The Voicebox Model and Its Applications

Leda Sari (Otter.ai, US)

May 08, 2025

* The works presented here were done at Meta along with many collaborators.

Collaborators

Voicebox Model:

- Matthew Le
- Apoorv Vyas
- Bowen Shi
- Brian Karrer
- Leda Sari
- Rashel Moritz
- Mary Williamson
- Vimal Manohar
- Yossi Adi
- Jay Mahadeokar
- Wei-Ning Hsu

Applications:

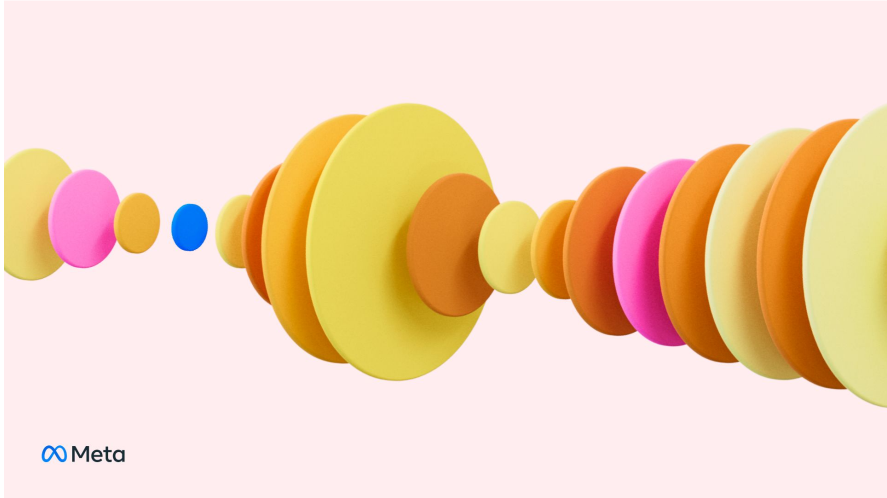
- Hira Dhamyal
- Roshan Sharma
- Shuo Liu
- Chunyang Wu
- Gil Keren
- Yuan Shangguan
- Nayan Singhal
- Suyoun Kim
- Ozlem Kalinli
- Daniel Lazar
- Trang Le
- Akshat Shrivastava
- Kwanghoon An
- Piyush Kansal
- Mike Seltzer

Outline

1. The Voicebox Model
2. Applications: Use of Voicebox generated data
    a. Automatic Speech Recognition (ASR)
    b. Spoken Language Understanding (SLU)
3. Extensions of the Voicebox Model
4. Potential Future Directions

