# VoiceTextBlender: Augmenting Large Language Models with Speech Capabilities via Single-Stage Joint Speech-Text Supervised Fine-Tuning



Yifan Peng<sup>1\*</sup>, Krishna C. Puvvada<sup>2\*</sup>, Zhehuai Chen<sup>2\*</sup>, Piotr Zelasko<sup>2</sup>, He Huang<sup>2</sup>, Kunal Dhawan<sup>2</sup>, Ke Hu<sup>2</sup>, Shinji Watanabe<sup>1</sup>, Jagadeesh Balam<sup>2</sup>, Boris Ginsburg<sup>2</sup> <sup>1</sup>Carnegie Mellon University <sup>2</sup>NVIDIA



### Background

Recent studies have augmented large language models (LLMs) with speech understanding capabilities, leading to SpeechLMs

- Strong performance on automatic speech recognition (ASR), translation (AST), and speech-based question answering (SQA)
- SpeechLMs are typically trained on speech supervised fine-tuning (SFT) data with multi-stage curriculum learning
- $\succ$  E.g., SALMONN adopts three-stage training: ASR pre-training  $\rightarrow$  instruction tuning  $\rightarrow$  activation tuning

Limitations

- Complicated multi-stage training pipeline
- Catastrophic forgetting of text-only capabilities



### **Proposed Method**

We propose VoiceTextBlender (VTBlender), a SpeechLM trained with single-stage joint speech-text SFT

## Learnable Frozen



Model architecture

Speech encoder: Canary (609M)

### Our method

- Streamlines the training process
- Enhances speech understanding performance
- Maintains original text-only performance

### Statistics of our training data mixture

Data Type	Dataset	<b>#Samples</b>	#Hours	Sampling Ratio		
Text-only SFT	Nemotron SFT	94.0k	-	0.1500		
ASR, AST	Canary	32.8M	85k	0.7556		
SQA generated from ASR data	Canary (subset) w/ gemma-2-27b-it	4.1M	20k	0.0378		
Mixed-modal SFT	Alpaca	55.3k	85	0.0189		
generated by TTS	Magpie	254.5k	461	0.0378		



SQA data generation

- LLM: Gemma (2.5B) with LoRA (36M)
- Modality adapter: Conformer (52M)

### Trained with NeMo

> 20 hours on 64 NVIDIA A100 GPUs (80GB)

User: This is so cool! I'm learning about the future of AI. Can you tell me more? System: Sure! Here is my response [...]

### Select a random sentence in user turn

User: This is so cool! I'm learning about the future of AI. Can you tell me more? System: Sure! Here is my response [...]

#### Replace it with the synthesized audio

User: This is so cool! [Audio] Can you tell me more? System: Sure! Here is my response [...]

### Mixed-modal SFT data generation

Experiments																		
ASR WER ↓					En-X	AST BL	EU ↑	X-En AST BLEU ↑			SQA ↑			Speech Only ↑	ech Only ↑ Text Only ↑			
Model	En	De	Es	Fr	De	Es	Fr	De	Es	Fr	SPGI	SQuAD2	AIR-Bench	Spoken IFEval	GSM8K	IFEval	BBH	MMLU

Speech and text baselines

Whisper-v3 1.5B	9.92	6.17	4.94	11.18	-	_	-	33.4	22.7	33.7	-	-	_	_	_	-	-	-
SALMONN 7B	20.84	40.83	37.47	36.78	18.0	17.1	27.8	5.1	7.1	3.3	0.778	0.597	-	0.147	_	-	-	-
SALMONN 13B	17.07	44.08	28.47	38.52	19.0	18.5	29.1	6.5	3.6	3.8	0.778	0.604	6.16	0.113	_	_	-	-
Qwen2-Audio 7B	8.78	7.67	5.65	9.49	24.8	18.9	27.7	30.7	22.2	29.6	0.810	0.656	7.24	0.140	_	_	-	-
Gemma 2.5B	-	_	_	-	_	-	-	-	-	-	-	-	-	_	0.2479	0.2089	0.3324	0.3554
Our method and ablation studies																		
VTBlender 3B	7.90	5.53	4.52	7.09	29.6	22.5	38.6	36.3	25.6	33.8	0.828	0.684	6.31	0.191	0.2358	0.2237	0.3003	0.3484
Ablation 1	7.83	5.50	4.36	7.11	30.6	22.6	38.9	36.3	24.9	33.7	0.828	0.687	6.26	0.181	0.0243	0.1294	0.0023	0.2457
Ablation 2	9.96	8.77	7.03	9.49	21.5	18.3	29.3	31.3	21.9	29.3	0.028	0.121	3.22	0.150	0.2479	0.2089	0.3324	0.3554
Ablation 3	8.92	6.67	5.39	7.97	26.3	20.5	34.4	34.5	23.2	31.4	0.666	0.529	5.29	0.135	0.0675	0.1867	0.2622	0.2586

