KIT's Submission to Cross-Lingual Voice Cloning in IWSLT 2026

Seymanur Akti[1,3], **Alexander Waibel**[1,2],
[1] Karlsruhe Institute of Technology (KIT)
[2] Carnegie Mellon University (CMU), [3] KIT Campus Transfer (KCT)
Correspondence: seymanur.akti@kit.edu

Abstract

Cross-lingual voice cloning aims to generate speech in a target language while preserving speaker identity from a source-language reference. This task is central to speech translation and is the focus of the IWSLT 2026 Cross-Lingual Voice Cloning track. A key challenge is maintaining intelligibility and naturalness in the presence of accent variation and domain-specific vocabulary. We build on a multilingual text-to-speech model, FishAudio-S2-Pro, and introduce language tag prompting to improve language control and reduce accent leakage. We further apply reinforcement learning (RL) fine-tuning for task adaptation and observe improvements in intelligibility. Finally, we propose a reference-conditioned lexical matching method that improves pronunciation of domain-specific terms when lexical overlap is present.

Results show that language prompting provides the largest gains, while lexical matching yields consistent improvements on matched subsets.

1 Introduction

Recent advances in multilingual text-to-speech (TTS) systems have enabled high-quality speech synthesis across multiple languages within a single model, while also supporting in-context voice cloning from reference audio. This capability is particularly important for speech translation pipelines, where cross-lingual voice cloning aims to generate speech in a target language while preserving speaker identity and speaking style from a source-language reference.

In IWSLT 2026 (Adelani et al., 2026), Cross-Lingual Voice Cloning track focuses on this setting, requiring synthesis in French, Arabic, and Chinese from English reference speech. The dataset includes diverse speaker accents and domain-specific terminology, which increases the difficulty of maintaining pronunciation consistency under cross-lingual conditions.

Cross-lingual voice cloning can be formulated using cascaded pipelines, where speech is first synthesized using a TTS system and then transformed using a separate voice conversion model. While such cascaded approaches offer modularity and flexibility, they introduce additional sources of error, including speaker leakage (Akti et al., 2025) and misalignment between linguistic content and prosody.

An alternative direction is end-to-end conditioning of speech synthesis on speaker characteristics. One line of work incorporates speaker information via explicit speaker encoders, which inject speaker representations into the synthesis model (Casanova et al., 2022; Lee et al., 2025; Li et al., 2023). This approach remains widely used due to its simplicity and stability; however, its voice cloning capability is constrained by the quality and expressiveness of the learned speaker embeddings.

More recent work has explored in-context learning approaches for zero-shot voice cloning, where models are conditioned directly on reference audio to generate speech across unseen speakers and languages. Systems such as VALL-E (Wang et al., 2023), F5-TTS (Chen et al., 2025), Qwen3-TTS (Hu et al., 2026), and CosyVoice3 (Du et al., 2025) follow this paradigm and demonstrate strong performance in preserving speaker identity in zero-shot settings. However, cross-lingual generation remains challenging even for large-scale models. A common issue is accent leakage, where phonetic characteristics of the source language persist in the generated speech. In addition, domain-specific words and named entities are often mispronounced, especially when they are out-of-distribution for the target language.

In this work, we present Karlsruhe Institute of Technology's (KIT) submission to the IWSLT 2026 Cross-Lingual Voice Cloning track. We build on a strong pretrained multilingual TTS model,

Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026), pages 78–83

FishAudio-S2-Pro[1], and focus on improving cross-lingual pronunciation through input-level conditioning and lightweight adaptation strategies.

2 Data

We evaluate our system on the ACL 60/60 dataset (Salesky et al., 2023)[2], which consists of long-form speech recordings derived from ACL conference talks. The dataset is designed for cross-lingual voice cloning and follows the IWSLT 2026 shared task setup, where English reference speech is paired with target text in French, Arabic, and Chinese.

The dataset presents several challenges. First, reference speech exhibits substantial variability in speaker accents, which introduces inconsistencies in speaker representation and complicates cross-lingual transfer. Second, the target text contains domain-specific terminology and named entities, many of which are either out-of-distribution for the target languages or appear in their original English form. This makes accurate pronunciation particularly challenging under cross-lingual conditions.

We use the ACL 60/60 development set for system development and reinforcement learning fine-tuning, and reserve the evaluation set for validation and analysis. All recordings are provided with gold utterance-level segmentation, and no additional segmentation or alignment is performed during fine-tuning. For each target text sample, the corresponding English reference speech-text pairs are used for conditioning.

In the blind test scenario, reference audio is provided in long-form format. To enable effective conditioning, we use an ASR model to segment the audio into smaller chunks around 2 to 10 seconds paired with their corresponding text transcriptions. We then perform inference-time retrieval over these segments when required by our lexical matching strategy. This retrieval is applied only at inference time and does not introduce additional supervision.

3 Model Adaptation

We build our system on top of FishAudio-S2-Pro, a multilingual text-to-speech model that supports in-context voice cloning from reference audio (Liao et al., 2026). The model is capable of synthesizing speech in multiple languages, including Arabic, French, and Chinese, making it suitable for the IWSLT cross-lingual voice cloning task.

FishAudio-S2-Pro allows flexible text prompting, enabling the use of free-form control tokens in the input. In our work, we leverage this capability to introduce explicit language tags for improved language control during generation. To simulate a realistic scenario where target speech is unavailable, we perform reinforcement learning fine-tuning instead of supervised training, aiming to maintain the original model quality while improving perceptual performance in cross-lingual settings.

3.1 Prompts with Language Tags

FishAudio-S2-Pro does not explicitly use language tags during training, and language identification is implicitly inferred from the input text. In cross-lingual voice cloning with autoregressive generation over mixed-language sequences, this can lead to accent leakage when multiple languages are present within a sequence.

To address this, we leverage the model's support for free-form prompting and introduce explicit language tags for each text input, including the reference text. We experiment with both English-language tags (e.g., [english], [arabic], [french], [chinese]) and native-script tags (e.g., [english], [العربية], [français], [普通话]). We find that native-script tags provide stronger conditioning signals and yield better cross-lingual pronunciation quality in our setup.

Overall, language tags act as a simple but effective control mechanism, guiding the model toward the desired phonetic realization and reducing cross-lingual interference.

An example of the constructed prompt for cross-lingual generation for submission is shown below:

Reference:
`[english] The little cat is sleeping under the table.`

Chinese (Mandarin) Target:
`[普通话]` 小猫正在桌子下面睡觉。

French Target:
`[français] Le petit chat dort sous la table.`

Arabic Target:
`[العربية]` القطة الصغيرة نائمة تحت الطاولة.

[1] fishaudio/fish-speech
[2] ymoslem/acl-6060

3.2 RL Fine-tuning

To adapt the model to the cross-lingual voice cloning task, we perform reinforcement learning (RL) fine-tuning using Group Relative Policy Optimization (GRPO) (Shao et al., 2024), which has already shown effectiveness in improving the base model. The objective is to adapt the model to the newly introduced language tags and cross-lingual inference without requiring supervised parallel data.

We define the reward function based on two components: (1) Character Error Rate (CER), computed using a multilingual ASR model (Pratap et al., 2024)[3], and (2) speaker similarity (SSIM), computed using a speaker verification model (Chen et al., 2022).[4]

3.2.1 Hyperparameters and Implementation Details

During RL fine-tuning, we adopt a more aggressive update strategy by optimizing attention, MLP, and output layers, compared to the original recipe which only updates MLP layers. We use LoRA with rank $r = 64$ and scaling factor $\alpha = 16$. The group size for GRPO was chosen as 8 and we use AdamW optimizer with learning rate $10^{-}5$

In addition, we experiment with different values of the KL divergence penalty coefficient, which controls deviation from the base model. Proper tuning of this coefficient improves training stability and leads to better adaptation to the target task. We used $\beta = 0.1$ as the penalty coefficient.

For reward calculation, the CER score is inverted, and both CER and SSIM are scaled to [0,1]. The final reward is computed as their average:

$$Reward = \frac{(1 - CER) + SSIM}{2} \quad (1)$$

3.3 Reference Speech Retrieval from Long Audio

To better handle domain-specific vocabulary in the blind test set, we introduce a reference speech retrieval strategy from long-form audio. Given long reference recordings, we first segment the audio into smaller chunks and transcribe them using the VibeVoice long-form ASR model (Peng et al., 2026).[5]

We then perform lexical matching between the target text and transcribed audio segments, selecting reference segments that contain overlapping words. This allows the model to observe correct pronunciations of domain-specific terms and rare words directly from the retrieved reference speech.

By conditioning on these retrieved acoustic segments, this approach improves pronunciation accuracy while preserving speaker-specific characteristics.

4 Evaluation Results

We evaluate our model on the ACL 60/60 evaluation subset, covering Arabic, Chinese, and French as target languages. Each sample is synthesized in all three languages given English reference speech. We compare three settings: (i) a baseline model (FishAudio-S2-Pro) without language tags, (ii) the same model with explicit language tags, and (iii) an RL fine-tuned model using language-tagged cross-lingual prompts.

We use the following evaluation metrics:

- **CER**: Character Error Rate computed using Whisper-Large-v3 (Radford et al., 2023)[6] for intelligibility assessment. We use CER instead of WER to ensure comparability across languages.
- **SSIM**: Speaker similarity calculated by cosine similarity between speaker embeddings extracted from source and generated speech using a pre-trained speaker verification model (Ravanelli et al., 2021).[7]
- **UTMOS**: Predicted mean opinion score using a pre-trained MOS estimation model (Saeki et al., 2022).

We use different ASR and speaker verification models from those used in the reward computation during RL fine-tuning to avoid evaluation bias and ensure a fair assessment of generalization performance.

The results in Table 1 show that introducing language tags on the base model improves CER for Arabic and French, while a slight degradation is observed for Chinese. This suggests that language conditioning might help reduce cross-lingual pronunciation drift even when it is used without prior

[3] facebook/mms-1b-all
[4] microsoft/wavlm-base-plus-sv
[5] microsoft/VibeVoice-ASR
[6] openai/whisper-large-v3
[7] speechbrain/spkrec-ecapa-voxceleb

Model	Lang. Tags	CER (%) ↓			SSIM (%) ↑			UTMOS ↑		
		ar	fr	zh	ar	fr	zh	ar	fr	zh
Baseline	✗	6.57	3.10	11.37	64.05	60.75	62.39	**2.94**	2.86	**2.90**
Baseline	✓	6.39	2.90	12.05	63.77	60.21	62.49	**2.94**	2.85	2.88
RL Finetuned †	✓	**6.38**	**2.78**	**10.99**	**64.15**	**60.83**	**62.52**	2.93	**2.88**	2.89

Table 1: Evaluation across languages. CER is reported in % (lower is better), while speaker similarity and UTMOS are higher-is-better metrics. † indicates submitted system.

tuning, although its effect is inconsistent across languages.

For speaker similarity, adding language tags leads to only minor variations, indicating that linguistic conditioning primarily affects phonetic realization while preserving speaker identity.

After RL fine-tuning with the language tags, we observe small but consistent improvements or stability across all languages. In particular, CER improves further compared to the language-tag-only setting, indicating that RL fine-tuning further refines pronunciation consistency under cross-lingual conditioning. Speaker similarity and UTMOS remain stable, suggesting that RL fine-tuning does not degrade speaker identity or perceptual quality.

4.1 Source Language Bias Analysis

We further find that language tags primarily reduce English pronunciation bias, leading to more natural target-language realization. Without language tags, generated speech often retains English-influenced phonetic characteristics, particularly in long-form reference conditions, which can amplify this bias during autoregressive generation. This effect is reduced when language tags are introduced, resulting in more consistent target-language pronunciation and prosody.

To evaluate this behaviour, we use a pre-trained language identification model[8] to measure target language confidence in generated speech. We hypothesize that leakage from the reference speech introduces pronunciation drift, which can negatively affect language identification performance.

Table 2 shows that language tags consistently improve target language identification probabilities across all languages, indicating reduced cross-lingual pronunciation leakage and improved language consistency.

At the same time, speaker-specific characteristics remain largely preserved, including speaking rate, accent-related phonetic traits, and speaking style. For example, speakers with strong regional accents tend to retain similar accent characteristics across target languages, indicating that language conditioning primarily affects linguistic realization rather than speaker identity.

[8] speechbrain/lang-id-voxlingua107-ecapa

Model	ar	fr	zh	avg
Normal Prompts	89.87	88.68	90.99	89.85
w/ Language Tags	**93.42**	**90.23**	**92.13**	**91.64**

Table 2: Target language identification probabilities (%) with and without explicit language tags. Language conditioning improves target language confidence across all languages.

4.2 Impact of Lexical Matching on Special Word Pronunciation

We further analyze pronunciation behavior on named entities and domain-specific terms under matched and non-matched reference conditioning. The matched setting is constructed using reference audio corresponding to the exact translation of the target text, such that domain-specific terms are already present in the reference prompt and provide explicit phonetic grounding. In contrast, the non-matched setting uses randomly selected reference audio from the same speaker, which does not necessarily contain the target lexical content.

As shown in Table 3, matched reference audio significantly improves pronunciation consistency for out-of-language words and acronyms. In this setting, the model is able to better preserve correct pronunciations when the reference speech provides aligned lexical and phonetic cues. We also observe that speaker-specific pronunciation patterns are preserved for named entities, reflecting the way the speaker produces these terms in the reference audio.

Entity	Matching	Non-matching
VALSE	/vɑls/	/viː eɪ ɛl ɛs iː/
LXMert	/ɛl ɛks mɜːrt/	/ɛl ɛks ɛm ɜːrt/
ViLBERT	/vɪlbɜːrt/	/viː ɛl bɜːrt/
Word2Vec	/wɜːd tu vɛk/	/wɜːd tu vɪk/
RNSum	/ɑːr ɛn sʊm/	ɑːr ɛn ɛs ʊm/
SVAMP	/swɑːmp/	/ɛs viː eɪ ɛm piː/

Table 3: Entity-level pronunciation comparison under matching and non-matching reference conditioning.

In contrast, under non-matched conditions, the model exhibits weaker phonetic grounding and tends to rely on spelling-based or language-default pronunciation strategies. This often results in letter-by-letter or segmented pronunciation of technical terms and acronyms. For example, tokens such as "seq2seq" or "word2vec" are interpreted compositionally, where the digit "2" is pronounced as a number rather than as part of the unit. This highlights the importance of reference-conditioned lexical grounding for accurate pronunciation.

5 Conclusion

We presented KIT's system for the IWSLT 2026 Cross-Lingual Voice Cloning task, focusing on improving multilingual in-context TTS through lightweight input conditioning and reinforcement learning fine-tuning. Starting from a strong pre-trained multilingual TTS model, we investigated the effects of explicit language tags and RL-based adaptation using GRPO.

Our results show that language tags help reduce cross-lingual pronunciation drift and improve target-language consistency, while RL fine-tuning further stabilizes performance across Arabic, French, and Chinese without degrading speaker similarity or perceptual quality. In addition, we analyzed pronunciation behavior under different reference conditions and found that lexical grounding plays an important role in correctly rendering domain-specific terms and acronyms, while speaker identity remains largely preserved.

Overall, our study highlights that simple conditioning strategies combined with targeted optimization can effectively improve robustness in cross-lingual voice cloning, particularly in handling pronunciation consistency and reference-induced bias.

Acknowledgments

This research is supported by the European Union's Horizon Europe programme grant agreement No. 101213369 (DVPS) and KIT Campus Transfer GmbH (KCT) in accordance with the collaboration with Carnegie-AI.

References

David Ifeoluwa Adelani, Antonios Anastasopoulos, Victor Agostinelli, 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, Danni Liu, and 26 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.

Seymanur Akti, Tuan Nam Nguyen, and Alexander Waibel. 2025. Towards better disentanglement in non-autoregressive zero-shot expressive voice conversion. *arXiv preprint arXiv:2506.04013*.

Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. 2022. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International conference on machine learning*, pages 2709–2720. PMLR.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, JianZhao JianZhao, Kai Yu, and Xie Chen. 2025. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6255–6271.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, and 1 others. 2025. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*.

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, and 1 others. 2026. Qwen3-tts technical report. *arXiv preprint arXiv:2601.15621*.

Sang-Hoon Lee, Ha-Yeong Choi, Seung-Bin Kim, and Seong-Whan Lee. 2025. Hierspeech++: Bridging the gap between semantic and acoustic representation of speech by hierarchical variational inference for zero-shot speech synthesis. *IEEE Transactions on Neural Networks and Learning Systems*.

Yinghao Aaron Li, Cong Han, Vinay Raghavan, Gavin Mischler, and Nima Mesgarani. 2023. Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. *Advances in neural information processing systems*, 36:19594–19621.

Shijia Liao, Yuxuan Wang, Songting Liu, Yifan Cheng, Ruoyi Zhang, Tianyu Li, Shidong Li, Yisheng Zheng, Xingwei Liu, Qingzheng Wang, and 1 others. 2026. Fish audio s2 technical report. *arXiv preprint arXiv:2603.08823*.

Zhiliang Peng, Jianwei Yu, Yaoyao Chang, Zilong Wang, Li Dong, Yingbo Hao, Yujie Tu, Chenyu Yang, Wenhui Wang, Songchen Xu, and 1 others. 2026. Vibevoice-asr technical report. *arXiv preprint arXiv:2601.18184*.

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.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, and 2 others. 2021. SpeechBrain: A general-purpose speech toolkit. *Preprint*, arXiv:2106.04624. ArXiv:2106.04624.

Takaaki Saeki and 1 others. 2022. Utmos: Utokyo-sarulab system for voicemos challenge 2022. In *Interspeech*.

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.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

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*.