

Speechless: Training Voice Assistants Without Speech for Low Resource Languages

A novel approach for training voice assistants in low-resource languages without requiring speech data. This method aligns synthetic semantic representations with pre-trained models, enabling LLMs to understand spoken instructions during inference.



The Challenge: Speech Instruction Data Scarcity

Voice Assistant Growth: Voice assistants powered by LLMs need speech instruction data for training.

BUT:



Data Scarcity

Speech instruction data is scarce, especially for low-resource languages.



TTS Limitations

Low-resource languages often lack high-quality text-to-speech models.

Introducing Speechless

</>

Quantizer Training

Train a quantizer using ASR data to align semantic and text representations.



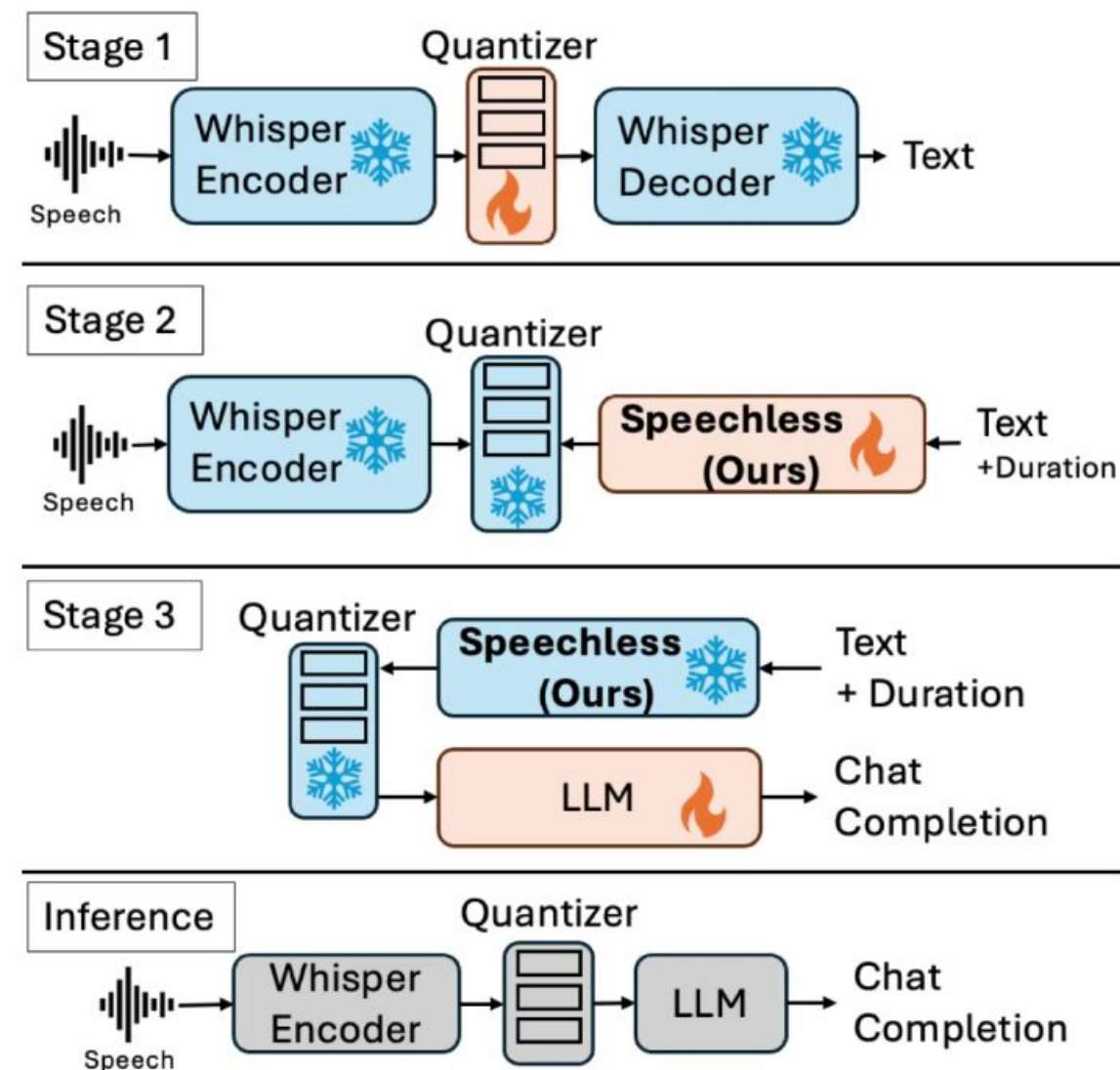
Speechless Training

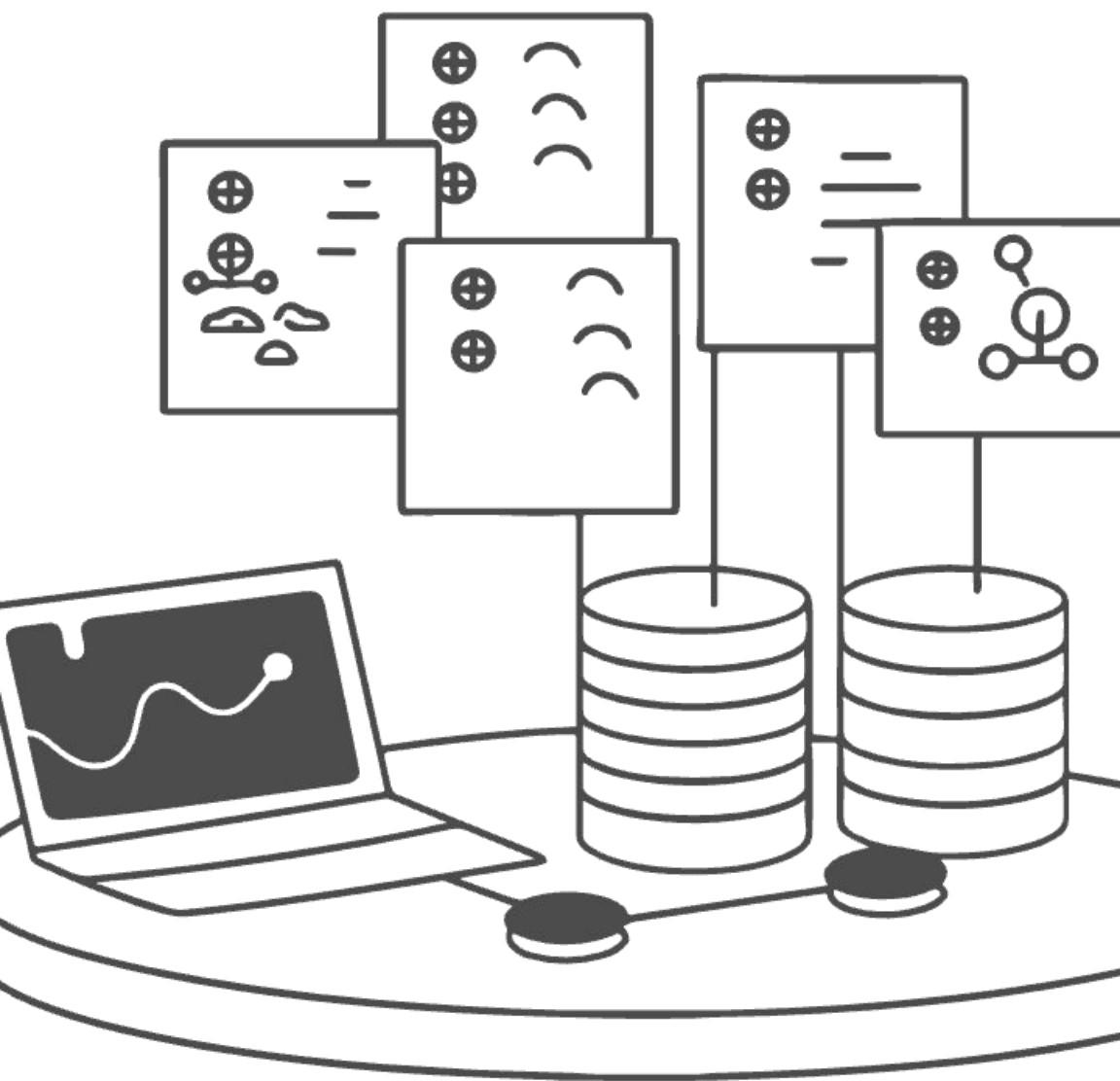
Train a model to map text to audio tokens without generating speech.



