One Voice, Many Tongues: Cross-Lingual Voice Cloning for Scientific Speech

Amanuel Gizachew Abebe
Shaggar Institute of Technology
amanuel.g.abebe1@gmail.com

Yasmin Moslem
Trinity College Dublin
yasmin.moslem@adaptcentre.ie

Abstract

Preserving a speaker's voice identity while generating speech in a different language remains a fundamental challenge in spoken language technology, particularly in specialized domains such as scientific communication. In this paper, we address this challenge through our system submission to the International Conference on Spoken Language Translation (IWSLT 2026), the Cross-Lingual Voice Cloning shared task. First, we evaluate several state-of-the-art voice cloning models for cross-lingual speech generation of scientific texts in Arabic, Chinese, and French. Then, we build voice cloning systems based on the OmniVoice foundation model. We employ data augmentation via multi-model ensemble distillation from the ACL 60/60 corpus. We investigate the effect of using this synthetic data for fine-tuning, demonstrating improvements in intelligibility (WER & CER) and speaker similarity (SIM), with gains varying across languages.

1 Introduction

The rapid advancement of speech synthesis has enabled zero-shot voice cloning, where a model generates speech in a target speaker's voice using only a few seconds of reference audio (Wang et al., 2023; Tan et al., 2021). This technology is particularly transformative for the scientific community, allowing for cross-lingual dissemination of research findings and enhancing accessibility for academic presentations. However, scientific speech presents unique challenges, including the prevalence of technical terminology, code-switching, specific prosodic patterns, and the need for high intelligibility across diverse languages.

The IWSLT 2026 Cross-Lingual Voice Cloning shared task (Adelani et al., 2026) challenges participants to clone voices for three diverse languages, namely Arabic, Chinese, and French. A primary bottleneck in adapting large-scale TTS models to these languages is the scarcity of high-quality paired training data that captures the nuances of academic discourse. While foundation models like OmniVoice (Zhu et al., 2026) provide broad multilingual coverage, they often require domain-specific fine-tuning to achieve optimal performance on scientific text.

In this work, we leverage ensemble distillation to address the data scarcity challenge and fine-tune our voice-cloning models. We utilize three zero-shot voice cloning models as "teachers" to synthesize data from the ACL 60/60 academic corpus (Salesky et al., 2023). By selecting the best output from this ensemble, we curate a high-fidelity synthetic dataset for fine-tuning. We then employ Parameter-Efficient Fine-Tuning (PEFT) via LoRA (Hu et al., 2022) to adapt the base OmniVoice model for each target language.

Our system demonstrates that even with a modest computational budget, targeted LoRA fine-tuning on ensemble-distilled data yields a highly competitive model that balances intelligibility and speaker similarity. Consequently, we submit this per-language fine-tuned OmniVoice model as our primary submission to the IWSLT 2026 shared task. To enable reproducibility, our code for data preparation, training, and evaluation is publicly available.[1]

2 Related Work

Zero-Shot and Multilingual Voice Cloning. The transition to zero-shot voice cloning has been accelerated by large-scale foundation models trained on diverse audio-text corpora. Early autoregressive models like VALL-E (Chen et al., 2025) demonstrated that conditioning on short reference utterances enables accurate timbre transfer. Recent architectures build upon this by leveraging discrete audio tokenization and treating speech

[1] https://github.com/Aman-byte1/multilingual-voice-cloning-training

Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026), pages 227–231

synthesis as a language modeling task. For instance, Qwen3 (Hu et al., 2026) and OmniVoice (Zhu et al., 2026) utilize dense transformer backbones (Vaswani et al., 2017) to achieve robust cross-lingual cloning. Similarly, models like XTTS-V2 (Casanova et al., 2024) and VoxCPM (Zhou et al., 2025) have introduced highly optimized pipelines for context-aware speech generation, while architectures such as Chatterbox (Resemble AI, 2025) provide robust synthesis for specific language families. While these foundation models generalize well, adapting them to specialized domains such as scientific and academic speech remains a challenge due to the complex phonetic realizations of technical terminology.

Data Distillation and Ensemble Selection. Data scarcity in specialized domains often necessitates synthetic data generation. Knowledge distillation (Hinton et al., 2015; Kim and Rush, 2016; Gandhi et al., 2023; Moslem, 2025) from powerful teacher models to smaller or specialized models is a common strategy. Our Best-of-N ensemble approach ensures that only the highest quality synthetic samples are used to fine-tune the final system, effectively bypassing the limitations of scarce parallel data. Specifically, we generate multiple candidate audios for each utterance and select the best one based on a combined score: character error rate for intelligibility (via Whisper (Radford et al., 2023)) and cosine distance for speaker similarity (via ECAPA-TDNN (Desplanques et al., 2020; Das et al., 2021)).

