Real-Time Spoken Instruction Following and Translation in Ugandan Languages

Benjamin Akera, Tim Wenjie Hu, Patrick Walukagga, Evelyn Nafula Ouma,
Gilbert Yiga, Engineer Bainomugisha, Ernest Tonny Mwebaze and John Quinn

Sunbird AI, Kampala, Uganda
Corresponding Authors: (info, bakera)@sunbird.ai

Abstract

Many languages are predominantly spoken rather than written, and to bring the benefits of LLMs to speakers of these languages, it is essential that models cater to the voice modality. The typical approach is to cascade ASR, LLM and TTS models together, though this results in systems with high latency, making them unsuitable for natural, real-time interaction. We describe results on taking the encoder part of a Whisper-based model trained to recognise ten languages common in Uganda, and using the Ultravox architecture to project its output directly to the input embedding space of a text model based on Qwen 3 32B, also trained to have comprehension of those languages. The result is a speech LLM with high accuracy and very low latency. For most spoken prompts, we can begin streaming a text response within as low as 50 ms, and a speech audio response within around one second, making real-time spoken interaction with an LLM possible for the first time in these languages. The model is available open source on Hugging Face.

1 Introduction

Speech LLMs are of particular significance for languages which are primarily spoken rather than written. When a practical application calls for speech input to an LLM in a low-resource language, the most common method is to separately train a speech recognition (ASR) model and optionally a machine translation (MT) model, then to simply chain them together such that text is input to the LLM. The drawbacks of this approach are that (1) latency is high, as the ASR and MT models must in turn complete generating tokens before the LLM can begin; (2) errors are amplified along the cascade of models, often leading to nonsensical output in the case of low-resource languages.

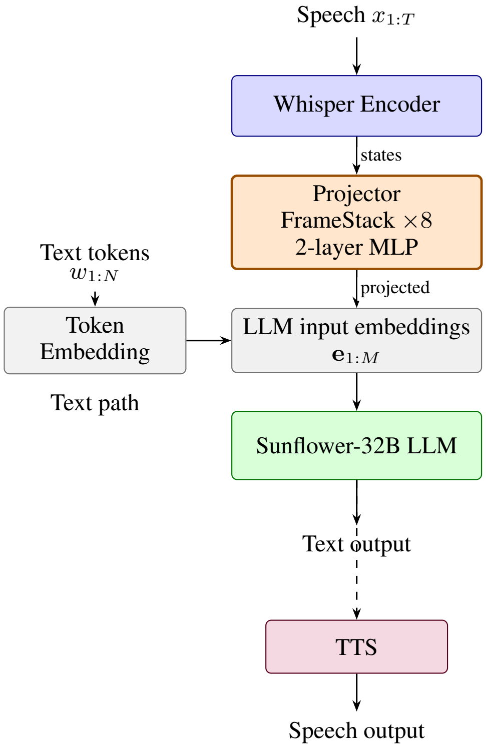

Figure 1: Ultravox-style speech LLM architecture: Whisper encoder states are mapped by a trainable projector into the text model's input embedding space, so that speech and text prompts can be concatenated. TTS is optional.

In this work, we apply recent findings in coupling ASR and LLM models at a deeper level than text tokens to a low resource setting, specifically Ugandan and East African languages. The Ultravox architecture (Fixie AI, 2024) is a linear multimodal projector that takes a high-dimensional embedding from the encoder part of an ASR model and maps it to the input dimension of an LLM. The resulting fused speech LLM model is trained so that the output is the same regardless of whether prompts are

Proceedings of the 7th Workshop on African Natural Language Processing (AfricaNLP 2026), pages 204–210
March 28, 2026 ©2026 Association for Computational Linguistics

input as text or read out as speech.

Starting by training both an ASR model and multilingual text LLM for our languages of interest (Luganda, Runyankore, Acholi, Ateso and Lugbara) we then apply this multimodal projector approach to create a single speech model that can respond to audio prompts directly, without needing a separate ASR step. We demonstrate that this results in a model which is very low latency and has high accuracy for speech translation and instruction following.

The contributions of this paper are:

- We describe the development of the first speech LLM model, to our knowledge, capable of real-time translation and responses to spoken instructions in some African languages.
- Rather than adding an Ultravox adapter to an off-the-shelf LLM such as Gemma or Llama, we demonstrate how this can be applied to a new multilingual LLM and ASR models separately trained to have comprehension of several new languages.
- We extend the training pipeline for Ultravox, specifically on the use of Group Relative Policy Optimisation for post-training the LLM.

