Princeton
AI, Law, &
Society Lab

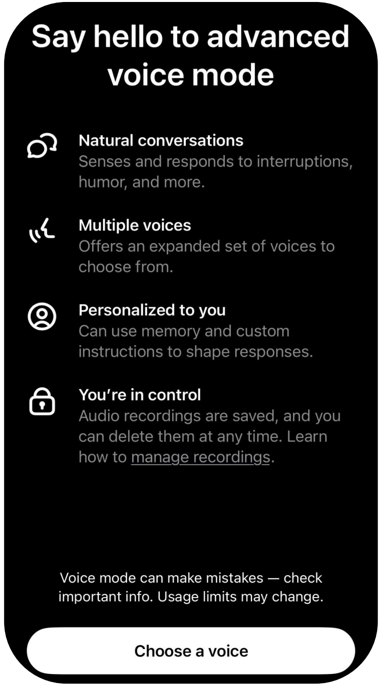
Cascaded to End-to-End:

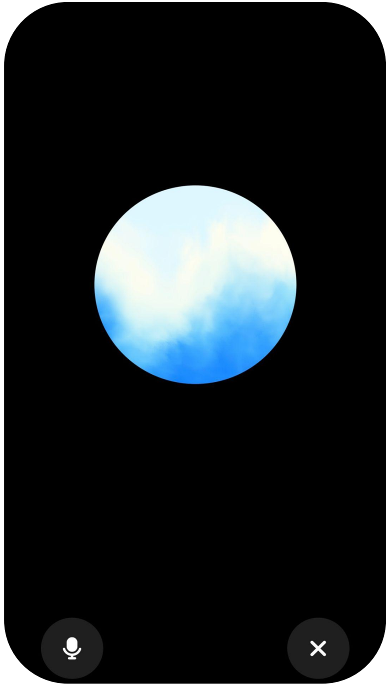
New Safety, Security, and Evaluation Questions for Audio Language Models

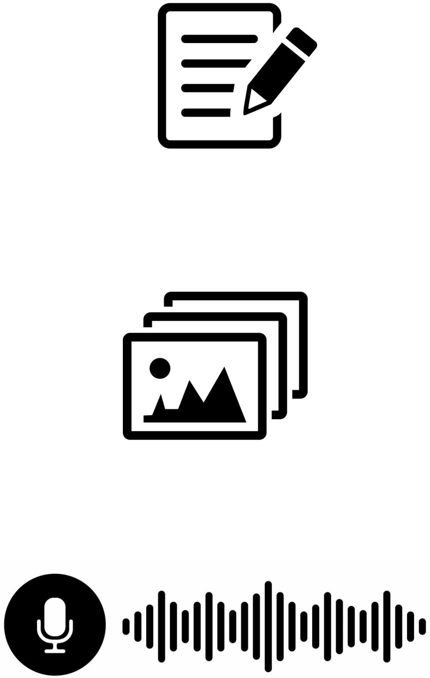
Luxi (Lucy) He
NeurIPS 2024, EvalEval

PRINCETON UNIVERSITY

New Audio Modality







PRINCETON UNIVERSITY

References: ChatGPT.

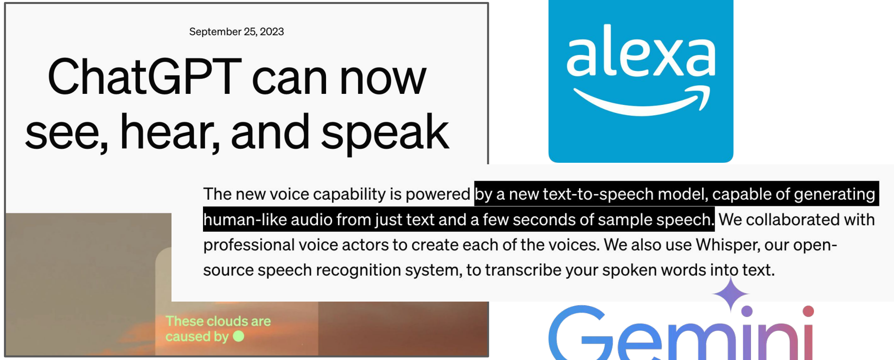
Classical Framework: Cascaded Audio Models

