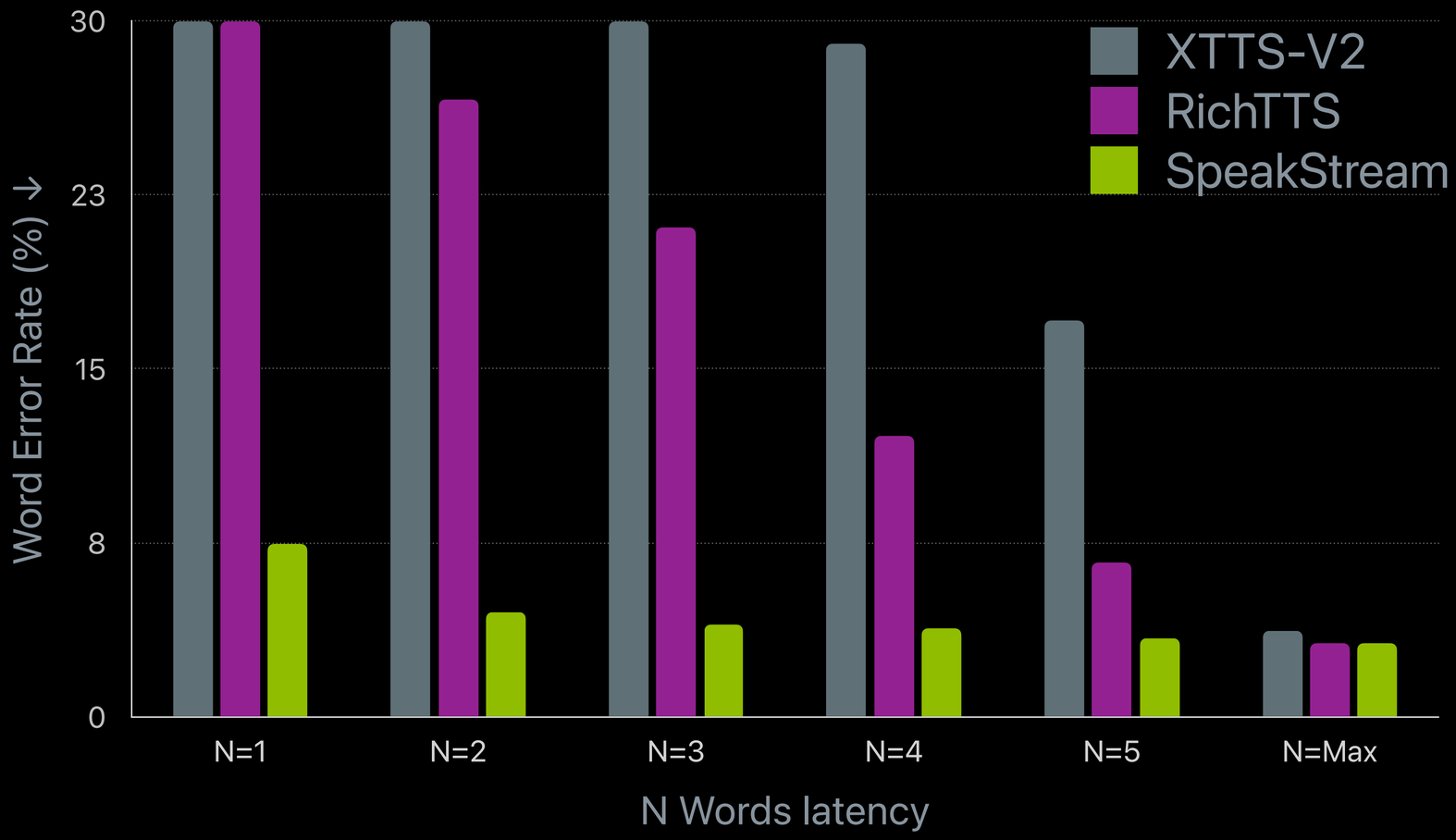
Latency : 45ms with **MLX** + M4

SpeakStream : LLM inspired streaming TTS on dMel



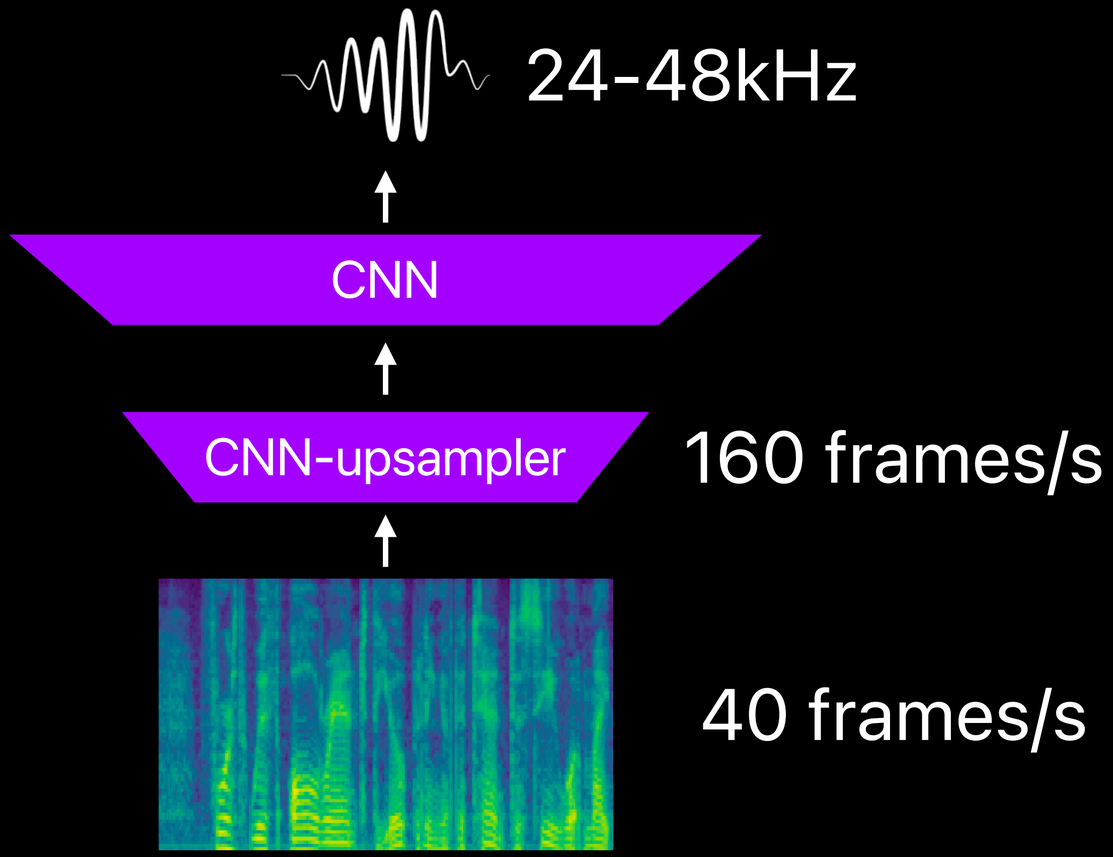
Latency : 45ms with **MLX** + M4

SpeakStream : LLM inspired streaming TTS on dMel



Latency : 45ms with **MLX** + M4

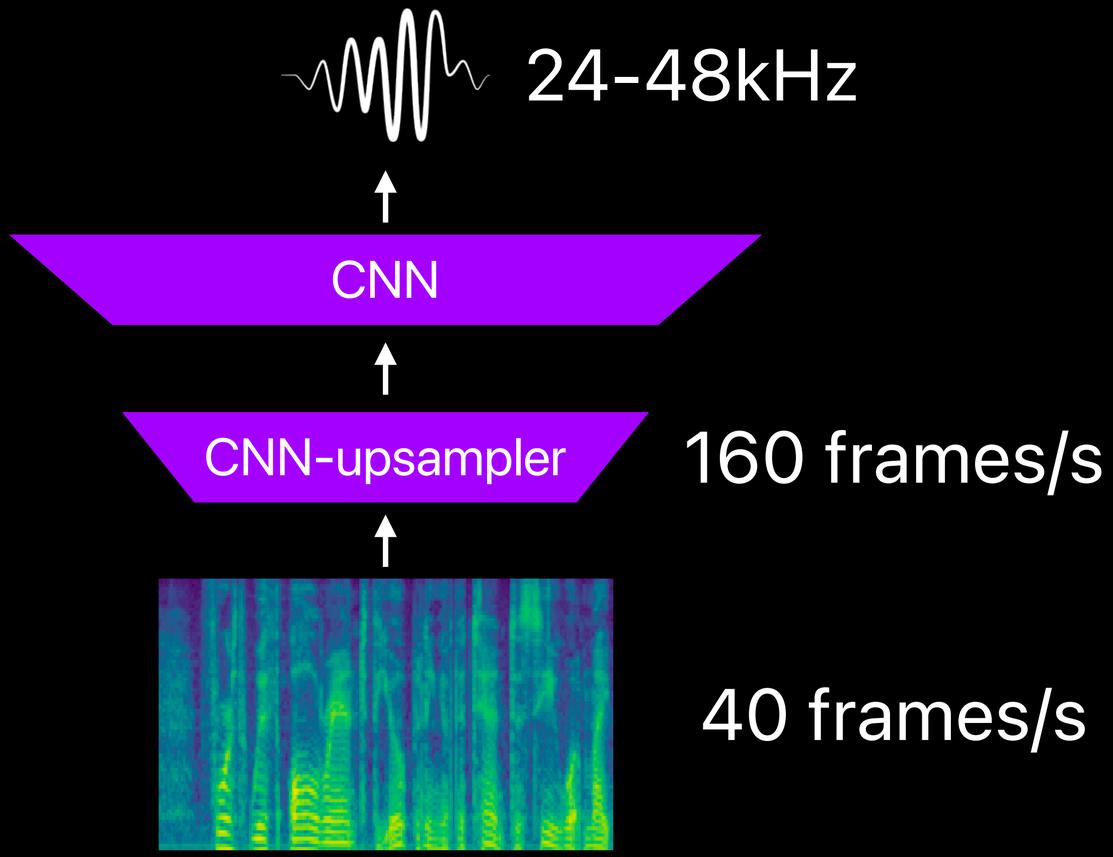
VocStream : streaming vocoder with 1 frame latency



