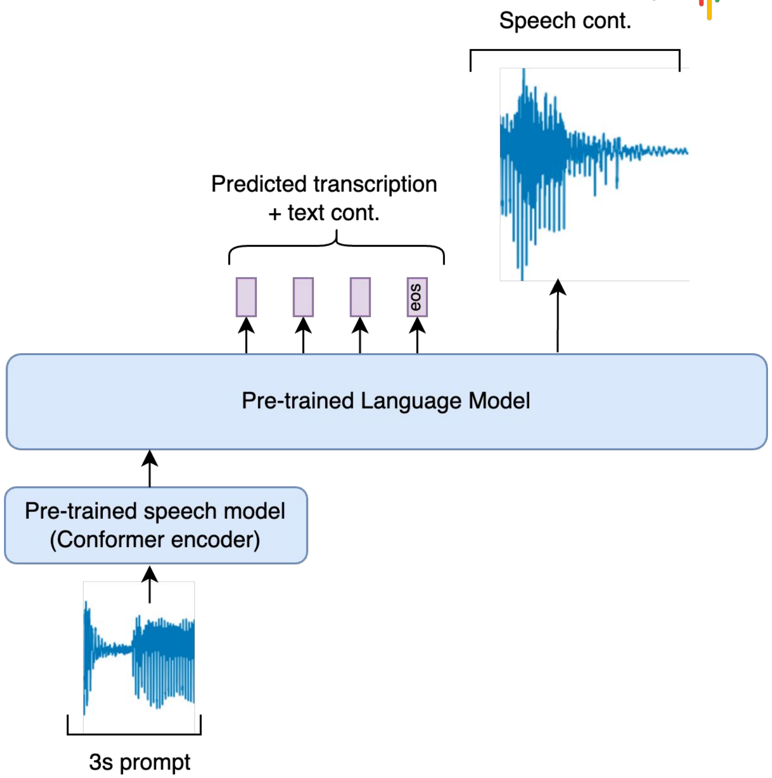
Spoken Question Answering and Speech Continuation Using Spectrogram-Powered LLM

Eliya Nachmani, Alon Levkovitch, Roy Hirsch, Julian Salazar, Chulayuth Asawaroengchai, Soroosh Mariooryad, Ehud Rivlin, RJ Skerry-Ryan, Michelle Tadmor Ramanovich