Furthermore, we make the model publicly available, which we hope will make progress towards important speech applications e.g. in education, healthcare and real-time translation.

2 Related work

Speech Language Models: Recent work has moved beyond cascaded ASR-LLM pipelines towards end to end speech understanding. Qwen Audio (Chu et al., 2023) and Qwen2-Audio (Chu et al., 2024) extend the Qwen LLM family with audio encoders demonstrating strong performance on speech recognition, translation and audio reasoning tasks across several languages. Moshi introduces a fully duplex spoken dialogue system capable of simultaneous listening and speaking with latencies under 200ms (Défossez et al., 2024). Most relevant to our work, Ultravox (Fixie AI, 2024) projects Whisper encoder outputs directly into an LLM's embedding space via a lightweight multimodal projector, avoiding explicit transcription while preserving the flexibility to use any open-weight LLM backbone. We adopt this architecture for its modularity: the audio encoder and LLM can be trained independently and then aligned through the projector alone.

Low Resource and African Language ASR: While large scale models like Whisper (Radford et al., 2023) and MMS (Pratap et al., 2024) include some African languages, coverage remains sparse and performance lags significantly behind high resource languages. Recent efforts have produced targeted datasets such as SALT (Owomugisha et al., 2023; Nakatumba-Nabende et al., 2024), Commonvoice (Ardila et al., 2020), Flores and the NLLB Project (Pratap et al., 2024) enabling fine-tuned models with substantially improved accuracy. However, these advances in ASR have not yet been coupled with LLMs capable of understanding and generating text in these languages. To our knowledge, no prior work has applied speech-native LLM architectures to regional African languages. Existing Speech LLMs focus predominantly on high resource languages or rely on translation to English as an intermediate step. We bridge independently trained components: a Whisper encoder fine-tuned on 10 Ugandan languages and an LLM with native comprehension of those languages, enabling direct spoken interaction without cascaded errors or translation bottlenecks. Our system achieves accuracy comparable with the cascade approach, but with extremely low latency of down to 50ms to first token.

3 Model

Our architecture (Figure 1) follows the Ultravox design (Fixie AI, 2024), comprising three components: (1) an audio encoder that converts speech to continuous representations, (2) a multimodal projector that maps these representations into the LLM's embedding space, and (3) a large language model that generates text conditioned on the projected audio. Following recent taxonomy (Arora et al., 2025), our system can be characterized as a *speech-aware language model*: an LLM augmented with a speech encoder that processes audio directly, rather than relying on cascaded ASR.

Audio encoder. We use the encoder from a Whisper Large V3 model (Radford et al., 2023) fine-tuned on ten Ugandan languages. Given input audio, the encoder produces a sequence of hidden states $\mathbf{h} \in \mathbf{R}^{T\times 1280}$, where T depends on audio duration. The decoder part of the Whisper model is discarded; the larger multilingual LLM that we add below becomes the new decoder. The encoder takes 16 KHz audio, used to compute an 80-dimensional log Mel spectrogram on 25 ms windows with a stride of 10 ms. The Whisper encoder layers downsample the time dimension by a factor of 2, so that each frame of the resulting encoding represents 20 ms of audio.

Multimodal projector. The projector bridges the audio and text representation spaces. Following Ultravox, we first apply a stacking operation that concatenates 8 adjacent frames, reducing sequence length to 6.25 frames per second while increasing dimensionality to $1280 \times 8 = 10240$. This is critical for efficiency: raw encoder outputs would otherwise produce prohibitively long sequences for the LLM's attention. The stacked features pass through a two-layer MLP with SwiGLU activation and RMS normalization, outputting embeddings matched to the LLM's hidden dimension.

The projector is trained to minimize cross-entropy loss on transcription tokens. Additionally, we employ KL divergence to encourage projected audio embeddings to match the text embeddings of equivalent transcriptions, ensuring modality alignment. The aim of this training is that the LLM should produce similar output logits for a given input prompt wording, regardless of whether those words are in text or spoken form.

Language model. We use Sunflower-32B (Akera et al., 2025), based on Qwen 3 32B and adapted for Ugandan languages through continued pretraining and instruction tuning on a corpus of Ugandan language data. During inference, projected audio embeddings replace placeholder tokens in the input sequence, and the LLM generates text autoregressively.