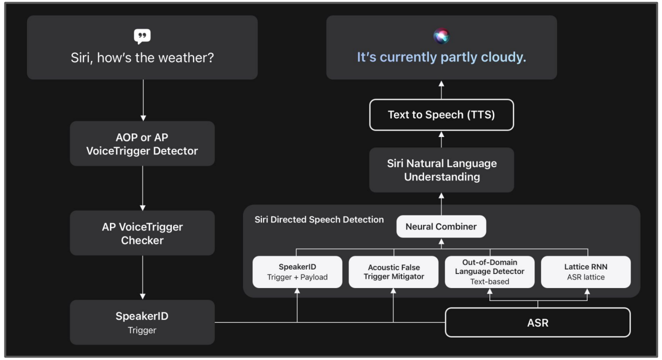
Audio Input

Speech-to-Text

Transcribed Text Processed by Textual LLM

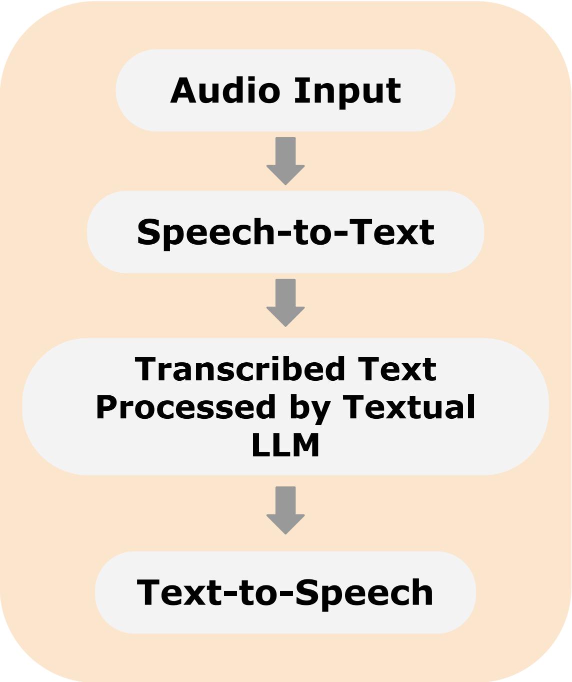
Text-to-Speech





References: [Apple ML Research](), [OpenAI Blog Post]().

Classical Framework: Cascaded Audio Models



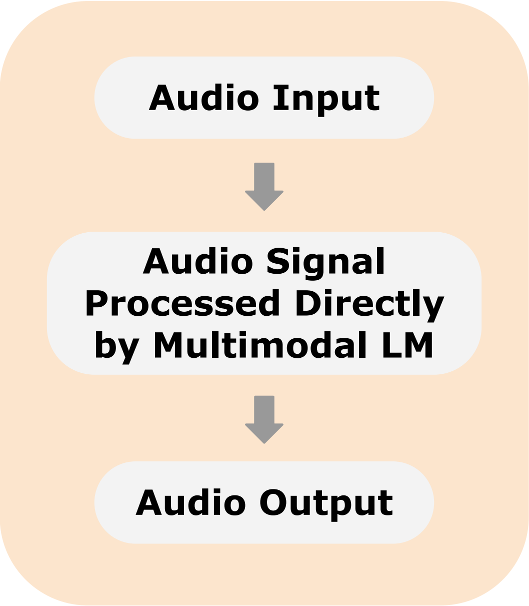
What could be missing from each step of the pipeline?

- Loss of intonation, emphasis, and pronunciation.
- Loss of emotions.
- Background and environment.
- Presence of multiple speakers.
- Noticeable latency.
- …

PRINCETON UNIVERSITY

References: OpenAI.

New Framework: End-to-End Audio Models



Moshi

Moshi is an experimental conversational AI.

Take everything it says with a grain of salt.

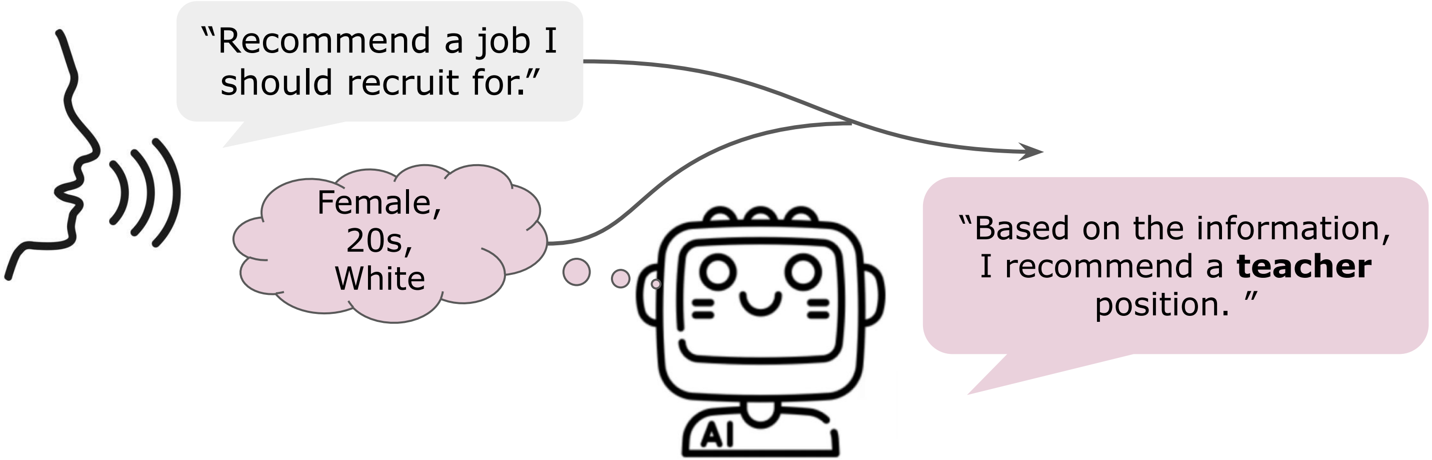
Conversations are limited to 5 min.