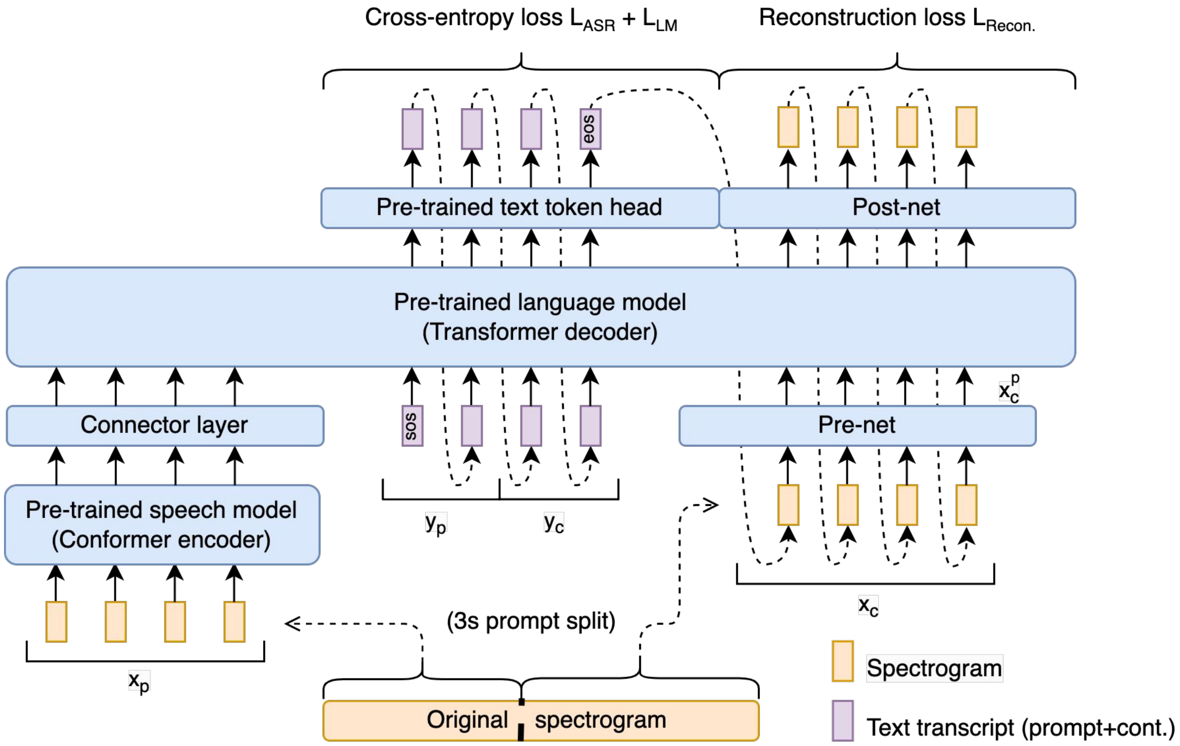
Google Research

SPECTRON

- ❑ Directly process spectrograms - input and output.
- ❑ Leverage the capabilities of a pre-trained speech encoder.
- ❑ Retain generative natural language abilities from a pre-trained LLM.
- ❑ End-to-end training.
- ❑ End-to-end spoken question answering.

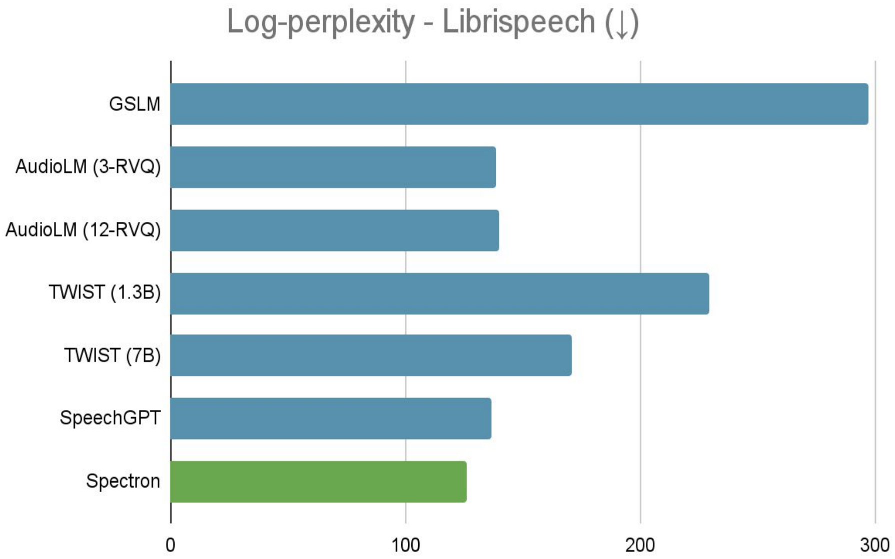

Speech cont.
Predicted transcription + text cont.
eos
Pre-trained Language Model
Pre-trained speech model (Conformer encoder)
3s prompt