Text-to-speech. Text-to-speech output is optional, but preferred to enable fully spoken interaction with the model. We use a version of SparkTTS fine-tuned to produce speech in Ugandan English, Luganda, Acholi, Lugbara, Runyankole and Ateso, and then apply inference optimisations to stream back audio sentence-by-sentence with latency of a few hundred milliseconds.

4 Training

Each component is trained independently before projector alignment.

4.1 Audio Encoder

The Whisper Large V3 model was fine-tuned on the SALT dataset (Owomugisha et al., 2023), which provides transcribed speech across ten languages common in Uganda: English, Luganda, Acholi, Lugbara, Ateso, Runyankole, Lumasaba, Lusoga, Swahili, and Kinyarwanda. Training incorporated language-specific tokens, enabling language-aware recognition without explicit identification as a preprocessing step. The resulting model achieves 1.8% WER on English, 14.2% on Luganda, and 11.1% on Kinyarwanda. A variant trained on approximately 1,400 hours of Kinyarwanda speech achieved 7.1% WER, placing first in the Digital Umuganda ASR competition.[1]

4.2 Language Model

Sunflower-32B (Akera et al., 2025) builds on Qwen 3 through continued pretraining on approximately one billion characters of Ugandan language text, including digitized books, transcripts, and parallel corpora across 31 languages. The model then undergoes supervised fine-tuning with LoRA for instruction-following, emphasizing translation tasks, followed by reinforcement learning via Direct Preference Optimization to reduce hallucinations. On translation benchmarks, Sunflower-32B achieves state-of-the-art performance in 24 of 31 Ugandan languages, outperforming GPT-4o and Gemini 2.5 Pro on both xx→eng and eng→xx directions.

4.3 Projector Alignment

With the encoder and LLM frozen, we train only the projector on paired speech-transcription data from SALT. Training follows a chat-based format where each example is structured as:

```
System: You are an ASR assistant that
transcribes speech
User: Transcribe in {lang}: <|audio|>
Assistant: {transcription}
```

We modify the Ultravox training recipe to use a specific prompt for translation ("Translate to [language]: [text to be translated]") which matches the Sunflower LLM instruction-tuning templates. This ensures the LLM recognizes the task framing from its prior training, reducing the burden on the projector to learn both modality alignment and task specification simultaneously.

[1] https://www.kaggle.com/competitions/kinyarwanda-automatic-speech-recognition-track-b

The <|audio|> placeholder is replaced by projected audio embeddings during the forward pass. We use language-aware prompts, inserting the detected language name (e.g., “Luganda”, “Acholi”) to provide the model with explicit language context.

Loss is computed only on the assistant response tokens; system and user tokens are masked. This focuses learning on the transcription task rather than prompt reconstruction. Audio samples are filtered to durations between 0.5 and 25 seconds.

We train for 2 epochs with batch size 4 and gradient accumulation over 8 steps (effective batch size 32). We use AdamW with learning rate 2×10^{-5}, cosine decay schedule, and 500 warmup steps.

Optionally, we apply LoRA to the encoder and LLM for joint fine-tuning: rank 32 for the audio encoder attention layers, and rank 64 for the LLM attention and feed-forward projections. This allows limited adaptation of the frozen backbones while keeping computational costs manageable.

4.4 LLM Post-training

The instruction-tuned LLM (Section 4.2) occasionally responds conversationally to prompts rather than following the literal task instruction. For example, given “Transcribe: Tell me a story about a goat”, the model may generate a story rather than echoing the input text. This behaviour, while inherent for question answering, degrades performance for transcription and translation tasks where verbatim output is required. We address this through two-stage post-training. First, we perform supervised fine-tuning on examples pairing prompts with the expected verbatim output. This establishes the basic transcription behavior. Second, we apply Group Relative Policy Optimization (GRPO) (Guo et al., 2025) with a reward function based on the negative edit distance between the model output and expected transcription. For each prompt, we generate multiple completions and reward those with lower Levenshtein distance to the target, penalizing deviations up to a maximum of 100 characters. This reinforcement learning phase strengthens the model’s adherence to literal transcription even when the input resembles a question or instruction.

5 Results

We evaluate on transcription (WER) and speech translation (BLEU) across the SALT test sets, covering 10 language pairs for translation (5 xx→eng and 5 eng→xx directions) and 8 languages for transcription.