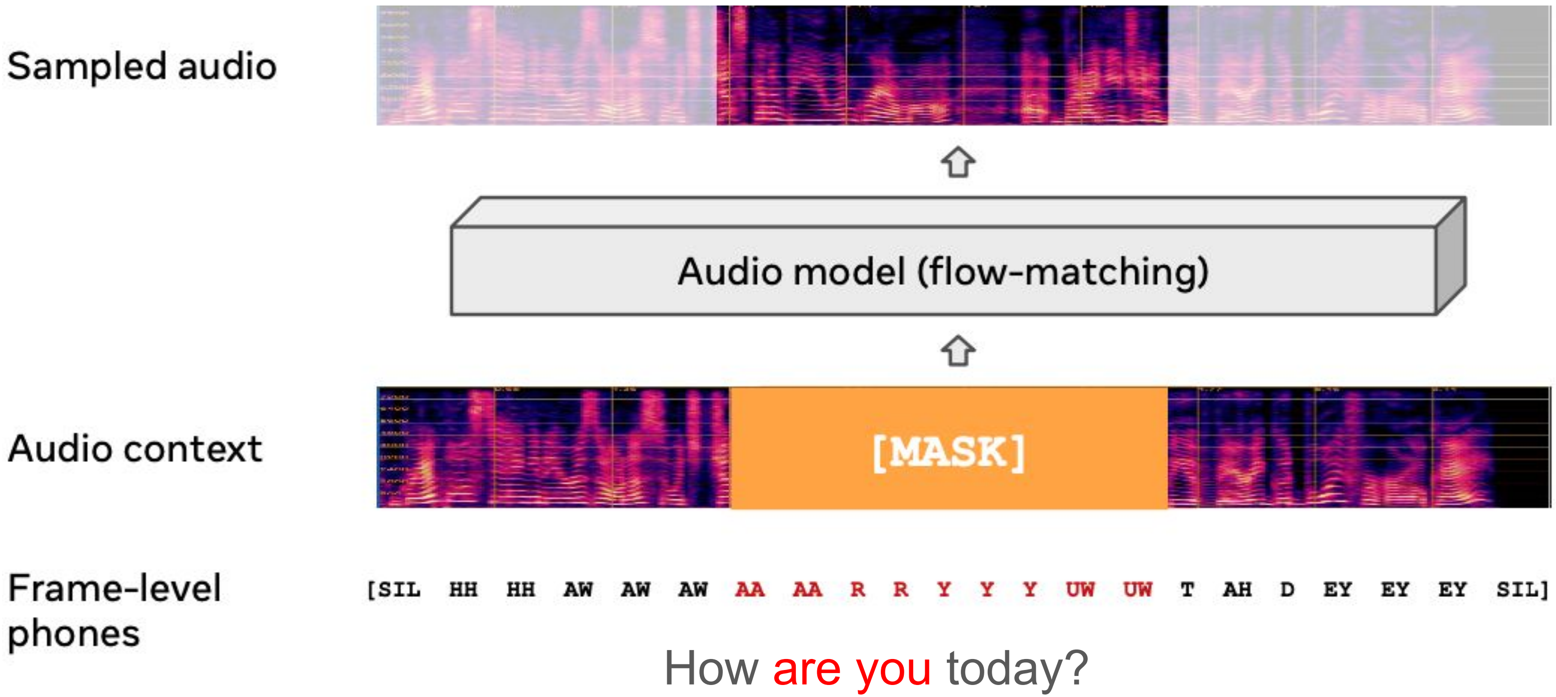
The Voicebox Model

- A text-guided generative model for speech
- A non-autoregressive model
- A flow-matching model
- Trained to infill speech
- Can be used for
  - Speech infilling
  - Speech denoising
  - Zero-shot TTS
  - Content editing



A text-guided generative model for speech

More specifically, the phonetic alignment



A flow-matching model: Definitions

- Time
- Probability Distribution
- A time-dependent vector field
- A flow
- The relationship
- Parametrized estimate
- Flow matching objective

$$t_0 \longrightarrow t_1$$

$$p_0 \longrightarrow p_1 \approx q$$

$$u_t$$

$$\phi_t$$

$$\frac{d\phi_t(x)}{dt} = u_t(\phi_t(x)), \text{ and } \phi_0(x) = x$$

$$v_t(x;\theta)$$

$$\mathcal{L}_{FM}(\theta) = \mathbb{E}_{t,p_t}||u_t(x) - v_t(x;\theta)||^2$$

A flow-matching model: CFM

- The probability path can be constructed via a mixture of simpler conditional paths (Lipman et al., 2022)

$$p_0(x \mid x_1) = p_0(x) \text{ and } p_1(x \mid x_1) = \mathcal{N}(x \mid x_1, \sigma^2 I)$$

$$\mathcal{L}_{CFM}(\theta) = \mathbb{E}_{t,q(x_1),p_t(x|x_1)}||u_t(x|x_1) - v_t(x;\theta)||^2$$

- It is easier to sample from the conditional distribution

A flow-matching model: the OT path

- How do you choose the path from $p_0 \longrightarrow p_1 \approx q$
- Optimal Transport Path (Lipman et al. 2022)

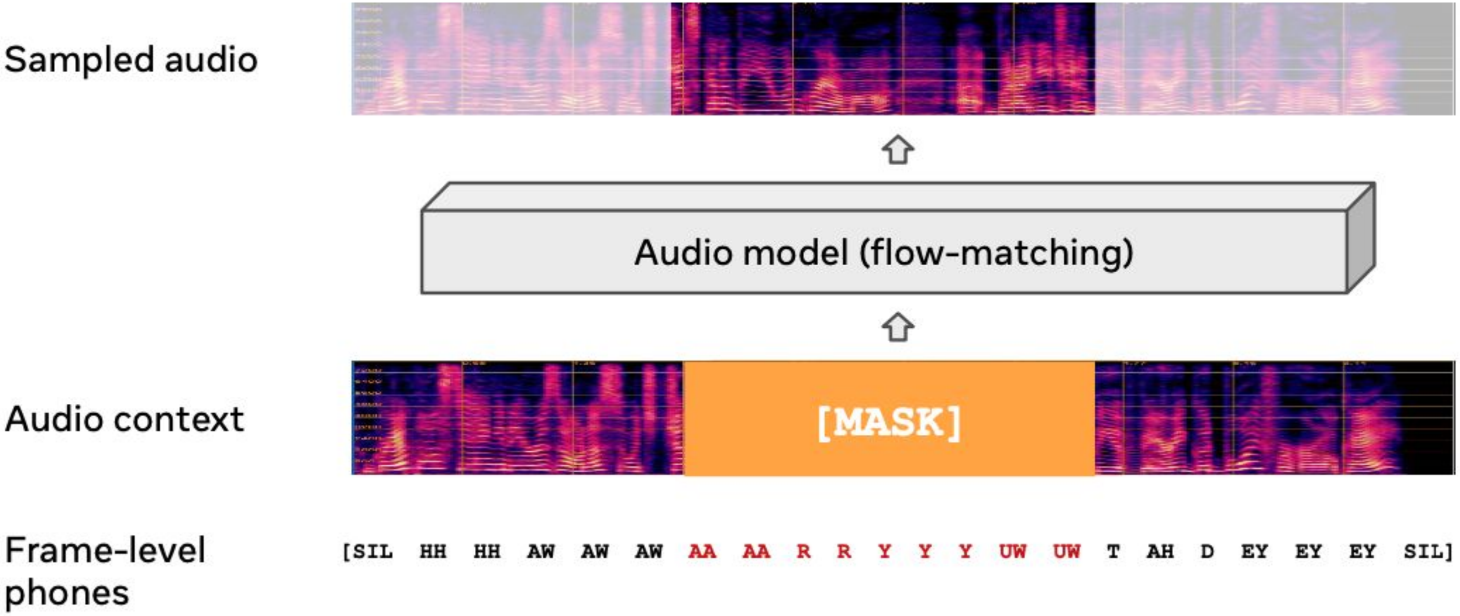
$$p_t(x \mid x_1) = \mathcal{N}(x \mid tx_1, (1-(1-\sigma_{min})t)^2 I)$$