Architecture



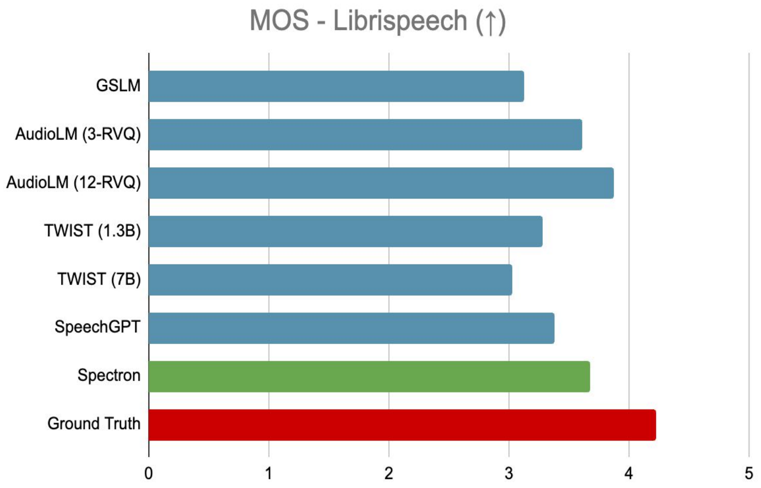
Speech Continuation - Semantic Quality

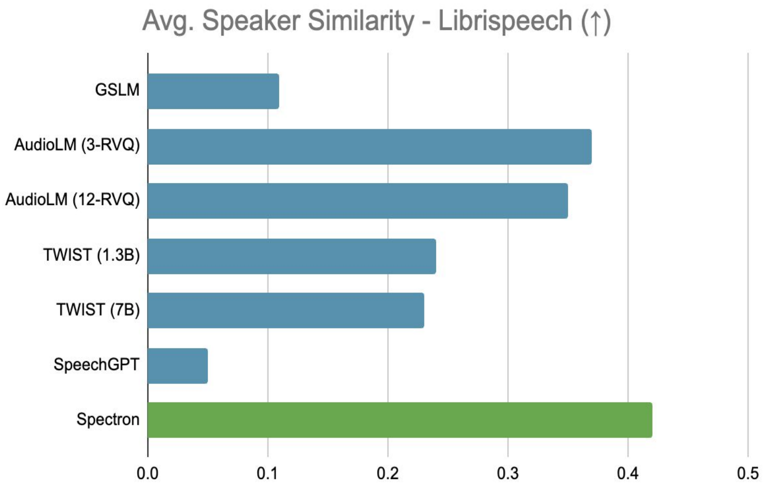
- ❑ Semantic Quality is measured by Log-Perplexity.
- ❑ Log-Perplexity computed using GPT2 on transcripted continuations.



Speech Continuation - Acoustic Quality

- ❑ Acoustic Naturalness is measured by Mean Opinion Score (MOS).
- ❑ Speaker fidelity is measured by Speaker Similarity.

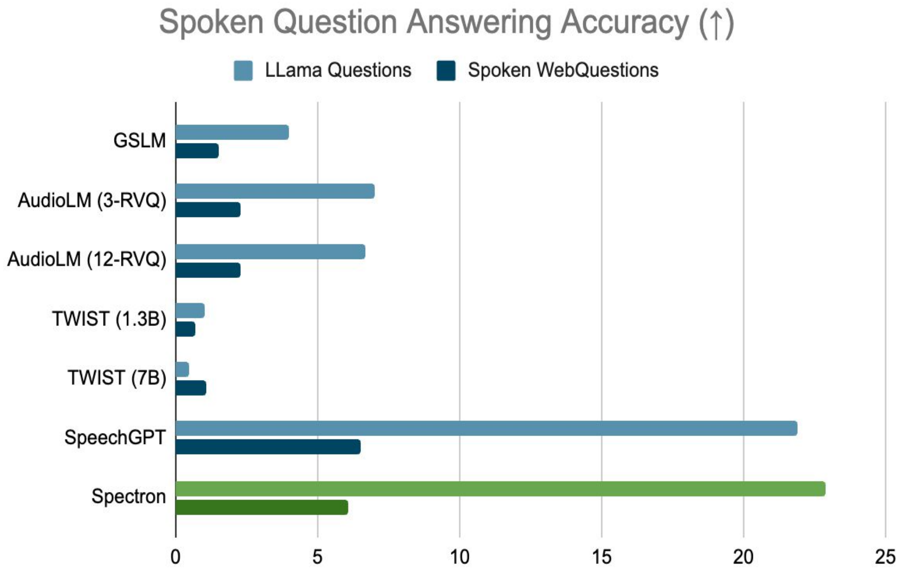




Spoken Question Answering

- Spoken WebQuestions: Spoken questions generated WebQuestions.

- LLama Questions: open-domain QA dataset that we had gathered from LLama-7B model.



Audio Samples

Promt	Generated Continuation
🔊	🔊
🔊	🔊
🔊	🔊