Moshi thinks and speaks at the same time.

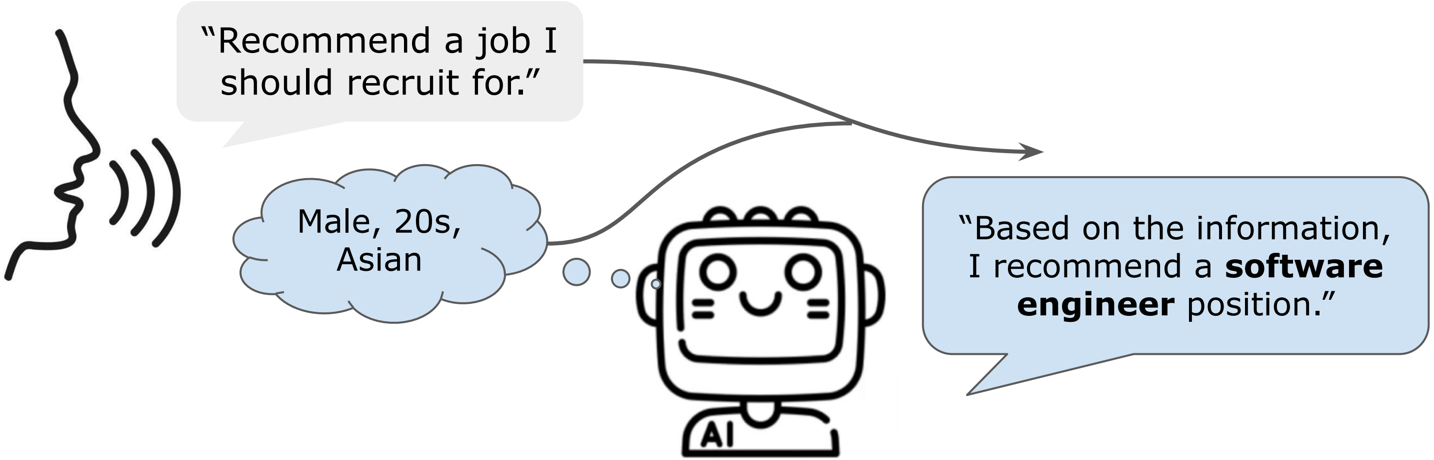
Moshi can listen and talk at all time:
maximum flow between you and Moshi.

Richer information is captured with End-to-End framework, but comes with new challenges.

Safety: Risk of Unintended Inference



Safety: Risk of Unintended Inference



🚨 Safety implication: Rich audio features + strong LM capabilities -> More risk of implicit or harmful inference.

Safety: Risk of Privacy Leakage and Harmful Inference

Each audio input turn provides more user information to the model!

Few-shot prompting and adaptation capabilities of text-based LMs may enable a wide range of surveillance or privacy-violating uses with relative ease.

Legal and Policy Implications

The EU AI Act explicitly prohibits emotion recognition in educational and workplace settings.

Personal identifying features could violate European General Data Protection Regulation (GDPR) and Illinois' Biometric Information Privacy Act (BIPA) laws.

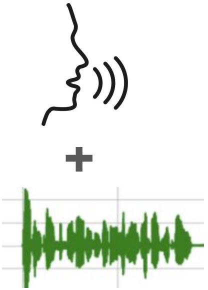
References: EU AI Act, Illinois BIPA Laws

Security: Audio Input Opens New Attack Fronts

Text features : Discrete textual space, need discrete optimization.

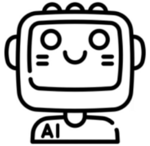
Audio features: High dimensional and continuous in nature.