Parameter-Efficient Fine-Tuning in Speech. While foundation models generalize well, adapting them to specific domains without catastrophic forgetting requires efficient techniques. Low-Rank Adaptation (LoRA) (Hu et al., 2022) has increasingly been used for fine-tuning large foundation models. By updating only a small subset of parameters within the self-attention and feed-forward layers, LoRA allows models to capture specific phonetic distributions and language nuances with minimal computational overhead. Innovations such as Rank Stabilization (Kalajdzievski, 2023) further enhance the stability of this fine-tuning process.

3 Data

In this section, we describe our training dataset that we augmented with knowledge distillation and used for fine-tuning our models.

Lang	Model	Wins	Percentage
AR	Chatterbox	0	0.0%
	OmniVoice	649	73.4%
	VoxCPM	235	26.6%
FR	Chatterbox	210	23.76%
	OmniVoice	674	76.24%
	VoxCPM	0	0.0%
ZH	Chatterbox	0	0.0%
	OmniVoice	601	68%
	VoxCPM	283	32%

Table 1: Distribution of selected samples in the Best-of-N distilled dataset broken down by language. Each language subset consists of 884 total utterances.

3.1 Source Dataset

We utilize the the ACL 60/60 dataset (Salesky et al., 2023) for multilingual translation of ACL 2022 technical presentations into several target languages. We use the development split, which consists of 884 utterances per language for Arabic (AR), French (FR), and Chinese (ZH), totaling 2,652 samples. Each sample includes the target text and a reference audio clip of the original speaker, providing a ground truth for speaker similarity.

3.2 Best-of-N Ensemble Distillation

To create a high-quality fine-tuning dataset, we implement an ensemble distillation pipeline using three teacher models:

- **OmniVoice**[2] (Zhu et al., 2026): A 0.6B parameter model based on Qwen3, supporting over 600 languages.
- **VoxCPM**[3] (Zhou et al., 2025): A cross-lingual model optimized for zero-shot timbre transfer.
- **Chatterbox**[4] (Resemble AI, 2025): An open-source architecture known for robust synthesis of European and Semitic languages.

For every utterance in the source dataset, each model generates a synthesis candidate. We then evaluate these candidates using a combined quality score:

$$S_{\text{comb}} = 0.5 \times (1 - \text{CER}) + 0.5 \times \text{SIM} \quad (1)$$

where CER is the character error rate produced by Whisper large-v3 (Radford et al., 2023) and

[2] https://huggingface.co/k2-fsa/OmniVoice
[3] https://huggingface.co/openbmb/VoxCPM
[4] https://huggingface.co/ResembleAI/chatterbox

SIM is the cosine similarity of speaker embeddings from an ECAPA-TDNN model (Desplanques et al., 2020). The candidate with the highest S_{comb} is selected for the fine-tuning set. As shown in Table 1, this strategy captures the strengths of multiple architectures, with non-primary models contributing over 27% of the final curated data.

4 Experiments

In this section, we elaborate on your experiments, including data processing, training configuration, and inference pipeline.

4.1 Data Preprocessing

Prior to training, we filter our Best-of-N distilled dataset using a minimum quality threshold to ensure high-fidelity audio retention (cf. Section 3.2). Then, we partition the dataset into standard training and development splits. Raw audio waveforms are tokenized into discrete acoustic tokens using a HIGGS-based tokenizer. This enables the base OmniVoice model to process speech synthesis as a discrete language generation task.

4.2 Per-Language Fine-Tuning Strategy

We fine-tune the OmniVoice model, which utilizes a Qwen3-0.6B backbone. Our approach relies on the observation that a single unified multilingual adapter can sometimes dilute language-specific phonological nuances. To address this, we train dedicated Low-Rank Adaptation (LoRA) modules exclusively for Arabic, Chinese, and French. By adapting the transformer's self-attention blocks, feed-forward networks, and audio projection layers, the model captures language-specific acoustic distributions efficiently without the risk of catastrophic forgetting.

4.3 Training Configuration

To maintain training stability with our small dataset, we utilize Rank-Stabilized LoRA (RSLoRA) and optimize the model using an autoregressive cross-entropy loss over the generated audio tokens. Fine-tuning is executed concurrently across NVIDIA A40 GPUs, dedicating a single GPU per language. The models are trained efficiently for exactly 400 steps using mixed precision and a cosine learning rate schedule, ensuring rapid convergence. Full replication details and precise hyperparameter configurations are available in our open-source code repository.