$$u_t(x \mid x_1) = \frac{x_1 - (1-\sigma_{\min})x}{1-(1-\sigma_{\min})t}$$

- A simple flow with a constant speed and direction
- Another alternative is a diffusion path with Gaussian conditional probability paths with specific choices of mean and variance (see Lipman et al.)

Flow-matching model in practice

- A neural network is used to parameterize the conditional vector field $v_t(x_t, x_{\text{ctx}}, z; \theta)$

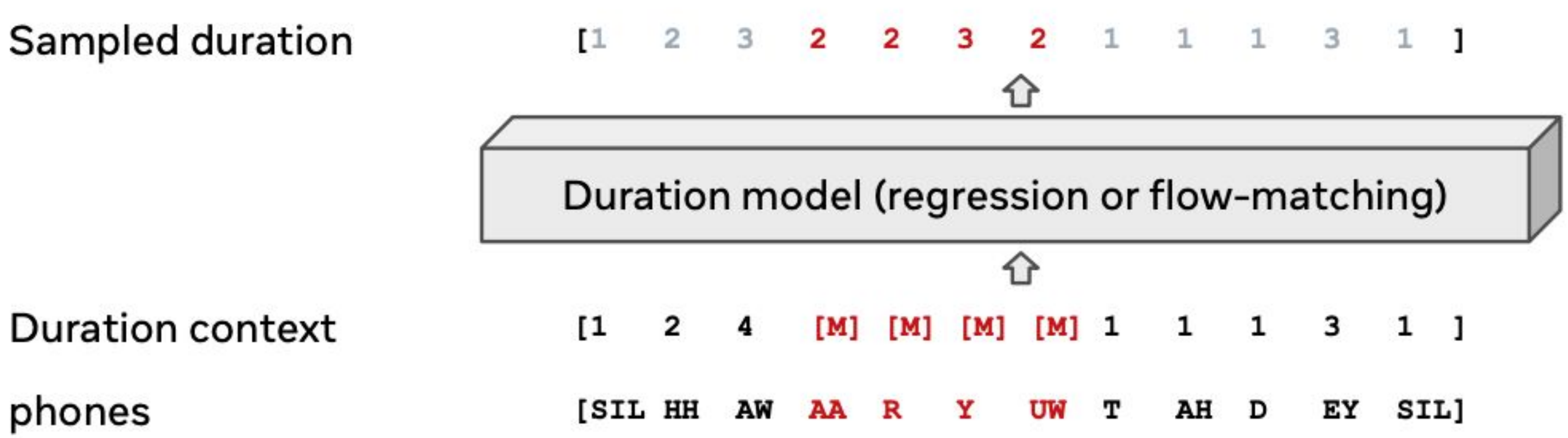


Inference

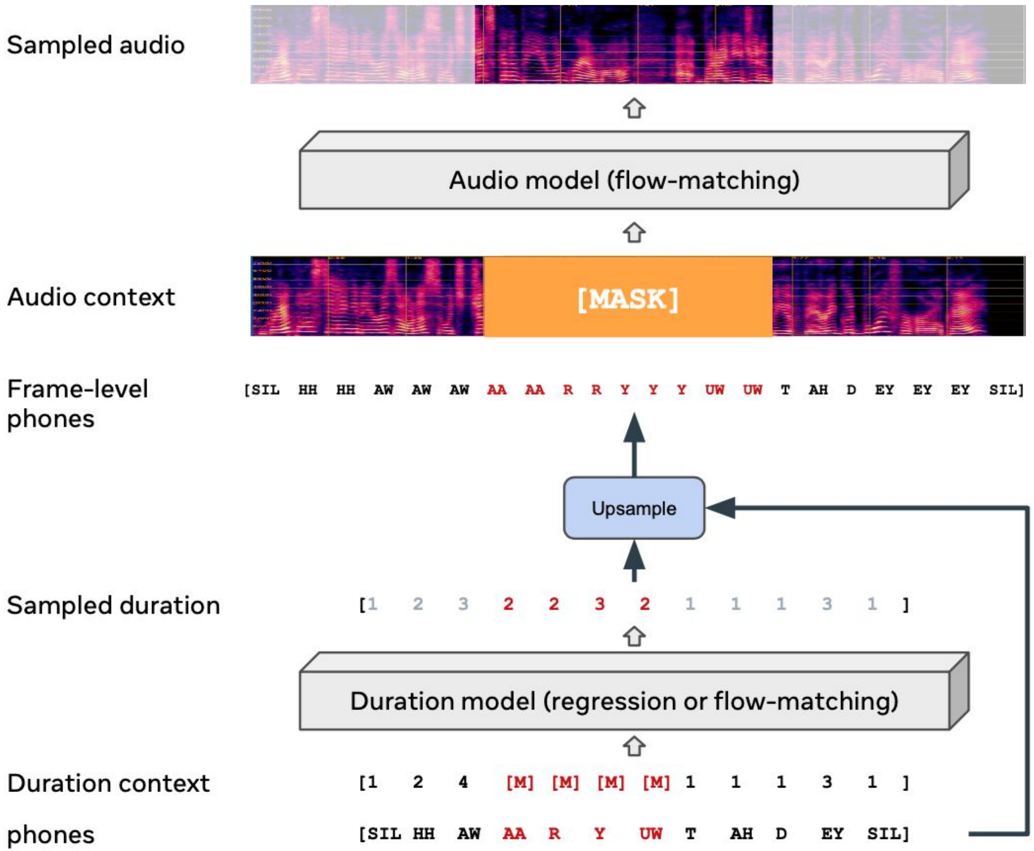
- And ODE solver computes $x_1 = \phi_1(x_0)$
- Starts from the noise samples
- Evaluates $d\phi_t(x_0)/dt$
  by number of function evaluation (NFE) times to approximate the integration from t = 0 to t = 1

Where does the alignment information come from?

- Forced alignment (e.g. Montreal forced aligner)
- Duration Model
  - A regression model based duration estimates
  - A **flow-matching model** based duration estimates



Putting the model together



Speech Denoising, Zero-shot TTS, and TTS

How to Evaluate this Model?

- Intelligibility: word error rate (WER) from an ASR system
- Coherence: speaker similarity based on speaker embeddings
- Subjective evaluations (MOS)

Training Setup

- Data
  - English-only model on 60K hours ASR-transcribed English audiobooks
  - Multilingual model on 50K hours of multilingual audiobooks from six languages: English (En), French (Fr), German (De), Spanish (Es), Polish (Pl) and Portuguese (Pt).