Original

Reconstruction

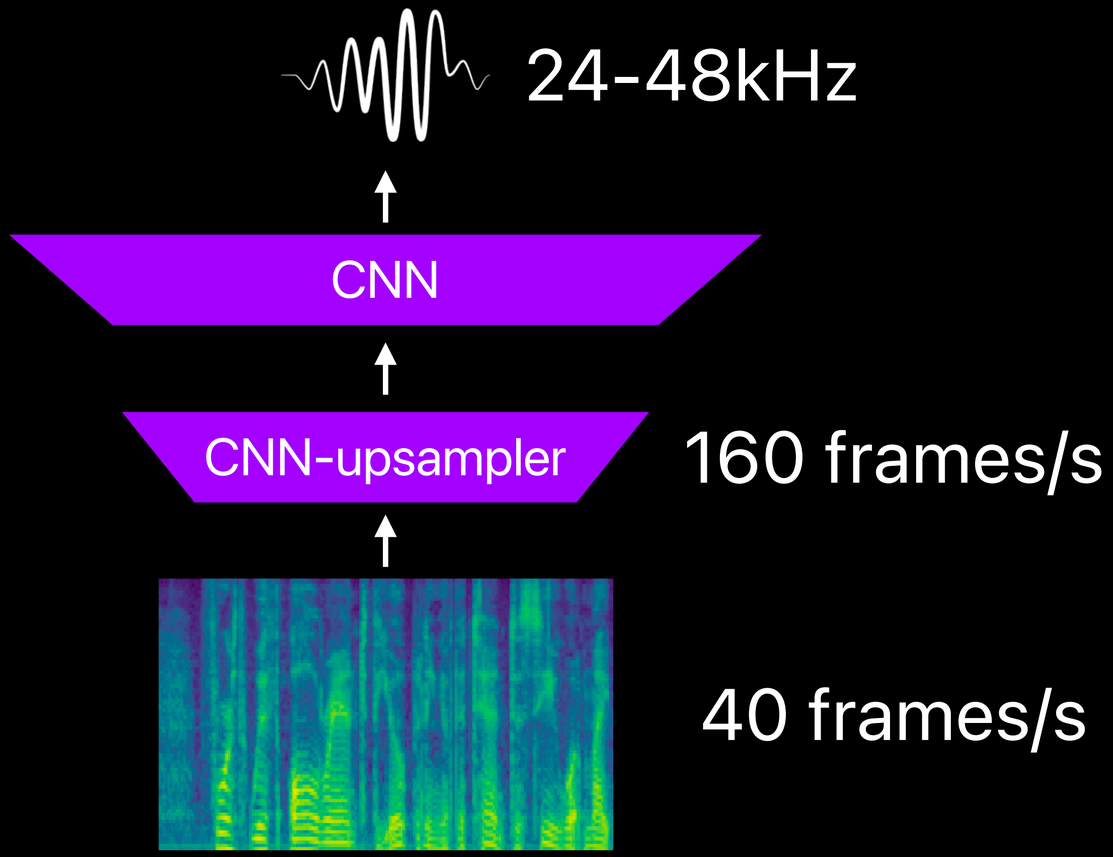
VocStream : streaming vocoder with 1 frame latency



Original

Reconstruction

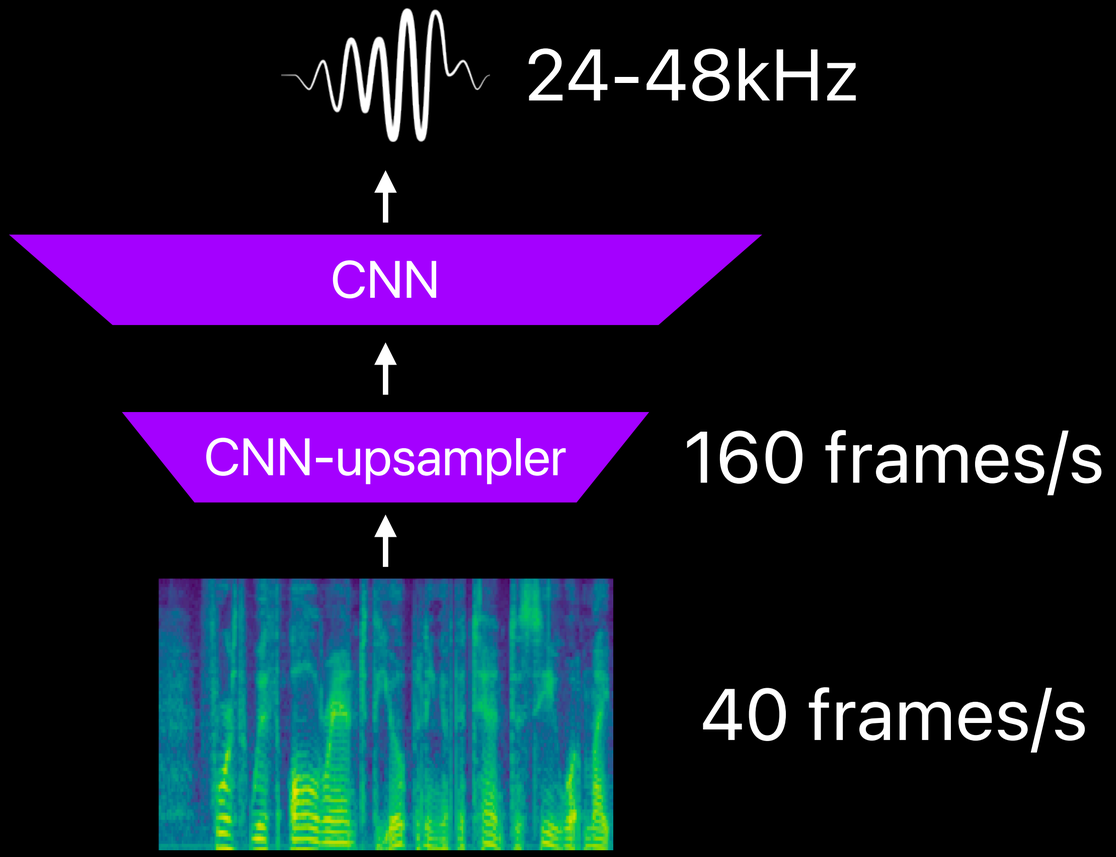
VocStream : streaming vocoder with 1 frame latency



Original

Reconstruction

VocStream : streaming vocoder with 1 frame latency



Original

Reconstruction

Latency: 15ms MLX + M4

ChipChat

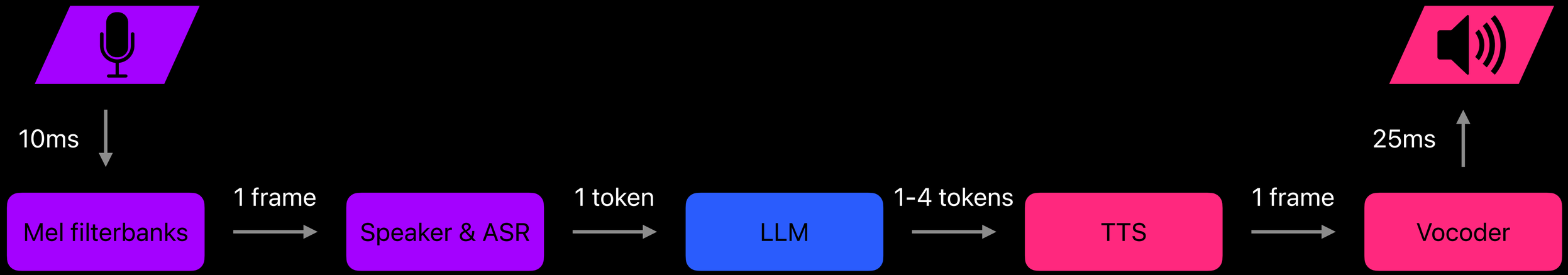
Likhomanenko, T. and et al. *ChipChat: Low-Latency Conversational Agent in MLX*
http://arxiv.org/abs/2509.00078, ASRU 2025 (demo track), *best demo paper award*