Table 1: Speech translation quality (BLEU ↑). *Ours*: Ultravox model. *Cascaded*: Whisper transcription followed by Sunflower translation. *Text-only*: Sunflower translating ground-truth text (oracle upper bound).

Direction	**Ours**	**Cascaded**	**Text-only**
lug → eng	38.2	38.3	42.1
eng → lug	30.9	31.3	34.3
nyn → eng	22.8	24.0	26.0
ach → eng	19.9	24.0	24.0
lgg → eng	18.4	17.0	23.0
teo → eng	18.2	17.5	26.0
eng → nyn	17.4	18.0	18.5
eng → lgg	18.3	17.5	20.0
eng → ach	16.6	16.8	17.5
eng → teo	15.0	16.0	17.0
Average	21.6	22.0	24.8

Speech translation. Table 1 compares our Ultravox model against two baselines: (1) a cascaded system that transcribes with Whisper then translates with Sunflower, and (2) an oracle that translates from ground-truth text. Our model matches the cascaded baseline across most language pairs (Table 1), achieving 38.2 BLEU on Luganda→English versus 38.3 for the cascaded system. We can therefore match the translation accuracy of a cascaded system, but with very low latency, enabling real-time speech translation in these languages for the first time.

Instruction following. A key advantage of our architecture is that the model naturally generalizes beyond the transcription and translation tasks seen during projector training. When we omit any text instruction and prompt the model with speech audio alone, it defaults to conversational assistant behavior, enabling free-form voice interaction. Table 2 shows an example where the model answers a factual question in Luganda. This emergent capability arises because we train only on response tokens, allowing the LLM’s instruction-following abilities to transfer directly to the speech modality.

Transcription. Table 3 compares our model against the full Whisper encoder-decoder model trained on identical data. For pure transcription, our model underperforms Whisper, as expected: we discard Whisper’s decoder (optimized for transcription) in favor of a general-purpose LLM that enables broader capabilities. When we look at the pattern of errors, the higher resulting word error

Table 2: Example model response when given an audio prompt in Luganda. When prompted with audio and no text instruction, the model defaults to responding as a conversational assistant. The response is generated directly from the audio, without any separate ASR step.

Input audio:

"Yingini y'emmotoka ekola etya?"
(*How does a car engine work?*)

Response:

Yingini y'emmotoka ekola ng'efulumya amaanyi okuva mu mafuta oba amasannyalaze, nga gakyusa amafuta oba amasannyalaze okugafuula amaanyi agasobozesa emmotoka okutambula...
(*The engine of a car works by converting energy from fuel or electricity into mechanical energy, which is then used to power the wheels of the car...*)

Table 3: Transcription quality (WER ↓, median %). The Whisper values are from a standard encoder-decoder model trained on the same data. Our model trades transcription accuracy for flexibility to support different tasks and lower latency.

Language	Ours	Ours (GRPO)	Whisper
English	0.0	0.0	0.0
Luganda	22.6	16.7	5.0
Kinyarwanda	27.8	28.2	1.0
Acholi	41.0	36.9	16.8
Runyankole	43.5	37.5	19.2
Lugbara	55.5	45.8	15.8
Lusoga	58.3	50.0	28.6
Ateso	62.5	58.6	28.6

Table 4: Latency on A100-80GB. Time-to-first-token (TTFT) measures server-side delay from request receipt to first generated token. The cascaded system cannot produce output until Whisper completes, making true TTFT undefined. Excludes cold-start (∼8s).

System	TTFT	Streamable
Ours (p50)	55 ms	✓
Ours (p90)	61 ms	✓
Cascaded	>1 s	×

rates have two causes. First, the model is prone to replying conversationally when asked to transcribe speech containing a question or something that sounds like an instruction. Second, the model often paraphrases the meaning of what was spoken in the text, rather than a word-for-word verbatim transcription. Initial experiments we have carried out with Group Relative Policy Optimisation indicate that this gap can be reduced, which will be the focus of future work.

Latency. Table 4 reports inference latency measured on an A100-80GB GPU using vLLM. Our model achieves a median time-to-first-token (TTFT) of 55ms after warm-up, enabling responsive streaming output. The first request after cold-start incurs ∼8s latency due to model loading and KV cache initialization; subsequent requests benefit from prefix caching (hit rate >85%).

The cascaded baseline cannot stream output until Whisper completes full transcription. Whisper's encoder-decoder architecture processes the entire audio context non-causally, precluding incremental output (Radford et al., 2023). This architectural constraint results in latencies exceeding one second for typical utterances, an order-of-magnitude slower than our end-to-end model. This difference is critical for real-time applications such as live translation and voice assistants.