- Montreal Forced Aligner (MFA) for the phonetic alignment
- HiFi-GAN as vocoder
- The flow-matching models are transformer based
- Other hyperparameters are available in the paper

TTS Quality Comparison

Table 2: English zero-shot TTS results on filtered LS test-clean. *obtained via personal communication.

Model	WER	SIM-o	SIM-r	QMOS	SMOS
Ground truth	2.2	0.754	n/a	$3.98_{\pm 0.14}$	$4.01_{\pm 0.09}$
cross-sentence					
A3T	63.3	0.046	0.146	-	-
YourTTS	7.7	0.337	n/a	$3.27_{\pm 0.13}$	$3.19_{\pm 0.14}$
VALL-E	5.9	-	0.580	-	-
VB-En	1.9	0.662	0.681	$3.78_{\pm 0.10}$	$3.71_{\pm 0.11}$
continuation					
A3T	18.7	0.058	0.144	-	-
VALL-E	3.8	0.452*	0.508	-	-
VB-En ($\alpha = 0.7$)	2.0	0.593	0.616	-	-

Use of the synthetic data

Table 6: Performance of ASR models trained on real or synthetic speech, tested on *real* speech and decoded with or without a 4-gram language model.

	WER on real data			
	No LM		4-gram LM	
ASR training data	test-c	test-o	test-c	test-o
Real audio (100hr)	9.0	21.5	6.1	16.2
Real audio (960hr)	2.6	6.3	2.2	5.0
VITS-LJ	58.0	81.2	51.6	78.1
VITS-VCTK	33.8	55.5	30.2	53.1
YourTTS (ref=LS train)	25.0	54.6	20.4	51.2
VB-En ($\alpha = 0$, dur=regr)	7.1	17.6	6.5	14.6
VB-En ($\alpha = 0$, dur=FM, $\alpha_{dur} = 0$)	3.1	8.3	2.6	6.7

Applications

Now, what can we do with these signals?