1. SOTA low latency

Every component is a standalone Python process that **streams** generated token/frame/audio

1. SOTA low latency

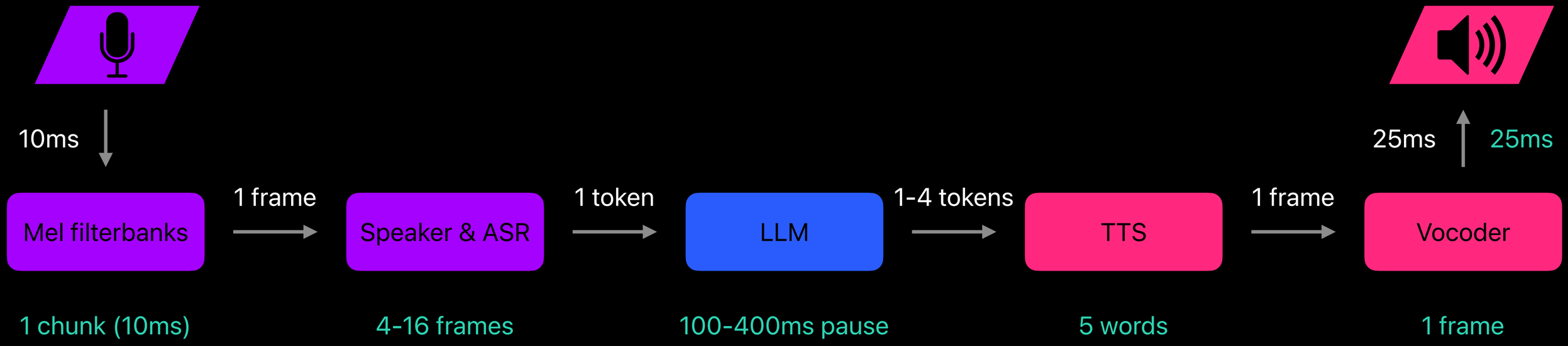
Every component is a standalone Python process that **streams** generated token/frame/audio



Sending to next process

1. SOTA low latency

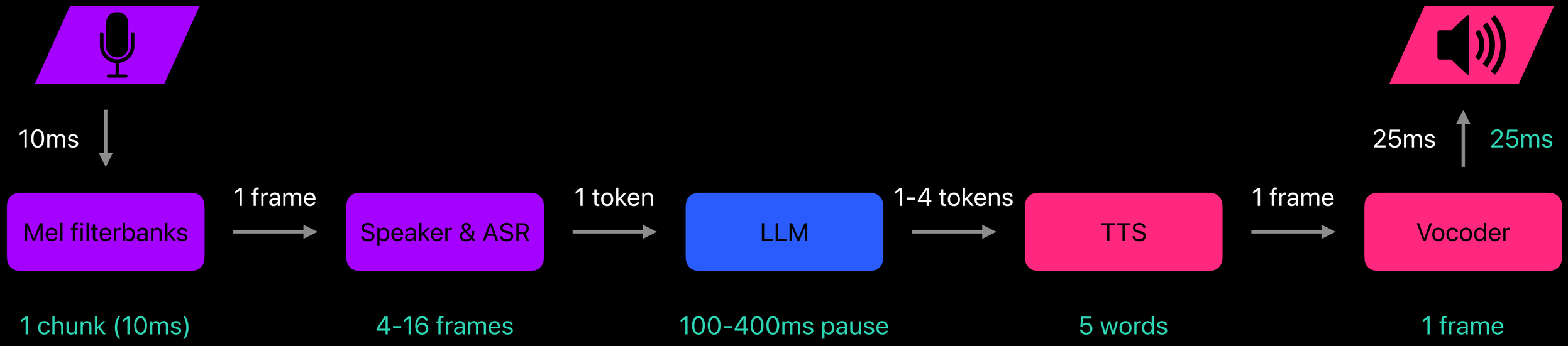
Every component is a standalone Python process that **streams** generated token/frame/audio



Sending to next process

Waiting for process to start generation

1. SOTA low latency

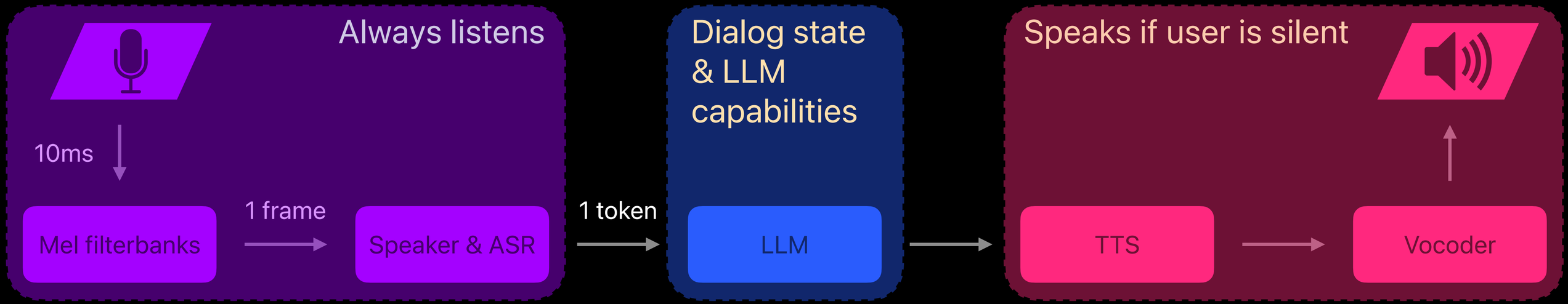


Low latency by design :