Results on speech-related tasks

✓ VTBlender 3B outperforms SALMONN 7B/13B and Qwen2-Audio 7B on most benchmarks

### Results on text-only tasks

 $(\mathbf{R})$ 

 $(\mathbf{8})$ 

 $(\mathsf{R})$ 

✓ VTBlender maintains similar text results as the LLM backbone

Model	Training Stages	SFT Data	LLM Training	Results of ablation study
VTBlender 3B	Single	Speech & text	LoRA	
Ablation 1	Single	Speech only	LoRA	LoRA alone cannot prevent catastrophic forgetting. Mixing in text-only SFT is very helpful
Ablation 2	Single	Speech only	Frozen	Freezing LM avoids catastrophic forgetting but significantly degrades speech performance
Ablation 3	Two	Speech only	Frozen → LoRA	Two-stage training helps a little, but is still much worse than our proposed method

Give me three suggestions. [audio]

(I don't feel well since last night, I had headache and I took a pill of Ibuprofen. I went to bed early.)

### Multi-turn, mixed-modal chat

Other demos can be found in the paper

- Understanding multi-speaker input
- ✓ ASR with unseen prompts
- ✓ AST in unseen directions
- ✓ Controlling output format, e.g., json
- ✓ Contextual biasing ASR
- ✓ Math and coding with mixed-modal input

I understand you're not feeling well. It's important to rest and seek medical attention. Here are a few things you can do:

- 1. Call your doctor or go to the emergency room immediately. They can assess your symptoms
- and recommend the best course of action.
- 2. Take a pain reliever. Over-the-counter pain relievers like ibuprofen or acetaminophen can help with headaches.
- 3. Stay hydrated. Drink plenty of fluids to prevent dehydration.
- Please remember to prioritize your health and seek professional help.

#### Give one more suggestion.

#### Here's another suggestion:

4. Rest and get enough sleep. Your body needs time to recover from illness.

Remember, it's always better to be safe than sorry. If you're unsure about your symptoms or need further assistance, don't hesitate to reach out to a healthcare professional.

#### [audio]

(Where can I find doctors near Santa Clara?)

I can't provide medical advice, and finding a doctor is crucial.
Here are some resources to help you find doctors near Santa Clara:
Your primary care doctor: They can refer you to a specialist if needed.
The Santa Clara County Health Department website: https://www.santaclaracounty.org/
The American Medical Association (AMA): https://www.ama.org/
The National Institutes of Health (NIH): https://www.nih.gov/
You can also use online search engines like Google to find doctors in your area.

- Tang, Changli, et al. "SALMONN: Towards generic hearing abilities for large language models." in Proc. ICLR, 2024.
- Chen, Zhehuai, et al. "SALM: Speechaugmented language model with in-context learning for speech recognition and translation." in Proc. ICASSP, 2024.
- Chu, Yunfei, et al. "Qwen2-audio technical report." arXiv preprint arXiv:2407.10759 (2024).
- Puvvada, Krishna, et al. "Less is More: Accurate Speech Recognition & Translation without Web-Scale Data." in Proc. INTERSPEECH, 2024.
- Gemma Team, et al. "Gemma: Open models based on gemini research and technology." arXiv preprint arXiv:2403.08295 (2024).
- Adler, Bo, et al. "Nemotron-4 340b technical report." arXiv preprint arXiv:2406.11704 (2024).