1. Directly use the signals in your TTS application
2. Indirectly consume the generated data
   Augment your training dataset for
   a. ASR
   b. SLU

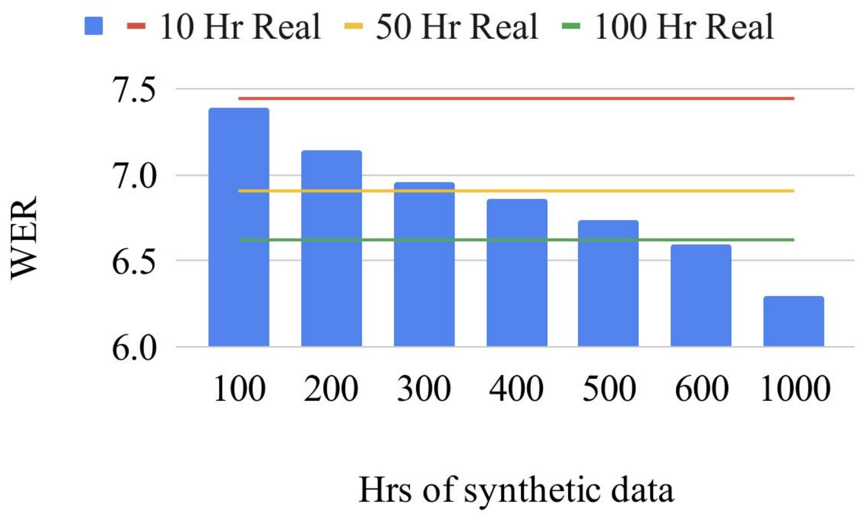
Using Voicebox-based Synthetic Speech for ASR Adaptation (Dhamyal et al., 2024)

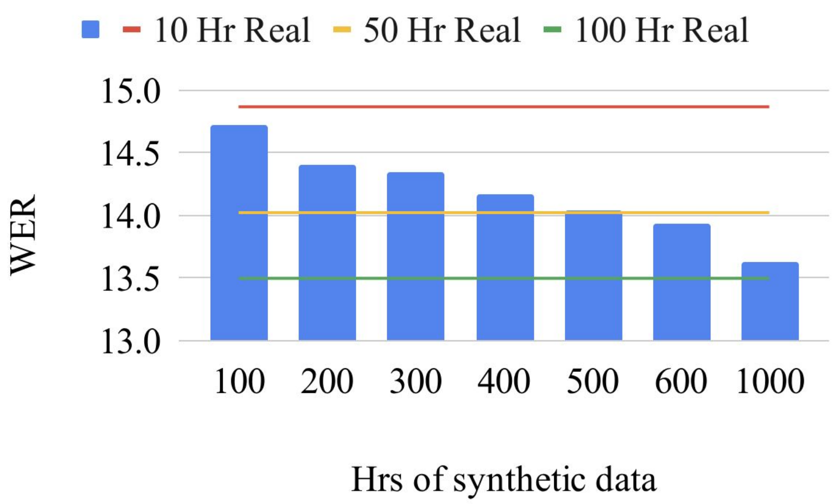
- How much synthetic data can match the WER performance of a real ASR model?
  - Synthetic data only model versus model trained on real data
  - Tests are performed on **real** data
- What kind of speech should we generate to improve ASR?
  - Lexically diverse data
  - Acoustically diverse data with a similar vocab
  - Combination

Experimental Settings

- Librispeech and the Libri-text (text for the Librispeech LM training data)
- An in-house RNN-T based ASR model
- A graphemic (not phonetic) version of the Voicebox model
- Comparisons
  - Real data-only baselines
  - Synthetic data-only (S)
  - Acoustic variability only (A)
  - Lexical variability only (L)
  - Acoustic and lexical variability (L + A)
- JAT model (Kim et al., EMNLP 2022) for the lexical variability experiments

How much synthetic data to match the real data?