<1s on Mac Studio M2 powered by **MLX**

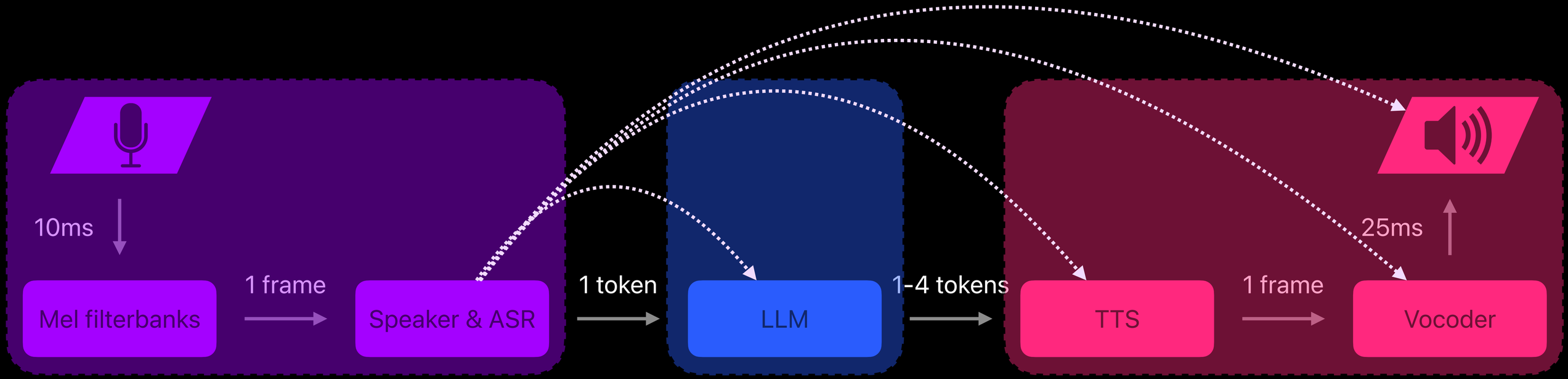
2. User-centric

1½ duplex



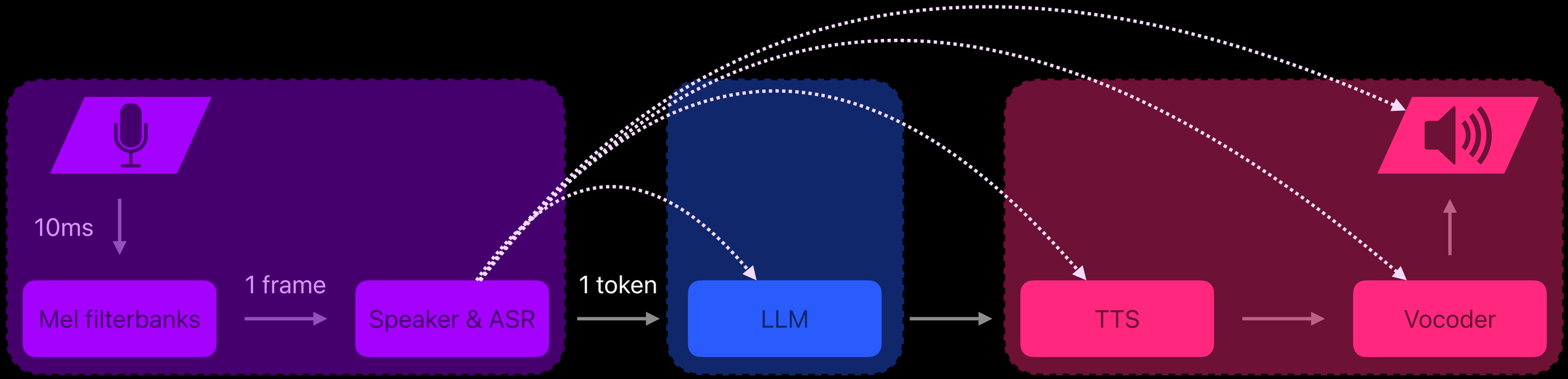
Agent always listens what user speaks

3. New interruption handling



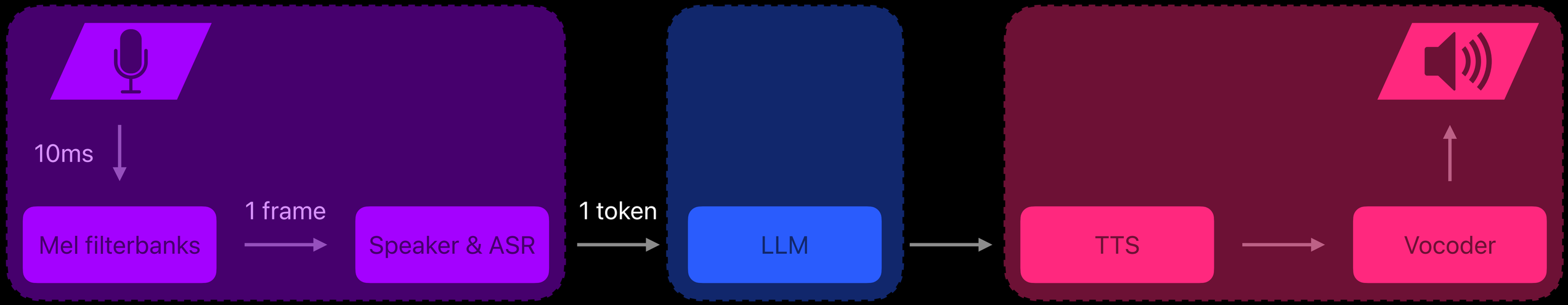
Sending interruption

3. New interruption handling



Stop generation

3. New interruption handling

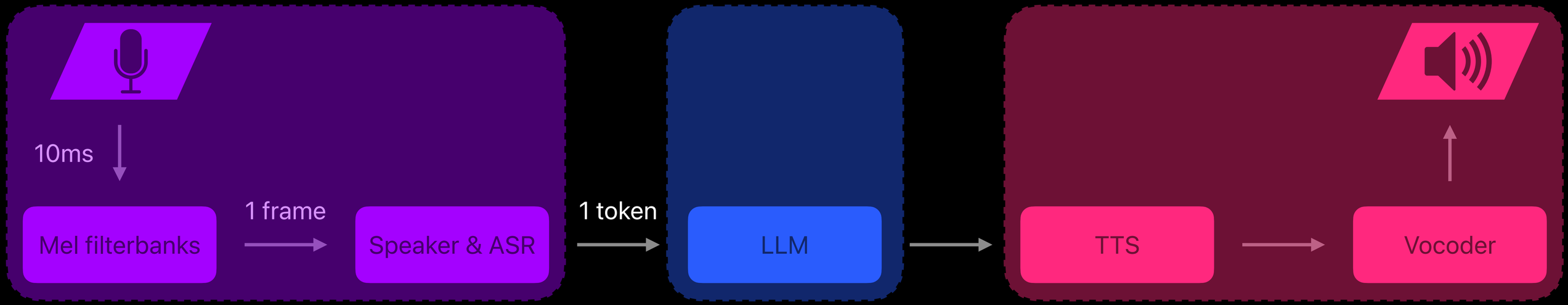


Let’s play a “guess movie” game !

Sure! This movie starts with “a long time ago in a gal...

Star Wars !

3. New interruption handling

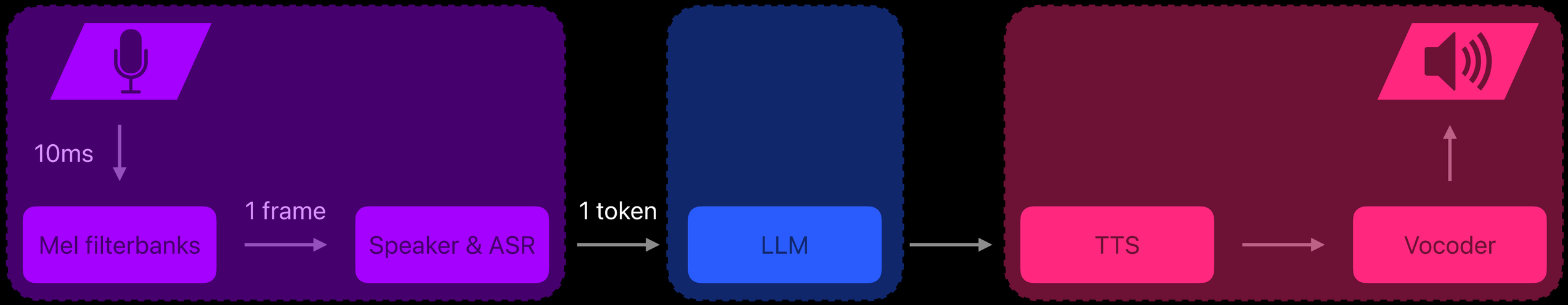


Let’s play a “guess movie” game !

Sure! This movie starts with “a long time ago in a galaxy far, far away…” and the main …

Star Wars !

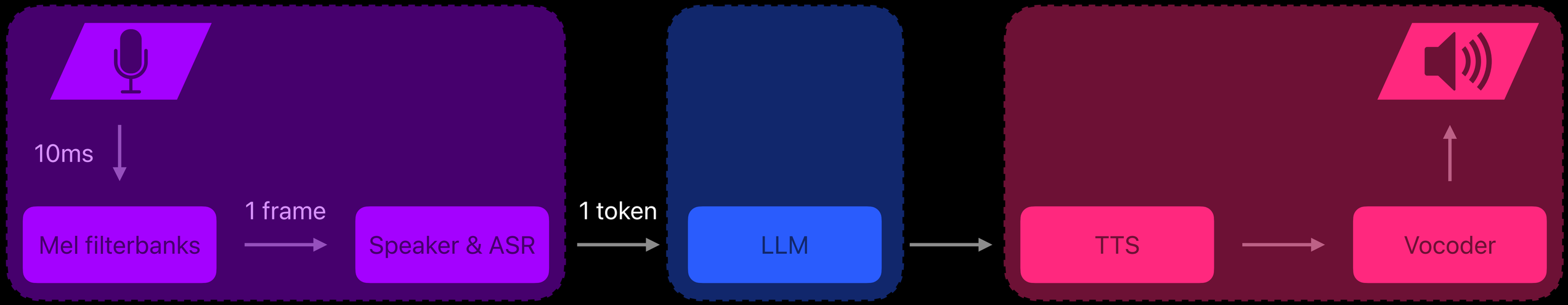
3. New interruption handling

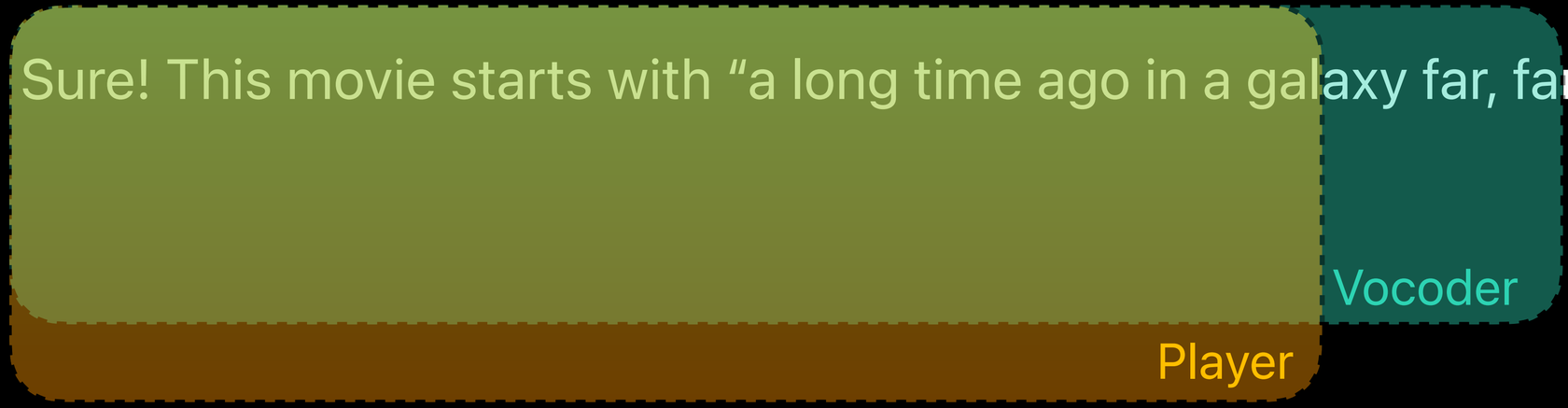


Sure! This movie starts with “a long time ago in a galaxy far, far away…” and the main …