Lang	Model	WER ↓	CER ↓	SIM ↑
AR	Chatterbox	0.250	0.086	0.680
	XTTS-V2	0.253	0.099	0.501
	VoxCPM2	**0.209**	**0.072**	0.607
	OmniVoice	0.238	0.076	**0.703**
FR	Chatterbox	0.111	0.045	0.619
	Qwen3-TTS	**0.050**	**0.011**	0.533
	XTTS-V2	0.082	0.031	0.445
	VoxCPM2	0.128	0.069	0.575
	OmniVoice	0.079	0.020	**0.753**
ZH	Chatterbox	–	0.203	0.653
	Qwen3-TTS	–	**0.090**	0.522
	XTTS-V2	–	0.176	0.511
	VoxCPM2	–	0.149	0.569
	OmniVoice	–	0.219	**0.702**

Table 2: Comparative evaluation against baselines on the `blindset-4` subset. OmniVoice achieves state-of-the-art speaker similarity across all languages, with competitive intelligibility for Arabic and French.

4.4 Inference Pipeline

The inference pipeline consists of three stages:

1. Reference extraction: We isolate a 20-second speech segment from the reference audio using energy-based Voice Activity Detection (VAD).

2. Text chunking: Long inputs are split into segments of up to 200 characters at sentence boundaries.

3. Synthesis: Each chunk is processed with a temperature of 0.8 and top-p of 0.9, and the resulting audio segments are concatenated.

5 Evaluation and Results

We evaluate our approach on the official blind test set, which consists of 49, 99, and 112 text segments collected from diverse scientific publications in Arabic (AR), French (FR), and Chinese (ZH), respectively, and 12 reference audio voices in English extracted from ACL 2023 presentations. To establish strong baselines, we compare against several state-of-the-art voice cloning models: Chatterbox (Resemble AI, 2025), Qwen3-TTS (Hu et al., 2026), XTTS-V2 (Casanova et al., 2024), and VoxCPM2 (Zhou et al., 2025) on a representative 4-speaker subset (`blindset-4`). For our primary contribution, we evaluate the base OmniVoice model and our LoRA-finetuned OmniVoice model on the complete blind test set (`blindset-full`). Metrics include Word Error Rate (WER) and Character

Lang	Model	WER ↓	CER ↓	SIM ↑
AR	OmniVoice (Baseline)	0.244	0.077	**0.734**
	OmniVoice (Finetuned)	**0.228**	**0.060**	0.726
FR	OmniVoice (Baseline)	**0.079**	**0.025**	0.753
	OmniVoice (Finetuned)	0.082	0.030	**0.760**
ZH	OmniVoice (Baseline)	–	**0.200**	0.719
	OmniVoice (Finetuned)	–	0.205	**0.732**

Table 3: Impact of LoRA fine-tuning evaluated on the complete `blindset-full` dataset.

Error Rate (CER) (Morris et al., 2004) for intelligibility (transcription accuracy), and speaker similarity (SIM) for cloning fidelity using ECAPA-TDNN embeddings. It is worth noting that Qwen3-TTS does not support Arabic synthesis. Also, the character-based CER metric is more representative of Chinese text quality than WER, as Chinese does not use word delimiters.

5.1 Results and Discussion

The quantitative results are presented in Table 2 (baseline comparison on the 4-speaker subset) and Table 3 (ablation of LoRA fine-tuning on the full blind set).

Quantitative Analysis. The results demonstrate the varying strengths of the evaluated architectures. In French, Qwen3-TTS achieves the lowest error rates (WER of 0.050), but our OmniVoice models deliver significantly higher speaker similarity (0.753) while maintaining strong intelligibility. For Chinese, the baseline models show lower error rates on the subset, but the fine-tuned OmniVoice model achieves the highest speaker similarity (0.719) across the entire test set. In Arabic, the fine-tuned OmniVoice model achieves the lowest CER (0.071) and strong speaker similarity (0.723), outperforming XTTS-V2 and VoxCPM2 in overall cloning fidelity.

Impact of LoRA Fine-Tuning. Across all three languages, our per-language LoRA fine-tuning strategy demonstrates improvements in either intelligibility or speaker similarity over the baseline OmniVoice model, with gains varying by language. For Arabic, fine-tuning yields clear improvements in WER and CER, while for French and Chinese, it primarily improves speaker similarity (SIM). This suggests a trade-off between the two objectives, though in all cases the fine-tuned model successfully adapts to the acoustic properties of the scientific domain without catastrophic forgetting of the source voice characteristics.

