Speech-Native Foundation Model Architecture

Audio is tokenized via an audio encoder (e.g. HuBERT) and combined with text instructions. The LMM autoregressively generates interleaved text and audio tokens; audio is passed through a vocoder for synthesis.