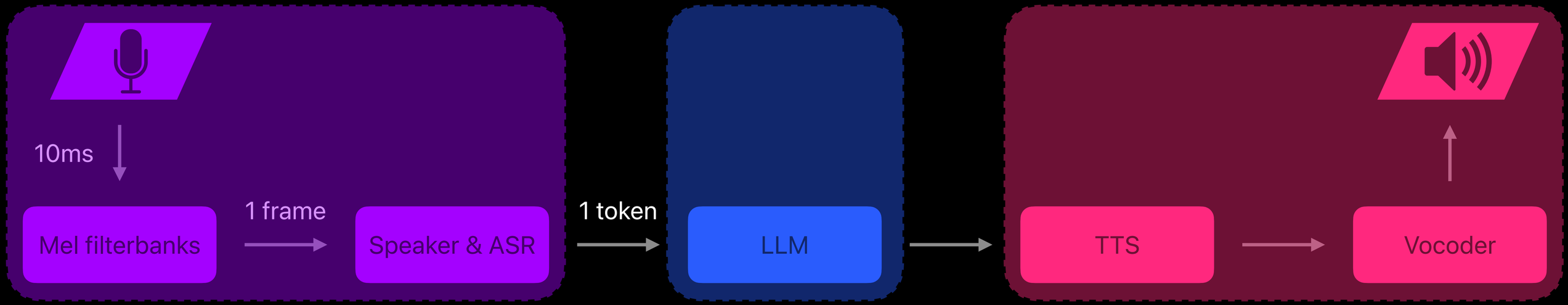
Player

3. New interruption handling

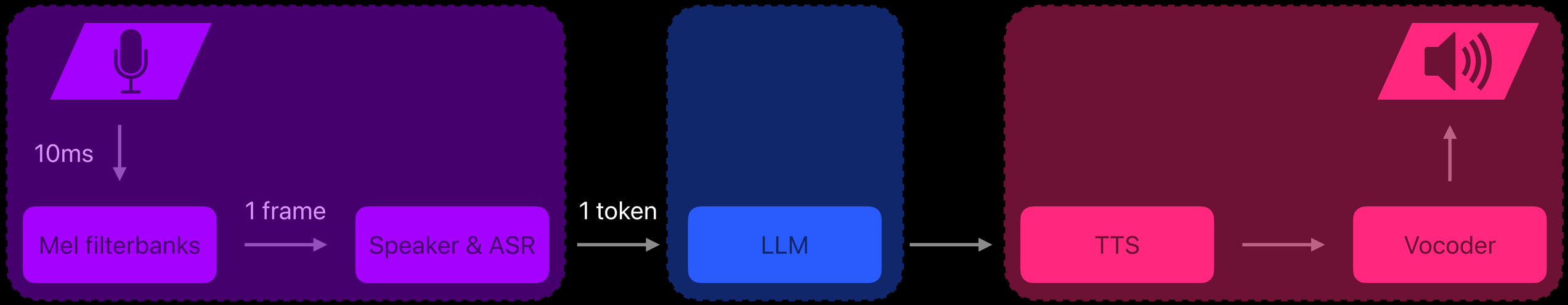




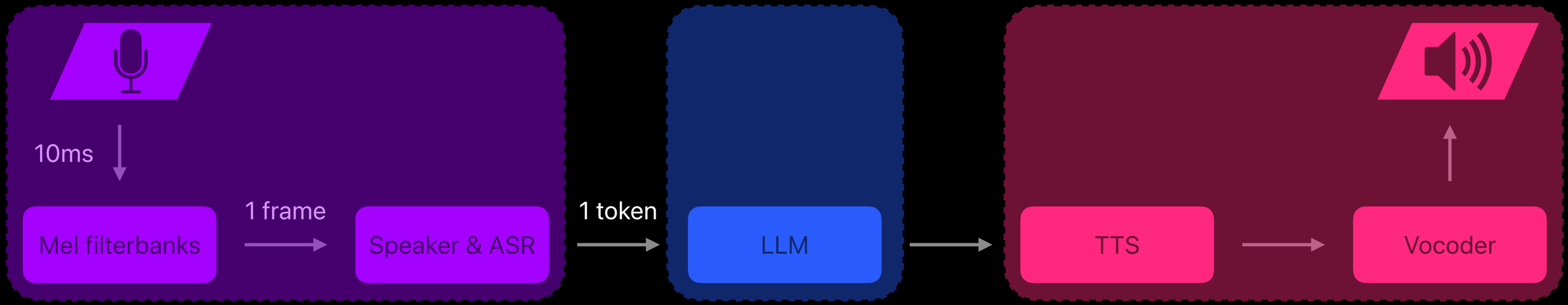
3. New interruption handling

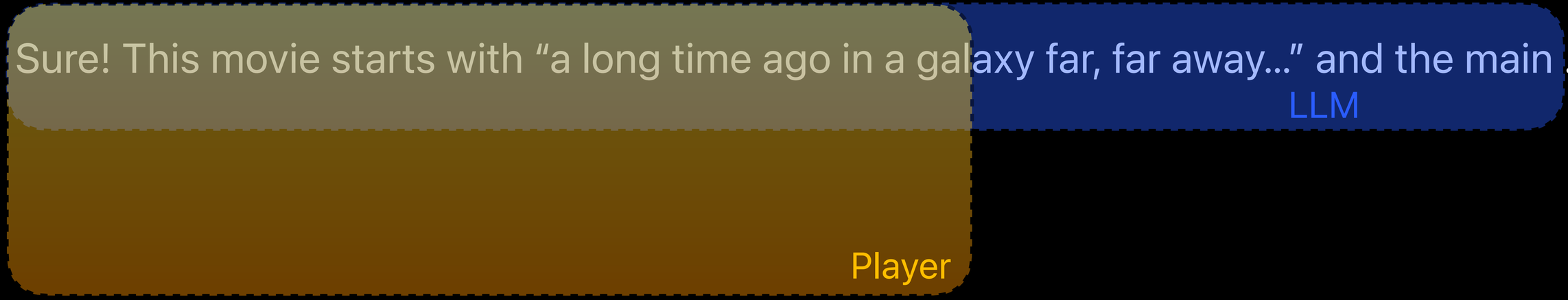


3. New interruption handling

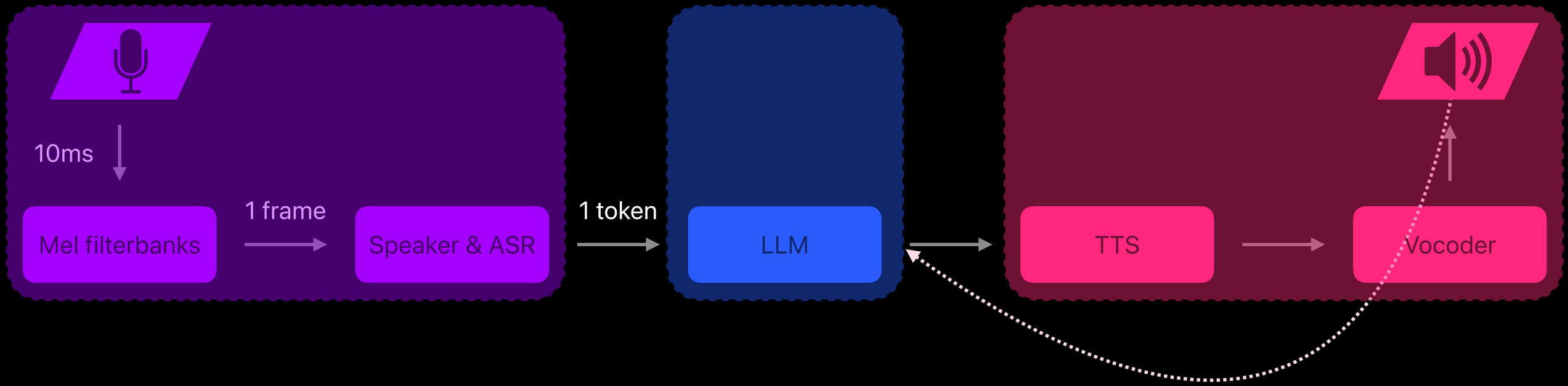


3. New interruption handling





3. New interruption handling

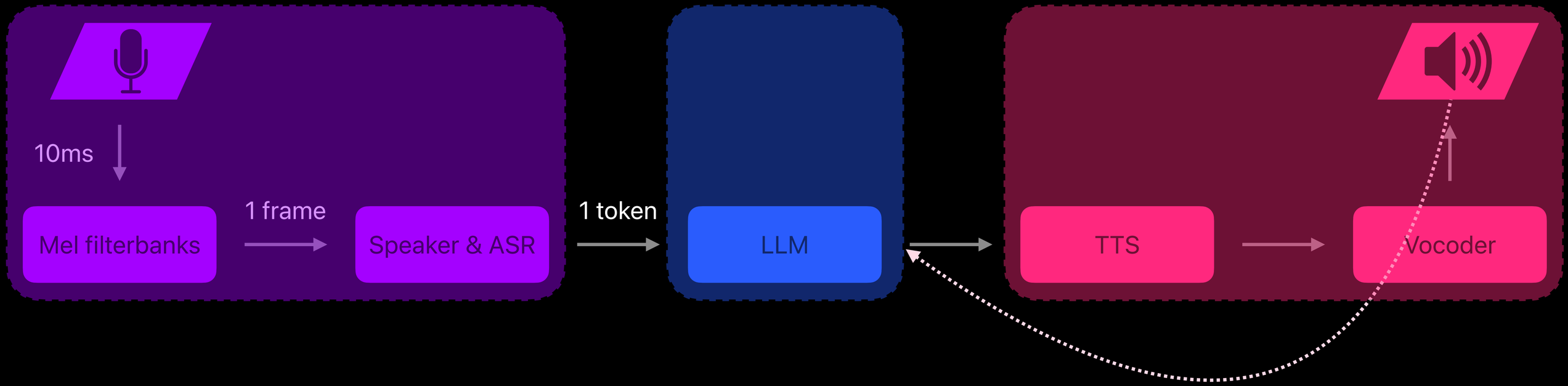


Sure! This movie starts with “a long time ago in a galaxy ~~far, far away…” and the main …~~

LLM

Sending interruption feedback

3. New interruption handling



Precise interruption with correct LLM history

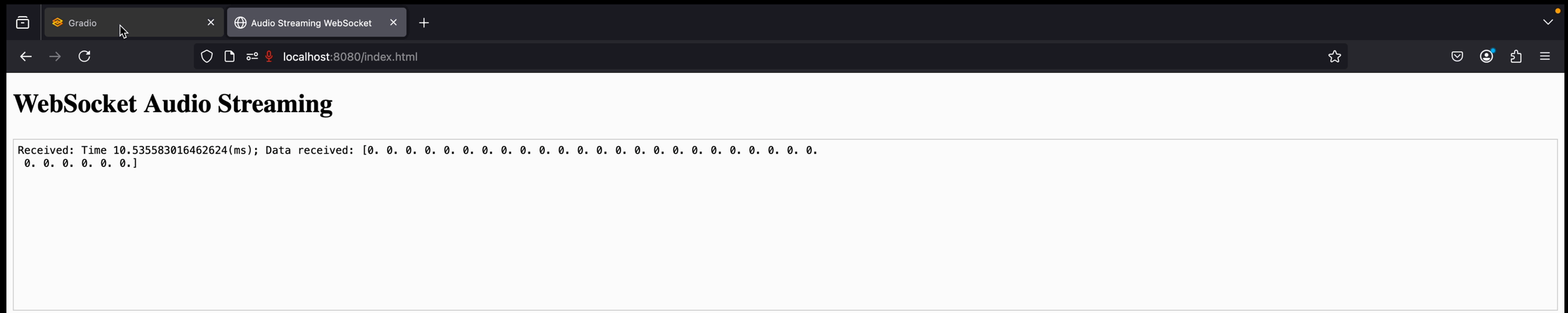
4.  & 

4.  & 

Mac Studio M2, powered by **MLX**

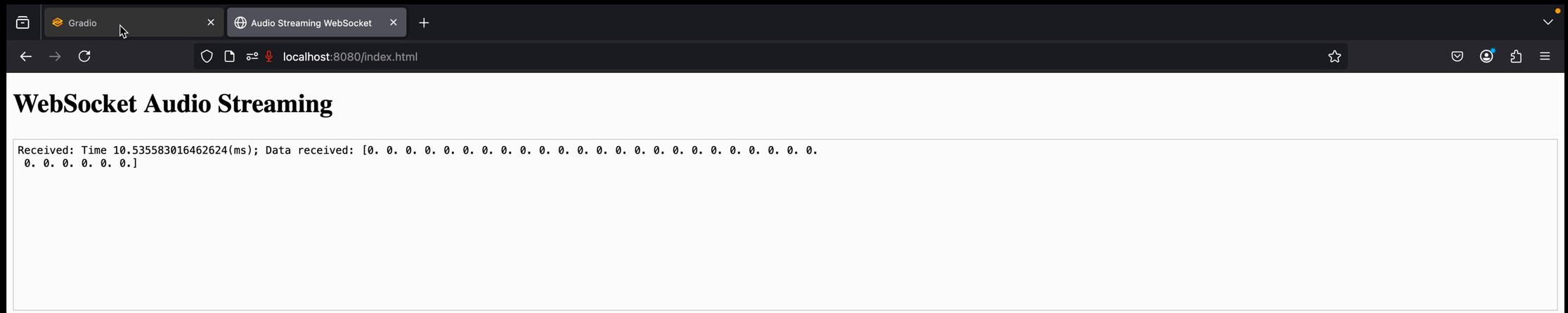
4.  & 

Mac Studio M2, powered by **MLX**



4.  & 

Mac Studio M2, powered by **MLX**



ChipChat models

Microphone : browser-based streaming with near-zero latency and built-in noise cancellation

ChipChat models

Microphone : browser-based streaming with near-zero latency and built-in noise cancellation

Mel filterbanks : 25ms with 10ms hop, running mean and variance for streaming

ChipChat models

Microphone : browser-based streaming with near-zero latency and built-in noise cancellation

Mel filterbanks : 25ms with 10ms hop, running mean and variance for streaming

ASR : Omni-router MOE model with CTC and fully causal masking

ChipChat models

Microphone : browser-based streaming with near-zero latency and built-in noise cancellation

Mel filterbanks : 25ms with 10ms hop, running mean and variance for streaming

ASR : Omni-router MOE model with CTC and fully causal masking

- 650M parameters and 40ms total stride

ChipChat models

Microphone : browser-based streaming with near-zero latency and built-in noise cancellation

Mel filterbanks : 25ms with 10ms hop, running mean and variance for streaming

ASR : Omni-router MOE model with CTC and fully causal masking

- 650M parameters and 40ms total stride
- Acts as VAD

ChipChat models

Microphone : browser-based streaming with near-zero latency and built-in noise cancellation

Mel filterbanks : 25ms with 10ms hop, running mean and variance for streaming

ASR : Omni-router MOE model with CTC and fully causal masking

- 650M parameters and 40ms total stride
- Acts as VAD
- Training on segments of <30s