6 Conclusion

We presented our system for the IWSLT 2026 Voice Cloning task. By combining multi-model ensemble distillation with per-language LoRA fine-tuning, we demonstrate a robust and efficient path for adapting large-scale TTS models to the scientific domain. Our results show improvements in intelligibility (WER & CER) and speaker similarity (SIM), with gains varying across languages. Future work will explore larger distilled datasets and human evaluations to further validate the perceptual quality of the synthesized scientific speech.

Limitations

Our study is limited by the scale of the distilled training dataset, and the use of automated metrics (Whisper/ECAPA) as proxies for human perception. While these metrics show clear trends, they may not capture all nuanced artifacts of synthesized speech. Furthermore, our per-language approach increases the number of model adapters compared to a unified multilingual approach.

Ethical Considerations

Voice cloning technologies possess dual-use capabilities. While our primary aim is to democratize scientific knowledge and enhance cross-lingual accessibility for academic presentations, the ability to synthesize highly accurate voice clones carries inherent risks of misuse, such as deepfakes, identity theft, or spreading misinformation. To mitigate these risks, we strongly advocate for the integration of synthetic speech watermarking and robust voice authentication protocols when deploying such systems in public-facing applications. Ultimately, this work is intended for beneficial applications, such as advancing scientific communication and supporting assistive technology.

References

David Ifeoluwa Adelani, Victor Agostinelli, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Sebastien Bratières, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, Marcello Federico, Marco Gaido, Mahendra Gupta, HyoJung Han, Ali Hatami, David Javorský, Yejin Jeon, Marek Kasztelnik, Antoine Laurent, Danni Liu, Nam Luu, Min Ma, Dominik Macháček, Marie Maltais, Evgeny Matusov, Chandresh Kumar Maurya, John P. McCrae, Chutong Meng, Mohammadamini Mohammad, and 23 others. 2026. Speech translation and metrics in 2026: Findings of the IWSLT campaign. In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, California, US. Association for Computational Linguistics.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and Julian Weber. 2024. XTTS: a massively multilingual zero-shot text-to-speech model. In *Proc. Interspeech 2024*, pages 4978–4982.

Sanyuan Chen, Chengyi Wang, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2025. Neural codec language models are zero-shot text to speech synthesizers. *IEEE Transactions on Audio, Speech and Language Processing*, 33:705–718.

Rohan Kumar Das, Ruijie Tao, and Haizhou Li. 2021. Hlt-nus submission for 2020 nist conversational telephone speech sre. *arXiv preprint arXiv:2111.06671*.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In *Interspeech 2020*.

Sanchit Gandhi, Patrick von Platen, and Alexander M Rush. 2023. Distil-Whisper: Robust knowledge distillation via large-scale pseudo labelling. *arXiv [cs.CL]*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, Xinyu Zhang, Pei Zhang, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. 2026. Qwen3-TTS technical report. *arXiv preprint arXiv:2601.15621*.

Damjan Kalajdzievski. 2023. A rank stabilization scaling factor for fine-tuning with LoRA. *arXiv preprint arXiv:2312.03732*.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

A.C. Morris, V. Maier, and P. Green. 2004. From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. In *Proc. Interspeech 2004*, pages 2765–2768.

Yasmin Moslem. 2025. Efficient speech translation through model compression and knowledge distillation. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 379–388, Vienna, Austria (in-person and online). Association for Computational Linguistics.

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*.

Resemble AI. 2025. Chatterbox-TTS. `https://github.com/resemble-ai/chatterbox`. GitHub repository.

Elizabeth Salesky, Kareem Darwish, Mohamed Al-Badrashiny, Mona Diab, and Jan Niehues. 2023. Evaluating multilingual speech translation under realistic conditions with resegmentation and terminology. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 62–78, Toronto, Canada. Association for Computational Linguistics.

Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. 2021. A survey on neural speech synthesis. In *arXiv preprint arXiv:2106.15561*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, and 1 others. 2023. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.

S. Zhou, Y. Zeng, and 1 others. 2025. VoxCPM: Tokenizer-free TTS for context-aware speech generation and true-to-life voice cloning. *arXiv preprint arXiv:2509.24650*.

Han Zhu, Lingxuan Ye, Wei Kang, Zengwei Yao, Liyong Guo, Fangjun Kuang, Zhifeng Han, Weiji Zhuang, Long Lin, and Daniel Povey. 2026. Omnivoice: Towards omnilingual zero-shot text-to-speech with diffusion language models. *arXiv preprint arXiv:2604.00688*.