6 Limitations

Several limitations warrant discussion. First, transcription accuracy lags behind the standalone Whisper model, reflecting the trade-off between latency and fidelity when discarding the decoder. Potential mitigations include adding another speech encoder, e.g. based on Wav2Vec2, and concatenating the input embeddings together for a richer representation of the input audio. Second, the model occasionally paraphrases utterances or responds conversationally to questions rather than transcribing verbatim. These behaviors are inherited from the instruction-tuned LLM and can be partially mitigated through reinforcement learning, as shown in Table 3. Finally, our evaluation focuses on translation and transcription; the model's capability for more complex spoken instructions (e.g., multi-turn dialogue, reasoning over audio) remains unexplored.

Efficiency considerations. Our model requires an A100-80GB GPU for inference, which limits accessibility in resource-constrained settings. Several paths exist for improving efficiency:

1. *Quantization* via GPTQ or AWQ could reduce memory requirements to 20GB (4bit quantization) with minimal quality loss.

2. *Smaller backbones*: The modular architecture allows independent scaling of each component. Using Whisper Small or Medium alongside a 7B–14B LLM would reduce memory requirements from 80GB to under 20GB, with expected trade-offs in transcription accuracy for lower-resource languages.

3. *Distillation*: The speech projector could potentially be transferred to a smaller student LLM. We leave systematic evaluation of these efficiency trade-offs to future work, but note that our architecture's modularity makes such adaptations straightforward.

7 Conclusion

We presented the first speech-native large language model for Ugandan languages, combining a Whisper encoder fine-tuned on SALT with the Sunflower-32B language model via the Ultravox multimodal projector. Our model responds to free-form speech instructions and matches cascaded Whisper→LLM translation quality while reducing time-to-first-token from over one second to 55 milliseconds, enabling real-time streaming applications for the first time in these languages. Our evaluation focuses on single-turn transcription and translation; multi-turn dialogue, spoken question answering, and constrained instruction-following remain to be systematically evaluated, pending the development of purpose-built benchmarks for Ugandan languages. Nonetheless, qualitative evidence suggests that the LLM's conversational capabilities transfer effectively to the speech modality.

Future work will extend coverage to additional African languages, improve transcription fidelity through reinforcement learning, and reduce TTS latency to enable full-duplex conversation. We release our models to support further research on speech-native LLMs for low-resource languages.

Acknowledgments

We acknowledge the linguists, educators, and translators from Makerere University whose expertise made this dataset possible. We also thank the Ultravox team for their great open source work, and the anonymous reviewers whose comments improved the paper.

References

Benjamin Akera, Evelyn Nafula Ouma, Gilbert Yiga, Patrick Walukagga, Phionah Natukunda, Trevor Saaka, Solomon Nsumba, Lilian Teddy Nabukeera, Joel Muhanguzi, Imran Sekalala, and 1 others. 2025. Sunflower: A new approach to expanding coverage of african languages in large language models. *arXiv preprint arXiv:2510.07203*.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the twelfth language resources and evaluation conference*, pages 4218–4222.

Siddhant Arora, Kai-Wei Chang, Chung-Ming Chien, Yifan Peng, Haibin Wu, Yossi Adi, Emmanuel Dupoux, Hung-Yi Lee, Karen Livescu, and Shinji Watanabe. 2025. On the landscape of spoken language models: A comprehensive survey. *arXiv preprint arXiv:2504.08528*.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, and 1 others. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*.

Fixie AI. 2024. Ultravox: A fast multimodal llm for real-time voice. https://github.com/fixie-ai/ultravox. Accessed: 2025-12-19.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Joyce Nakatumba-Nabende, Claire Babirye, Peter Nabende, Jeremy Francis Tusubira, Jonathan Mukiibi, Eric Peter Wairagala, Chodrine Mutebi, Tobius Saul Bateesa, Alvin Nahabwe, Hewitt Tusiime, and 1 others. 2024. Building text and speech benchmark datasets and models for low-resourced east african languages: experiences and lessons. *Applied AI Letters*, 5(2):e92.

Isaac Owomugisha, Benjamin Akera, Ernest Tonny Mwebaze, and John Quinn. 2023. Multilingual model and data resources for text-to-speech in ugandan languages. In *4th Workshop on African Natural Language Processing*.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, and 1 others. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.