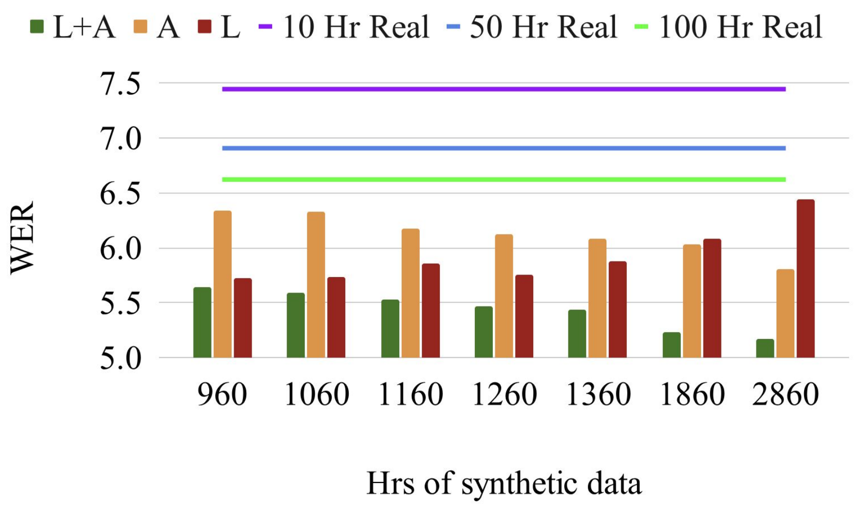


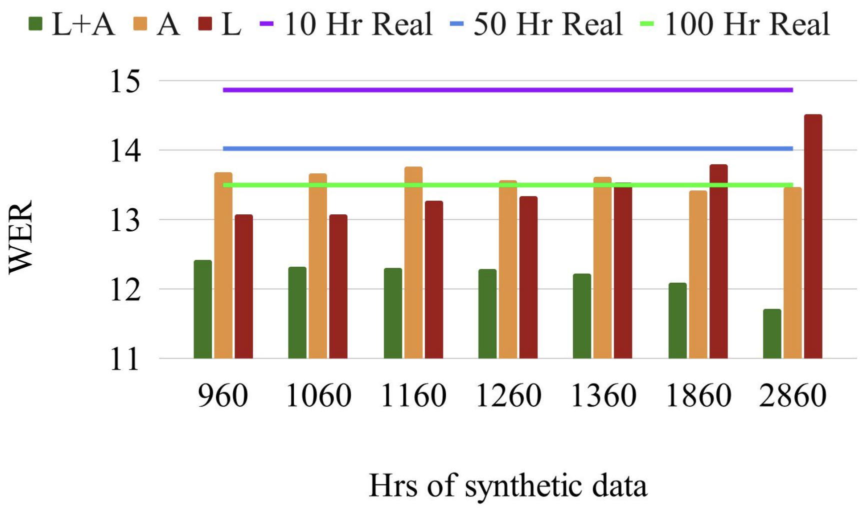


test-clean
~7x more synthetic

test-other
~10x more synthetic

What kind of synthetic data?





WER(L+A) is lower than WER(A) or WER(L)

Voicebox for ASR

- In clean ASR conditions, we need about 7x synthetic data to match the WER performance of the real data on real test sets.
- In **noisy ASR** conditions, we need about **10x synthetic data** to match the WER performance of the real data on real test sets.
- Lexical and acoustic diversity are both crucial

Improving Spoken Semantic Parsing using Synthetic Data from Large Generative Models (Sharma et al., 2024)

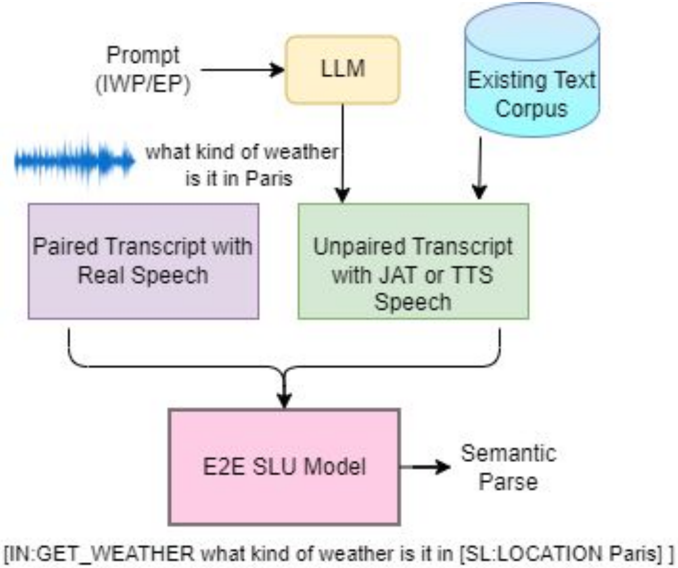
- Spoken Semantic Parsing (SSP) is the SLU task that involves transforming a recording to a machine-comprehensible parse tree
- Requires triplets of (speech, transcript, semantic parse)
  - An audio file saying “I would like to fly from San Francisco to Montreal”
  - I would like to fly from San Francisco to Montreal
  - I would like to <intent: fly> from <from_entity: San Francisco> to <to_entity: Montreal>
- Limited amounts of such paired data

Problems addressed in this paper

- Q1: How can we use unpaired data?
  - ASR (speech + transcript)
  - NLU (text + semantic parse)
  - Some paired data (speech + semantic parse)
- Q2: How to deal with existing domains (ED) versus new domains (ND)?