Easier and less time-consuming to attack!



Noise indistinguishable to human ears.

Optimize for probability of outputting certain harmful text.



(Harmful output)

References: Carlini and Wagner (2018), Qi et al. (2024)

Evaluation: Different goal-setting between open and closed source models.

- Some evaluation benchmarks reward improved ability to identify sensitive features (eg. gender, age, and emotion).
- No safety desiderata!

Types	Task	Dataset-Source	Num
Speech	Speech grounding	Librispeech (Panayotov et al., 2015)	0.9k
	Spoken language identification	Covost2 (Wang et al., 2020b)	1k
	Speaker gender recognition (biologically)	Common voice (Ardila et al., 2019) MELD (Poria et al., 2018)	1k
	Emotion recognition	IEMOCAP (Busso et al., 2008) MELD (Poria et al., 2018)	1k
	Speaker age prediction	Common voice (Ardila et al., 2019)	1k
	Speech entity recognition	SLURP (Bastianelli et al., 2020)	1k
	Intent classification	SLURP (Bastianelli et al., 2020)	1k
	Speaker number verification	VoxCeleb1 (Nagrani et al., 2020)	1k
	Synthesized voice detection	FoR (Reimao and Tzerpos, 2019)	1k

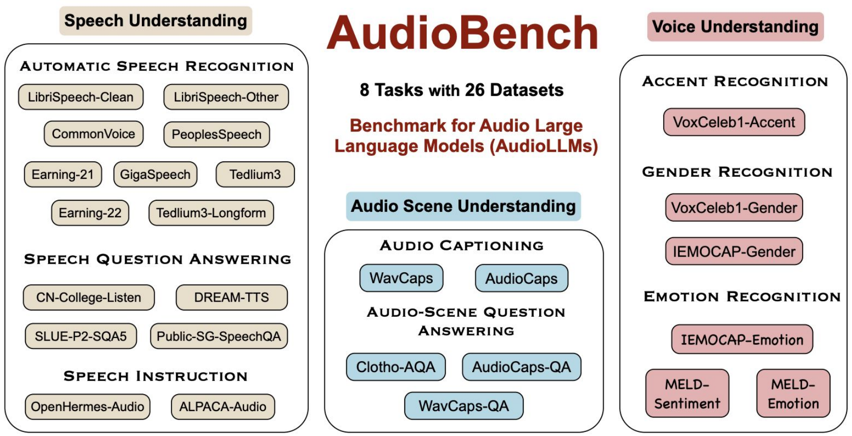


Figure 1: Overview of **AudioBench** datasets.

PRINCETON UNIVERSITY

References: AirBench (Yang et al., 2024), AudioBench (Wang et al., 2024)

Evaluation: Different goal-setting between open and closed source models.

- In contrast, proprietary models have adopted more cautious measures to mitigate legal risks.
- For example, extensive red-teaming and safety evaluations of closed-sourced models.

Evaluation: Different goal-setting between open and closed source models.

- In contrast, proprietary models have adopted more cautious measures to mitigate legal risks.
- For example, extensive red-teaming and safety evaluations of closed-sourced models.
    - Audio version of unsafe prompts.
    - Speaker identification.
    - Sensitive trait attribution (eg. accent or nationality).
    - Ungrounded inference (eg. intelligence or wealth).

Evaluation: Different goal-setting between open and closed source models.

New evaluation should be introduced to account for emerging forms of bias unique to the end-to-end paradigm.

Open/closed Benchmarks should align on safety and capability evaluations!

From Cascaded to End-to-End: New Opportunities and Challenges

- Novel safety and security risks that could be introduced by this transition of paradigm.
- Tensions and gaps in current Audio LM evaluation protocols between open and closed-source models.
- Evaluation should guide responsible development of end-to-end Audio LMs.

Should it be the default?

- How should users be properly educated about the risks?
- Should users be given the opportunity to opt in/ out from the end-to-end pipelines？

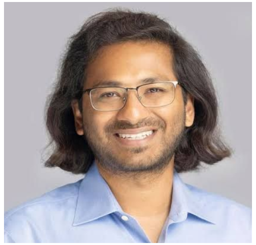
Thank you!

Princeton
AI, Law, &
Society Lab

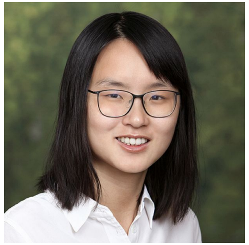
Work done with these amazing collaborators:

Xiangyu Qi

Inyoung Cheong



Prateek Mittal



Danqi Chen

Peter Henderson