ChipChat models

Microphone : browser-based streaming with near-zero latency and built-in noise cancellation

Mel filterbanks : 25ms with 10ms hop, running mean and variance for streaming

ASR : Omni-router MOE model with CTC and fully causal masking

- 650M parameters and 40ms total stride
- Acts as VAD
- Training on segments of <30s
- Cache resetting when pauses observed

ChipChat models

Microphone : browser-based streaming with near-zero latency and built-in noise cancellation

Mel filterbanks : 25ms with 10ms hop, running mean and variance for streaming

ASR : Omni-router MOE model with CTC and fully causal masking

- 650M parameters and 40ms total stride
- Acts as VAD
- Training on segments of <30s
- Cache resetting when pauses observed

Speaker model : 3s latency at the beginning, after — 1s striding window

ChipChat models

LLM : SAGE with 8x7B Mixtral model

ChipChat models

LLM : SAGE with 8x7B Mixtral model

TTS : SpeakStream with 5 words latency

ChipChat models

LLM : SAGE with 8x7B Mixtral model

TTS : SpeakStream with 5 words latency

Vocoder : VocStream (based on ParallelWaveGAN) with 1 frame latency

ChipChat models

LLM : SAGE with 8x7B Mixtral model

TTS : SpeakStream with 5 words latency

Vocoder : VocStream (based on ParallelWaveGAN) with 1 frame latency

Player : negligible latency

ChipChat models

LLM : SAGE with 8x7B Mixtral model

TTS : SpeakStream with 5 words latency

Vocoder : VocStream (based on ParallelWaveGAN) with 1 frame latency

Player : negligible latency

Viewer : Gradio to show turns in text form, their motivation and emotions states, latencies of different processes, speaker annotations

ChipChat models

LLM : SAGE with 8x7B Mixtral model

TTS : SpeakStream with 5 words latency

Vocoder : VocStream (based on ParallelWaveGAN) with 1 frame latency

Player : negligible latency

Viewer : Gradio to show turns in text form, their motivation and emotions states, latencies of different processes, speaker annotations

Interruption handling :

ChipChat models

LLM : SAGE with 8x7B Mixtral model

TTS : SpeakStream with 5 words latency

Vocoder : VocStream (based on ParallelWaveGAN) with 1 frame latency

Player : negligible latency

Viewer : Gradio to show turns in text form, their motivation and emotions states, latencies of different processes, speaker annotations

Interruption handling :

- we don’t vocalize any remaining frames as soon as user starts to speak

ChipChat models

LLM : SAGE with 8x7B Mixtral model

TTS : SpeakStream with 5 words latency

Vocoder : VocStream (based on ParallelWaveGAN) with 1 frame latency

Player : negligible latency

Viewer : Gradio to show turns in text form, their motivation and emotions states, latencies of different processes, speaker annotations

Interruption handling :

- we don’t vocalize any remaining frames as soon as user starts to speak
- system clears key-value LLM’s cache to the vocalized-only state

ChipChat latency

	Input	Wait time for input	Output	Inference time per output	Runtime up to the point
	1ms	0ms	1ms	0ms	0ms
Mel	10ms	10ms	1 frame	1ms	11ms
ASR	16 frames	160ms	4 tokens	13ms	165-175ms
LLM State	100-400ms pause	[pause]	states	~560ms	~560ms
LLM	N/A	0ms	1 token	16ms	~576ms
TTS	5 words	[5 words]	1 frame	20ms	~880ms
Vocoder	1 frame	25ms	25ms	13ms	~920ms
	25ms	25ms	25ms	0.2ms	~920ms

Conclusions

Simple discretization with dMel, and independent prediction of channels per step offers an alternative to current tokenization schemes and models in building generative models of speech, including streaming models

- model-free
- robust
- streamable
- multilingual

Redesigned parts of cascaded systems with a focus on overall system performance offer a compelling alternative for next-generation conversational agents

Collaborators


Luke Carlson


Richard Bai


Zijin Gu


Zak Aldeneh


Yizhe Zhang


Shiladitya Dutta


Han Tran


Navdeep Jaitly


Ruixiang Zhang


Huangjie Zheng


Samy Bengio

Code and demos : stay tuned !

SpeakerIPL

Omni-Router ASR OSS

SAGE OSS

dMel Demo

dMel OSS

SpeakStream Demo

Bonus

Closing the Gap Between Text and Speech Understanding in LLMs

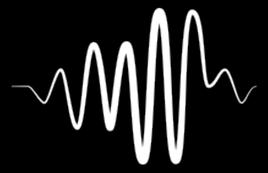
Identify causes of the gap and develop ways to close it

Closing the Gap Between Text and Speech Understanding in LLMs

Identify causes of the gap and develop ways to close it

Closing the Gap Between Text and Speech Understanding in LLMs

Identify causes of the gap and develop ways to close it



Text-Speech LLM

Albert Einstein is a physicist...

Who is Albert Einstein?

Text-Speech LLM

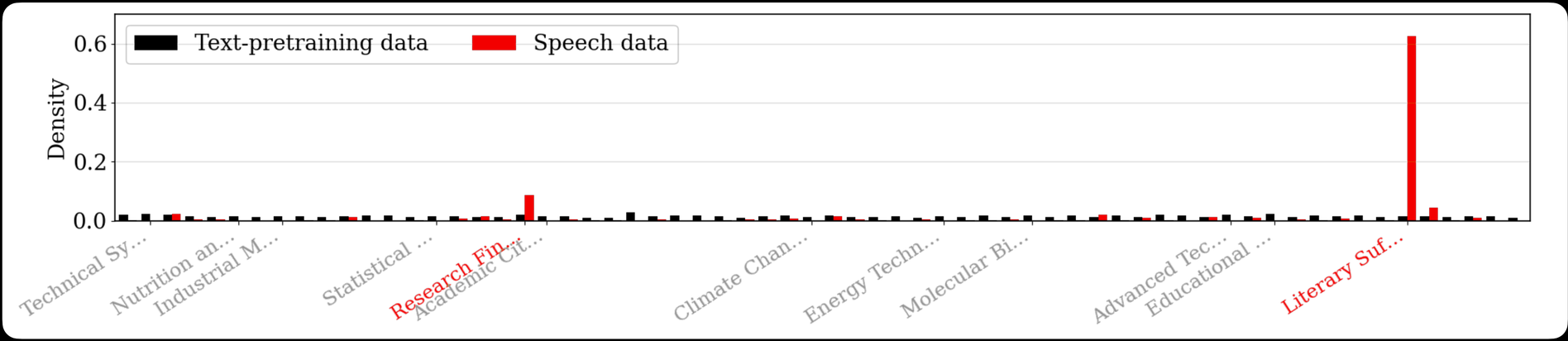
Albert Einstein is a physicist...

Who is Albert Einstein?

LLM

Albert Einstein is a physicist...

Closing the Gap Between Text and Speech Understanding in LLMs



Available speech datasets are **narrow in scope** and do not capture the full general language distribution

Prior efforts to address the gap

Transfer learning

- poor knowledge transfer
- forgetting

Data scaling

- costly and biased toward synthetic speech

Analysing the gap

Measure:

- Speech-text misalignment
- Forgetting of text knowledge

$$M = \sum_{(\mathbf{w},\mathbf{c})\sim\mathbb{Q}} D_{KL}\big(P_\theta(w_{i+1} \mid \mathbf{w}_{\le i}) \,\big\|\, P_\theta(w_{i+1} \mid \mathbf{c}_{\le i})\big),$$

$$F = \sum_{\mathbf{w}\sim\mathbb{Q}} D_{KL}\big(Q_\phi(w_{i+1} \mid \mathbf{w}_{\le i}) \,\big\|\, P_\theta(w_{i+1} \mid \mathbf{w}_{\le i})\big),$$

Where,

P_θ : speech-adapted LLM;
Q_ϕ: pretrained text LLM;
w_i : text token;
c_i : multimodal context

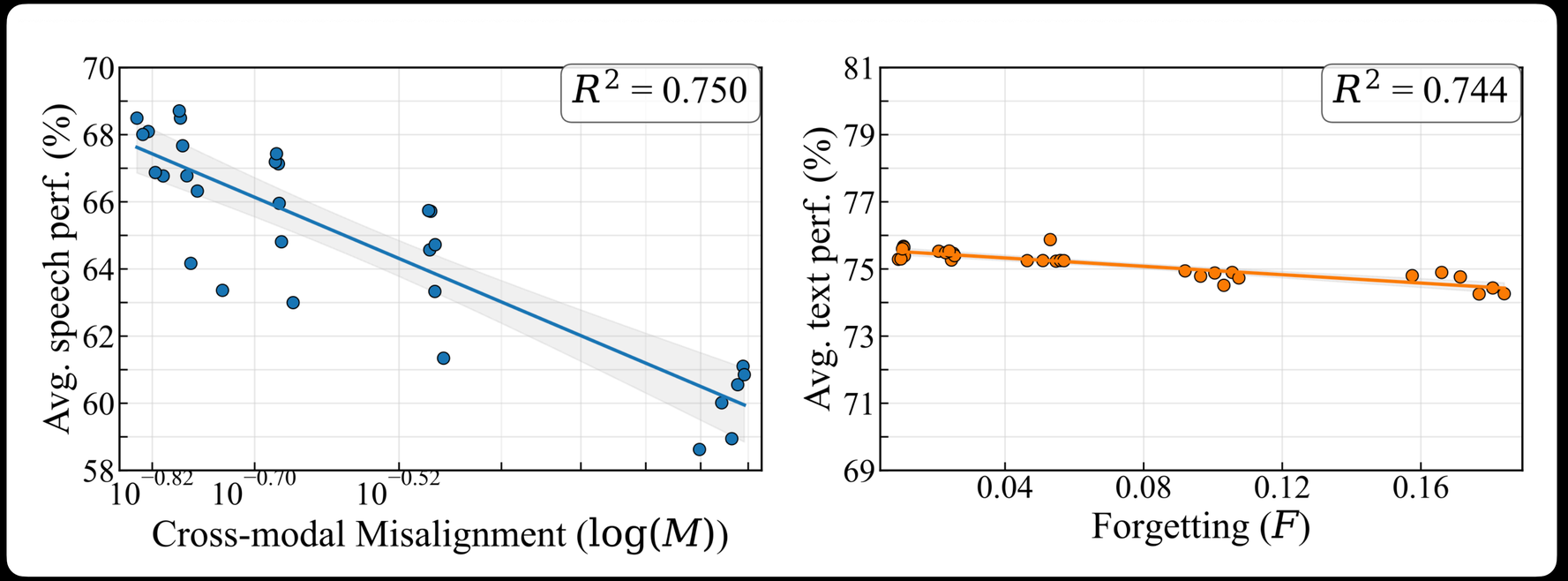
Analysing the gap

Measure:

- Speech-text misalignment
- Forgetting of text knowledge

$$M = \sum_{(\mathbf{w},\mathbf{c})\sim\mathcal{Q}} D_{KL}\big(P_\theta(w_{i+1} \mid \mathbf{w}_{\le i}) \,\big\|\, P_\theta(w_{i+1} \mid \mathbf{c}_{\le i})\big),$$

$$F = \sum_{\mathbf{w}\sim\mathcal{Q}} D_{KL}\big(Q_\phi(w_{i+1} \mid \mathbf{w}_{\le i}) \,\big\|\, P_\theta(w_{i+1} \mid \mathbf{w}_{\le i})\big),$$