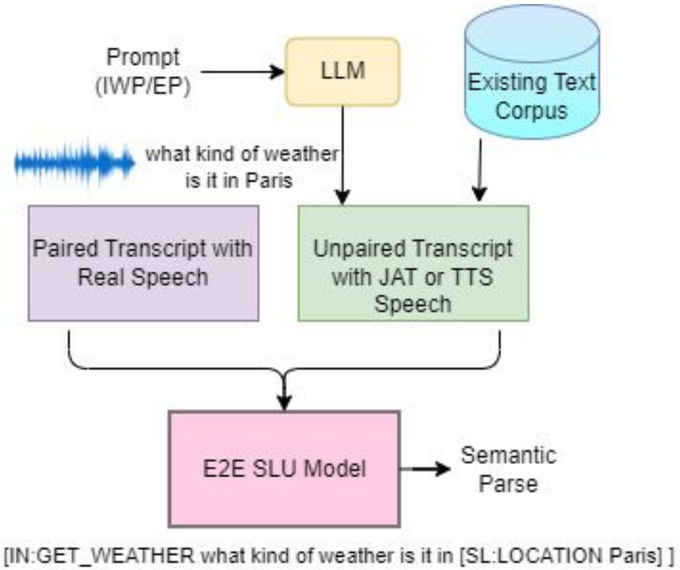
ASR data to Spoken Semantic Parsing

- Available: Speech → Text
- Compute: Semantic parsing from the transcript
  - Existing domains: use the existing semantic parsing models
  - New domains: prompt an LLM and ask it to generate the semantic parse
    - Intent-word-based prompting (IWP)
    - Exemplar-based prompting (EP)



NLU data to Spoken Semantic Parsing

- Available: Text → Semantic Parse
- Compute: Speech signal corresponding to the the input text by **Voicebox**



Experimental Settings

- STOP dataset: 100 hours of real speech for spoken semantic parsing (8 domains: alarm, event, messaging, music, navigation, reminder, timer, and weather, 28 unique intents, 82 slot types)
- Evaluation criteria:
  - Exact Match (EM)
  - EM (w/ ASR error)
  - EM (w/o ASR error)

Results: In domain

Table 1: *Comparing JAT and TTS as speech representations for unpaired text from ED and ND. Number of paired and unpaired utterances, and Exact Match (EM) is reported*

	Model	#Pair/#Unpair	EM	EM(No Err)	EM w/ Err
ED	Baseline	60.4k / 0	64.25	80.51	24.37
	w/ JAT	60.4k / 60.4k	66.92	83.90	25.25
	w/ TTS	60.4 / 60.4k	**67.05**	**83.88**	**25.80**
ND	Baseline	60.7k / 0	33.28	41.32	13.54
	w/ JAT	60.7k / 60.1k	57.74	73.34	19.50
	w/ TTS	60.7k / 60.1k	**63.95**	**80.70**	**22.88**
	Topline	120.9k / 0	67.67	84.52	26.34

Use of Voicebox based TTS data on an unseen domain

Table 4: *Using TTS to generate speech for LLama 2.0 text when unpaired text is in an unseen new domain*

Model	#Utts(Weather)	Weather EM	Overall EM
STOP 7 dom.	0	0	54.61
+ 3 real example-TTS	360	48.18	61.80
+ Exemplar LLama2-TTS	2,910	**50.82**	**62.29**
Topline: STOP Weather-TTS	2,910	63.80	66.33

Voicebox for SSP

- For unpaired text in new domains, TTS outperforms JAT by 6% absolute EM overall, with a gain of 30.6% EM over a paired baseline.
- With LLM-generated data and TTS, SSP can be improved by 1.4% EM and 2.6% EM absolute for existing and new domains, respectively.

Extensions of the Voicebox Model

- Reducing the need for a forced aligner
- More controllability
- Text description as prompt
    - Audiobox

Future Directions

- Long form speech generation without loss of speaker consistency
- TTS is one approach to use unpaired text data in speech but we still cannot fully rely on synthetic data (e.g. 10x synthetic data to match the real data)
    - How can we improve the generative models to make them closer real speech?
    - Can we use this “performance matching factor” as an evaluation metric for synthetic speech?