LLM Fine-tuning

Fine-tune LLM using audio tokens generated by Speechless.





Stage 1: Training a Quantizer

Residual Vector Quantizer

Transforms high-dimensional speech representations into discrete tokens while preserving meaning.

Iterative Refinement

Creates coarse representation first, then progressively refines through subsequent codebooks.

Expanded Capacity

Quadrupled codebook size from 512 to 2048 entries for low-resource languages.

Stage 2: Training Speechless



Text-to-Semantics Translation

Functions like a machine translation model



Token Generation

Generates semantic tokens similar to Whisper Encoder output



Length Management

Uses duration tokens to handle text-speech length mismatch

Stage 3: Training the LLM

Data Collection
Combined multiple instruction datasets
in English and Vietnamese

Fine-tuning
Applied standard speech instruction
tuning pipeline

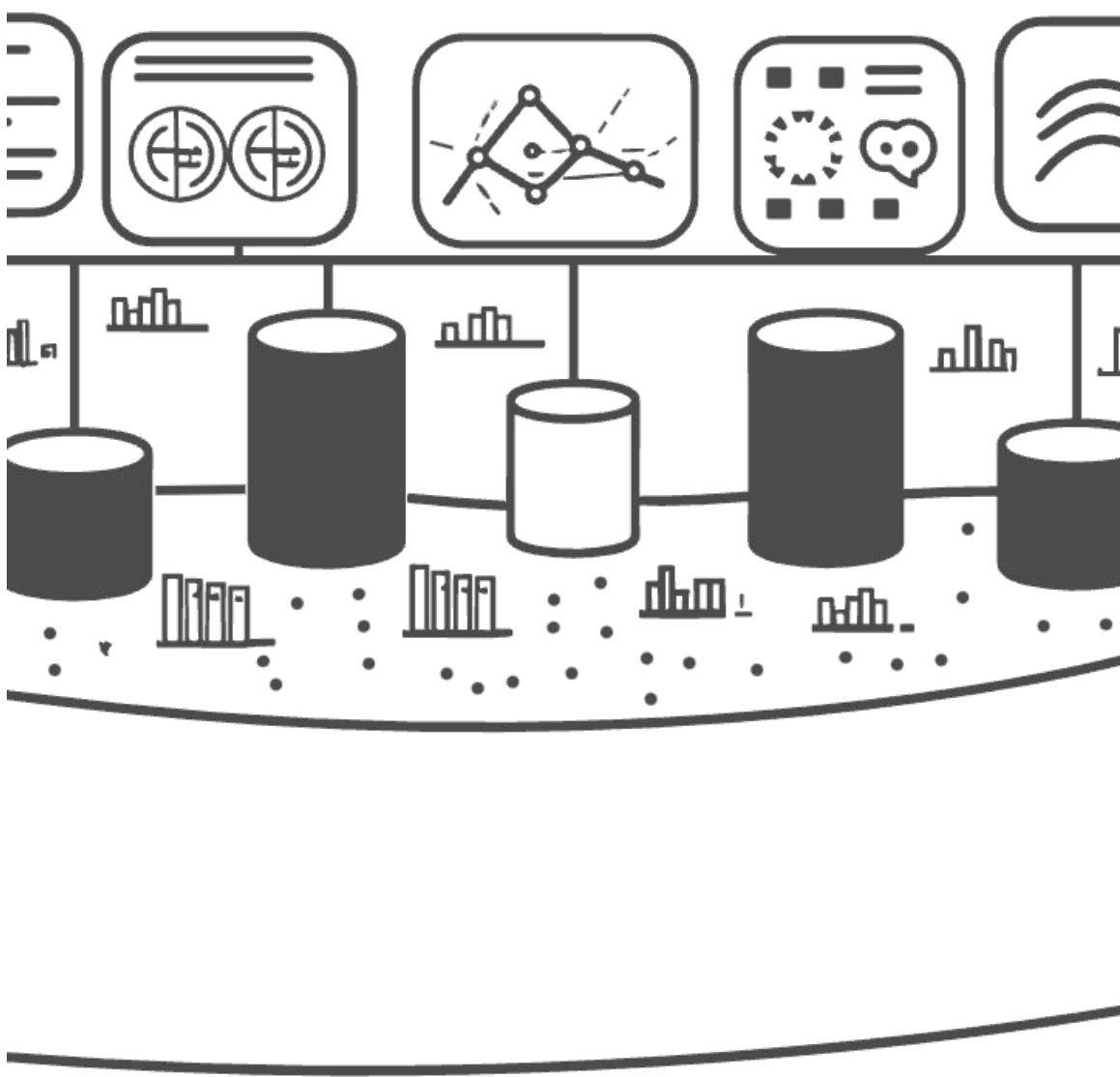


Data Filtering
Removed excessive prompts and
non-audible content

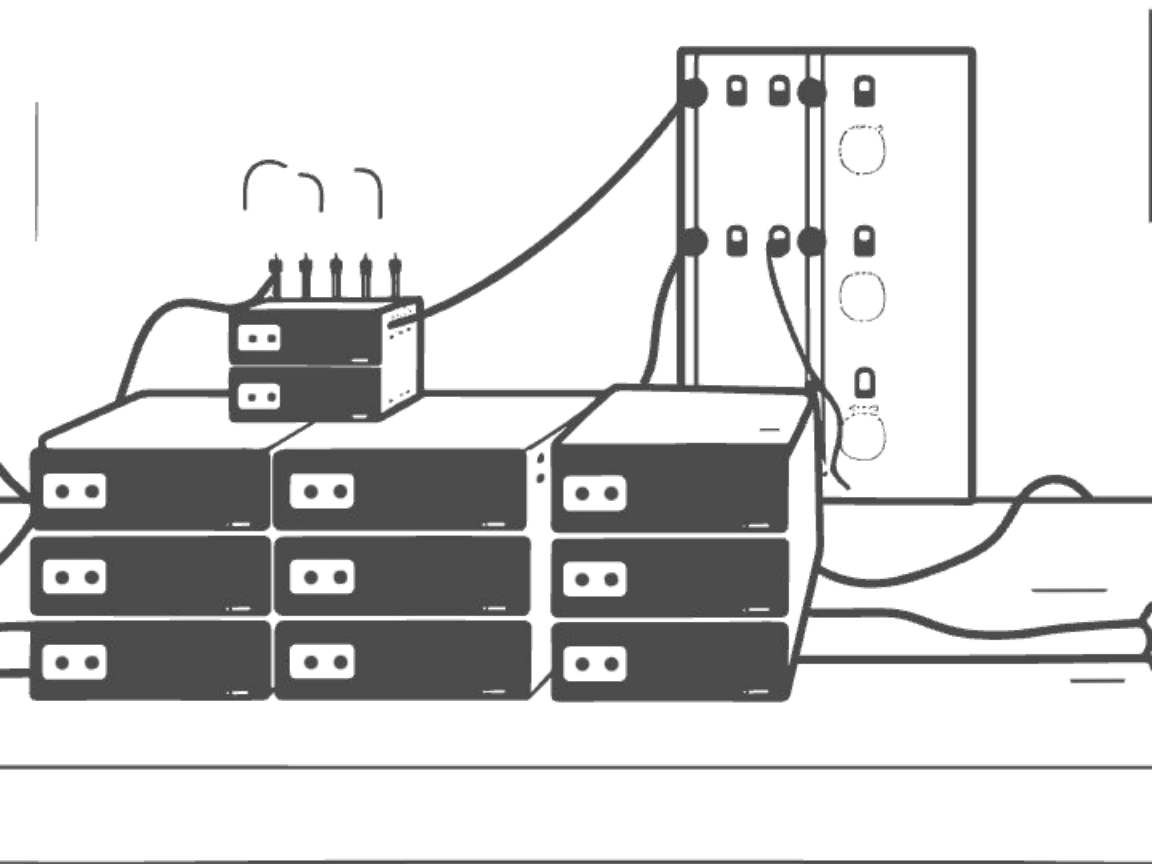
Tokenization
User turns tokenized to discrete
speech tokens using Speechless