Analysing the gap

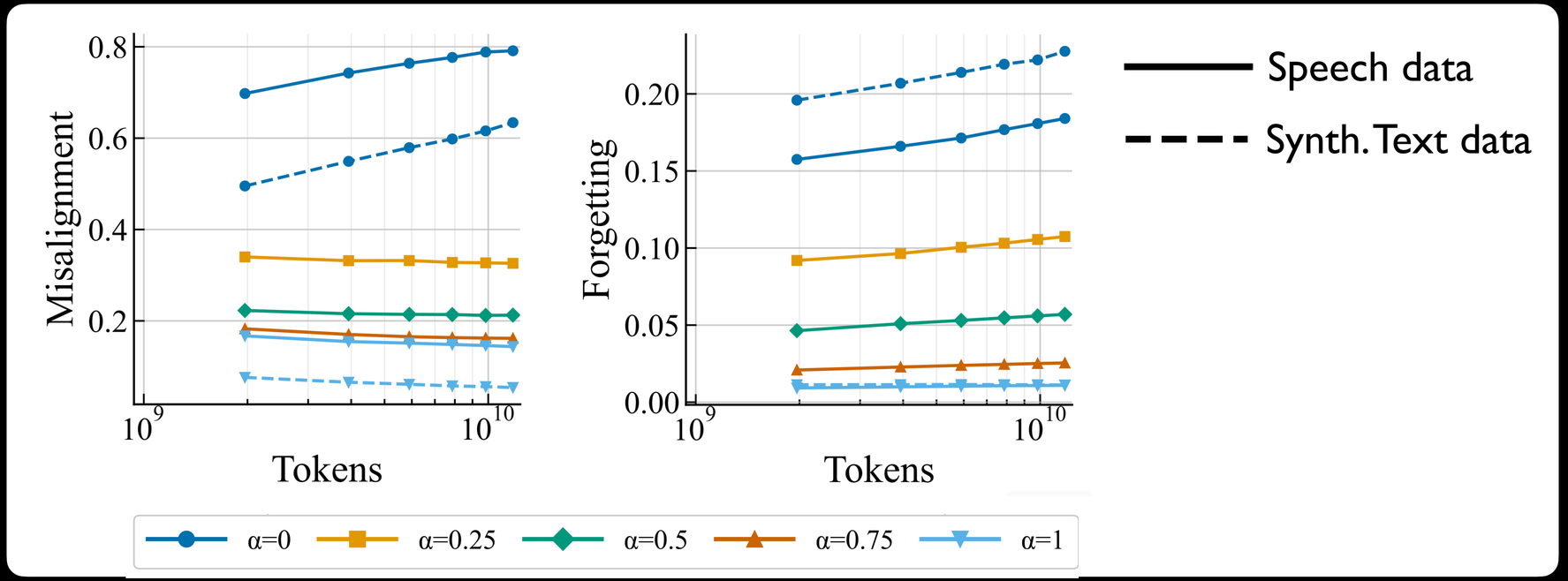
Cross-modal distillation (higher α) and domain-matched data reduce misalignment and forgetting

Measure:

- Speech-text misalignment
- Forgetting of text knowledge

$$M = \sum_{(\mathbf{w},\mathbf{c})\sim\mathcal{Q}} D_{KL}\big(P_\theta(w_{i+1} \mid \mathbf{w}_{\le i}) \,\big\|\, P_\theta(w_{i+1} \mid \mathbf{c}_{\le i})\big),$$

$$F = \sum_{\mathbf{w}\sim\mathcal{Q}} D_{KL}\big(Q_\phi(w_{i+1} \mid \mathbf{w}_{\le i}) \,\big\|\, P_\theta(w_{i+1} \mid \mathbf{w}_{\le i})\big),$$



Closing the gap

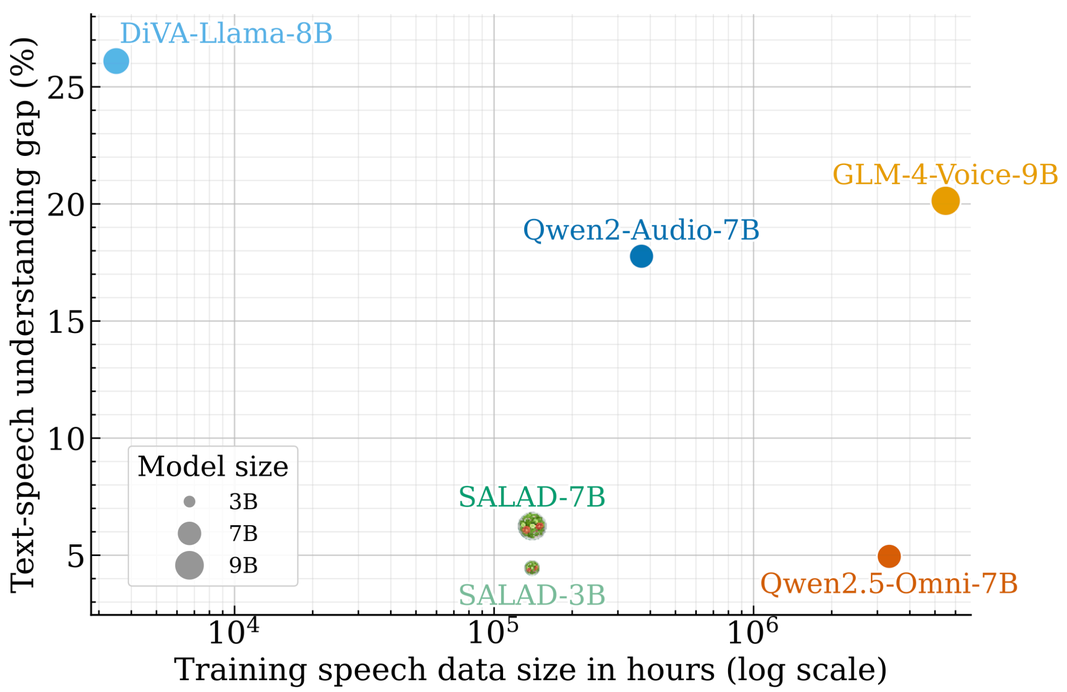
SALAD: Sample-efficient Alignment through Learning with Active selection and cross-modal Distillation

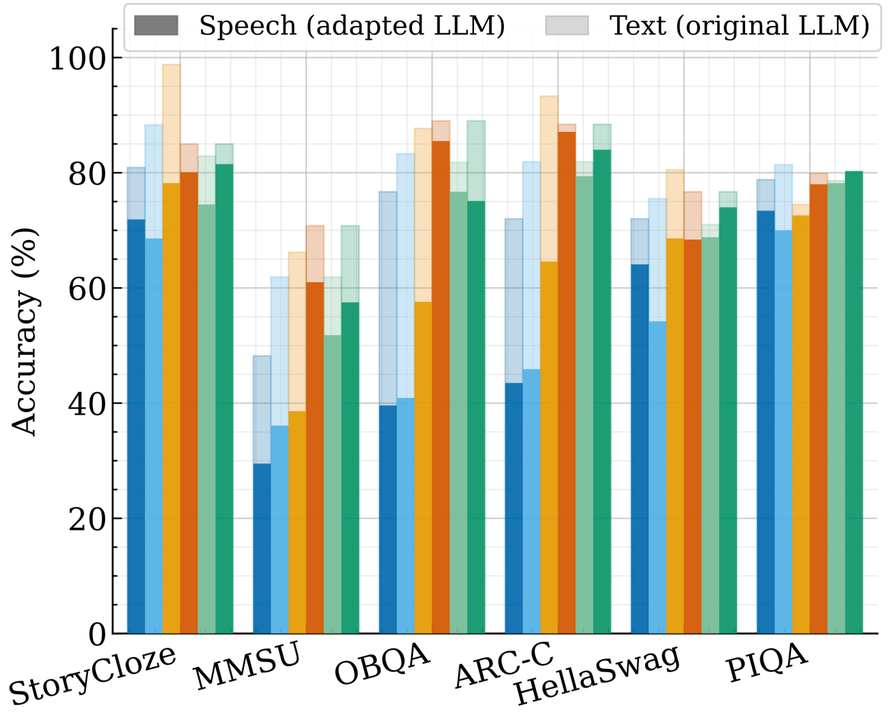
Cross-modal distillation improves alignment and reduces forgetting

Active selection efficiently matches domains using targeted synthetic speech, avoiding large-scale synthesis

Closing the gap

SALAD: Sample-efficient Alignment through Learning with Active selection and cross-modal Distillation





Qwen2-Audio-7B DiVA-Llama-8B GLM-4-Voice-9B

Qwen2.5-Omni-7B SALAD-3B SALAD-7B

Apple Machine Learning Research (MLR)

Fundamental, Open-Ended Research Group

Have a lively internship program — work with us on interesting research papers