Links and Other Resources

- https://lsari.github.io/voicebox_talk_may_2025/

Professional Activities

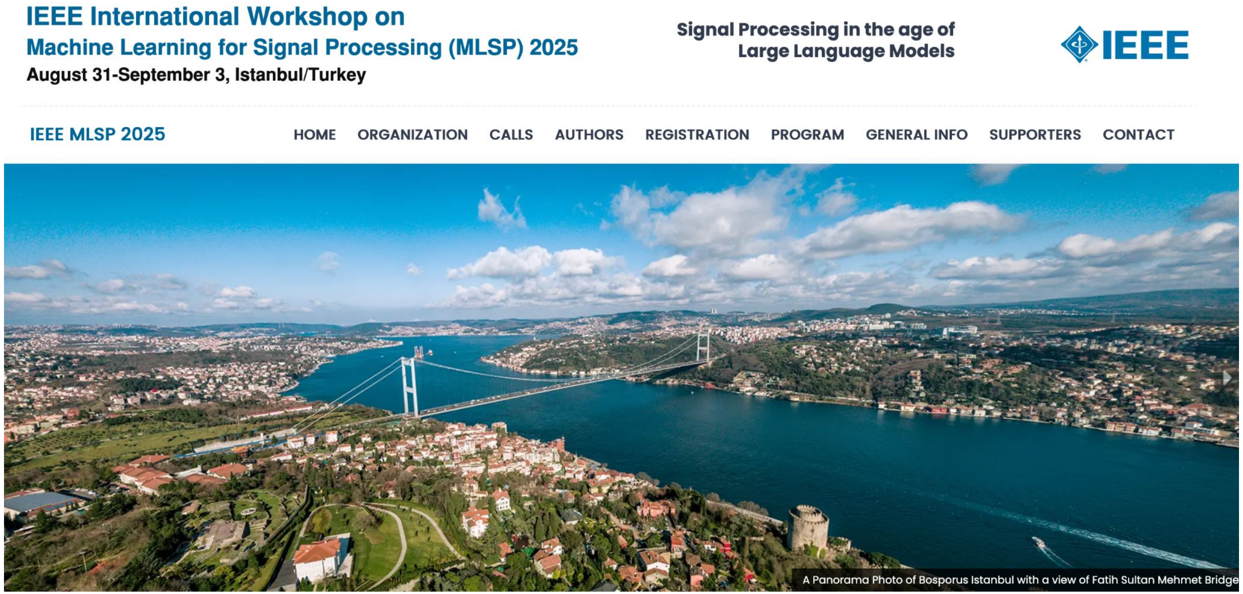
Young Female* Researchers in Speech Workshop (YFRSW) 2025 (The Netherlands)

- A satellite event of **Interspeech** since 2016
- Application deadline has passed for this year
  - Current female undergraduate and master's students, please keep following us for future years
- Current female PhD students, we are looking for volunteers for the PhD student panel discussion!
- Sponsorship opportunities are still available for 2025 (both industry and academia)!

**The workshop is open for marginalized genders, including women, as well as non-binary and gender non-conforming people who are comfortable in a space that is centered on women's experiences in the speech science and technology community. We aim to offer an inclusive and accessible program. If you are unsure if this workshop is for you, please don't hesitate to reach out to us!*

IEEE MLSP 2025 (Istanbul, TR)

- **Theme: Signal Processing in the Age of LLMs**
- Sponsorship opportunities are available!
- Registrations will open soon!



IEEE ASRU 2025 (Hawaii, USA)

- **Call for demos/system/data papers!!!**
- **Deadline: June 25, 2025**
- NEW this year:
  - Will be part of IEEEXplore proceedings
  - Short-paper format (3 pages)
  - Single-blind review

ASRU 2025

Workshop on Automatic Speech Recognition and Understanding

6-10 December 2025 || Honolulu, HI, USA

CALL FOR DEMO/SYSTEM/DATA PAPERS

IMPORTANT DATES

June 25, 2025*
Demo & challenge papers due

August 6, 2025*
Paper notification of acceptance

(All midnight AoE)

CHALLENGE, SPECIAL SESSION, AND DEMONSTRATION CHAIRS

Shinji Watanabe
CMU, US

Jingdong Chen
NPU, CN

Omid Sadjadi
AWS AI, US

Leda Sari
Otter AI, US

IEEE Signal Processing Society

IEEE

ASRU 2025 invites submissions to a dedicated track for **demonstration, system, and data description papers**.

In previous ASRU workshops, the demo track was held as a separate session without inclusion in the official workshop proceedings or IEEE Xplore. However, with the growing impact of automatic speech recognition and understanding systems in real-world applications—and the emergence of large language models—this category of work has become increasingly important to our community. This is especially true for researchers in industry, and the redesigned demo track aims to encourage their participation by providing a platform to share valuable research and development outcomes with the broader ASRU community.

In recognition of this trend, ASRU 2025 is **elevating the demo track to an official, peer-reviewed track**. Accepted papers will be included in the **workshop proceedings and published in IEEE *Xplore***, alongside regular papers.

We welcome submissions in the following categories:

- Descriptions of speech and language processing systems and demonstrations
- Applications of speech and language technologies
- Development of spoken and multimodal language models
- Software or toolkits for speech and language processing
- Optimization techniques for large-scale training and complex system development
- Methods for efficient inference and deployment
- Collection, description, and curation of datasets for speech, language, and multimodal data
- Tools for system visualization and evaluation
- Benchmark creation, description, and evaluation

We look forward to your contributions to this exciting and evolving area of research.

Paper Format:

- **Review Process:** Single-blind
- **Length:** Up to 3 pages for main content, plus 1 additional page for references
- **Template:** Same formatting guidelines as regular ASRU papers

2025.IEEEASRU.ORG