Datasets Used

Stage 1	viVoice (868k Vietnamese utterances), LibriTTS-R (112k English samples)
Stage 2	880k samples from Vivoice, 112k from LibriTTS-R Clean
Stage 3	Pretrain: 880k from Vivoice, 112k from LibriTTS-R Clean, 2.4M from MLS Eng SFT: Ichigo (English), Sailor and Viettel x NVIDIA (Vietnamese)



Training Resources



104h

Stage 1

75h for Phase 1, 29h for Phase 2
on 8 A6000 GPUs

60h

Stage 2

Training on 6 A6000 GPUs with
batch size of 48 per GPU

280h

Stage 3

240h pretraining on A6000, 40h
fine-tuning on H100 GPUs

ASR Performance Results

Speechless shows competitive performance across datasets, with particularly strong results for Vietnamese compared to Whisper.

Table 1: All results are in percentages. Comparative analysis of model performance for general, noisy, and multilingual ASR using the LibrisSpeech (LS), VoiceBank+DEMAND (VBD), and CommonVoice (CV) datasets respectively. All results are derived from processed labels and predictions. Both labels and predictions are lower-cased and all special characters are removed.

Model	Config	LS test-clean		VBD clean		VBD noisy		CV En		CV Vi	
		CER	WER	CER	WER	CER	WER	CER	WER	CER	WER
Whisper (M)	Zero-shot (greedy)	1.21	2.85	1.45	4.99	2.13	6.17	4.21	5.98	15.00	25.43
	Zero-shot (beam-10)	0.92	2.51	1.33	4.80	1.94	5.91	3.21	5.22	13.72	24.18
Whisper (M) Quantized	Greedy Inference	3.45	6.74	3.27	7.12	9.32	15.76	4.33	7.27	28.11	36.53
	Beam-Search (n=10)	2.42	5.52	2.89	6.76	6.63	12.34	3.24	7.01	24.16	34.84
Speechless	Greedy Inference	2.47	4.65	1.01	2.32	-	-	3.54	8.03	2.69	5.90
	Beam-Search (n=10)	2.08	4.21	1.52	3.92	-	-	2.92	6.56	3.77	7.08

VoiceBench Results

Table 2: *VoiceBench [28] Results. These are results based on spoken questions and text answers. Experiments other than ours were performed by the VoiceBench authors. SD-QA and CommonEval have a human audio source, while the rest use Google TTS.*

Model Name	AlpacaEval	CommonEval	SD-QA	OpenBookQA	AdvBench
Baichuan-Omni-1.5	4.50	4.05	43.40	74.51	97.31
GLM-4-Voice	3.97	3.42	36.98	53.41	88.08
Qwen2-Audio	3.74	3.43	35.71	49.45	96.73
VITA-1.0	3.38	2.15	27.94	29.01	26.73
Moshi	2.01	1.60	15.64	25.93	44.23
Whisper-v3-turbo+LLaMA-3.1-8B	4.55	4.02	58.23	72.09	98.46
LLaMA-Omni	3.70	3.46	39.69	27.47	11.35
Speechless-llama3.1-8B-instruct (Ours)	3.86	2.51	35.00	26.15	62.88

Text Benchmark Performance

While Speechless shows strong voice instruction results, there's a performance trade-off when evaluating on text-only benchmarks.

Table 3: MMLU and VMLU Benchmarks. These are text-based benchmarks for comparing the performance degradation due to speech instruction tuning

Model Name	MMLU	VMLU
meta-llama3.1-8B-instruct	69.40	50.69
Speechless-llama3.1-8B-instruct	62.27	43.22

Limitations & Future Work

Text Performance

Noise Robustness

Language Expansion

Further exploration needed for highly noisy or diverse linguistic contexts.

Future work will focus on expanding to a broader range of languages and dialects.

Thank you for listening



Join our team now via:

