HW-TSC's Submission to the IWSLT 2026 Cross-Lingual Voice Cloning Track

Yu He[1,2], **Daimeng Wei**[1], **Jiaxin Guo**[1], **Yuanchang Luo**[1], **Hengchao Shang**[1], **Zongyao Li**[1], **Zhiqiang Rao**[1], **Jinlong Yang**[1], **Zhanglin Wu**[1], **Boqi Huang**[1,2], **Xiaoqing Lan**[1,2]
[1]Huawei Translation Service Center, Beijing, China
[2]School of Software, Northwestern Polytechnical University, Xi'an, China
{heyu97, weidaimeng, guojiaxin1, luoyuanchang1, shanghengchao, lizongyao, raozhiqiang, yangjinlong7, wuzhanglin2, huangboqi1, lanxiaoqing}@huawei.com

Abstract

This paper presents HW-TSC's submission to the IWSLT 2026 Cross-Lingual Voice Cloning Track. The Cross-Lingual Voice Cloning Track includes three target languages: Arabic, Chinese, and French. We take part in two language tasks of this track, namely Chinese and French. We employ the Qwen3-TTS-12Hz-1.7B-Base multilingual model as the core voice cloning model. To tackle problems such as excessively long duration of the original reference audio and scattered features, we design a sliding-window audio segmentation preprocessing method, which continuously splits long audio into standardized short segments with overlapping redundancy. This method avoids feature attenuation caused by overly long audio and maximizes the preservation of complete timbre information through step overlap. To select the outputs with the highest timbre similarity from numerous synthetic results, this study conducts voiceprint recognition based on the Enhanced Context-Dependent Adversarial Time Delay Neural Network (ECAPA-TDNN), with cosine similarity as the core quantitative evaluation metric, and selects the result with the highest similarity as the optimal output.

1 Introduction

In recent years, text-to-speech (TTS) systems evolve from merely generating natural and high-fidelity speech to enabling controllable and expressive voice synthesis (Sun et al., 2026). As a crucial research direction in personalized speech synthesis, voice cloning aims to accurately capture and replicate the unique timbre, vocal habits, intonation fluctuations and prosodic characteristics of a target speaker with only a small amount of reference audio, thereby achieving efficient and natural customized speech generation unrestricted by text content and language types (Pallewatta et al., 2025). The task of cross-lingual voice cloning focuses on transferring speaker identity across different languages while preserving speech naturalness and intelligibility. Early speech synthesis and transcription technologies mainly rely on rule-based frameworks and manually curated customized datasets (Azzuni and Saddik, 2025). Although such methods deliver stable performance in standardized and constrained experimental environments, they fail to fully depict the diverse expressions of natural language, rich emotional nuances, and distinctive vocal features of individual speakers. With the rapid advancement of large-scale pre-trained speech models and generative large language models, a great number of powerful large models emerge, including Qwen3-TTS (Hu et al., 2026), Qwen3-Omni (Xu et al., 2025), IndexTTS2 (Zhou et al., 2026), CosyVoice3 (Du et al., 2025), OpenVoice2 (Qin et al., 2023) and VoxCPM (Zhou et al., 2025). General voice cloning technologies centered on few-shot, zero-shot and cross-lingual paradigms achieve rapid development, breaking the data constraints and language barriers of conventional voice cloning methods. These approaches enable speaker timbre transfer with extremely limited reference audio and significantly improve the generalization performance of speech synthesis systems.

Existing methods still face numerous technical challenges in complex scenarios (Huang et al., 2026). First, feature decoupling remains difficult. Speech signals contain multiple coupled attributes, including textual content, speaker timbre, emotion and prosody, making it hard for models to fully separate linguistic information from speaker identity features. Second, these methods lack robustness under few-shot and low-quality reference conditions. When reference audio is short or contaminated by background noise in low-resource settings, the generated speech commonly suffers from reduced timbre similarity and unnatural, robotic vocal artifacts. Collectively, these bottlenecks impose a clear upper limit on the overall naturalness and authenticity of cloned speech.

Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026), pages 97–102
July 3-4, 2026 ©2026 Association for Computational Linguistics

In the IWSLT 2026 Cross-Lingual Voice Cloning task, we adopt the end-to-end multilingual model Qwen3-TTS-12Hz-1.7B-Base as the voice cloning backbone for both Chinese and French target language tracks.

2 TTS Model

For the two tracks targeting Chinese and French as the target languages, we adopt the multilingual model Qwen3-TTS-12Hz-1.7B-Base as the core voice cloning model. First, it fully covers Chinese and French and exhibits reliable speech naturalness and content accuracy in bilingual scenarios. Second, its native zero-shot voice cloning capability matches our overall technical route, eliminating extra adaptation work. In addition, its streamlined end-to-end architecture simplifies system integration, which is favorable for fast iteration and deployment in the competition. The Qwen3-TTS speech synthesis model takes the Transformer architecture as its fundamental backbone and consists of three key components: a text encoding module, a speaker encoder, and an acoustic decoder. With a parameter scale of 1.7 billion, it realizes efficient prosody modeling and high-naturalness speech synthesis in end-to-end speech generation tasks. Through a decoupled speech representation system, Qwen3-TTS dynamically extracts, reorganizes and transfers acoustic features during the inference phase. The underlying design of Qwen3-TTS-12Hz-1.7B-Base thoroughly avoids the cascaded structure of acoustic models and vocoders adopted in conventional TTS frameworks. It utilizes the 16-layer multi-codebook encoder Qwen3-TTS-Tokenizer-12Hz to compress speech signals into discrete token sequences, and applies a multi-codebook language model to directly model the generation logic of these discrete tokens. Throughout the entire pipeline, speaker identity is no longer treated as a trainable classification label, but as a group of decoupled, editable and transferable latent vectors.

3 Methods

3.1 Data Preprocessing

High-quality and fine-grained audio data is the fundamental prerequisite for high-precision timbre cloning. The competition organizer provides 12 original reference audio clips for timbre, each lasting approximately five minutes. To obtain standardized reference samples suitable for voice cloning, we adopt an overlapping sliding window strategy for the fine-grained preprocessing of long audio. In our experiments, the window length is set to 10 seconds with a sliding step of 5 seconds. With this configuration, each reference audio is divided into around 60 segments. We keep all segments throughout the entire pipeline. Long reference audio contains subtle variations in speaking rate, timbre and prosody across different time periods, and retaining all segments enables us to fully capture the speaker's vocal characteristics and prevent incomplete feature coverage caused by partial sampling. While reducing the number of segments can cut computational costs, it may lead to the loss of critical fine-grained voice features and lower cloning similarity. Since our top priority is to achieve the best cloning performance, we choose to preserve all segments. This overlapping segmentation method fully retains the speaker's timbre, prosodic patterns and voiceprint details. It also expands the volume of reference samples and effectively avoids timbre feature loss resulting from non-overlapping cropping.

The detailed preprocessing pipeline is described as follows:

1) Traverse all WAV audio files in the original audio directory and load each file at its original sampling rate to avoid audio distortion and feature deviation introduced by resampling.

2) Calculate the total audio duration according to the total number of sampling points, and convert the time-based window size and sliding step into the corresponding number of sampling points to adapt to time-domain audio sequence processing.

3) Perform segmented clipping with the sliding window mechanism. Starting from the beginning of the audio, the framework extracts 10-second audio segments each time and slides forward by 5 seconds for continuous clipping until the window boundary exceeds the total audio length.

With the above preprocessing operations, each long original reference audio is divided into a large number of short standardized audio segments. This process constructs a diverse and comprehensive reference timbre sample library, provides sufficient data support for large-scale subsequent voice cloning generation, and effectively improves the stability and generalization performance of timbre replication.

3.2 TTS-Based Voice Cloning

We adopt the Qwen3-TTS-12Hz-1.7B-Base model to conduct the cross-lingual voice cloning task and uniformly complete voice cloning for both Chinese and French. The same implementation pipeline is applied to bilingual tasks.

1) Text corpus and reference sample construction. We utilize the official standard text data chinese.txt and french.txt provided by the competition. Each independent sentence in the files corresponds to one target speech to be synthesized. The original text order and semantic information are strictly retained to ensure that the generated content fully meets the task requirements. Meanwhile, we traverse all overlapping audio segments obtained through sliding window segmentation during data preprocessing and summarize all reference voice clips item by item to build a comprehensive voice reference task list. It covers vocal details and prosodic features of the target speaker at different periods and provides sufficient reference sources for diverse timbre replication.

2) Batch speech synthesis. Each standardized audio slice serves as an independent timbre reference to perform batch inference and generation for all text sentences. During the inference process, the model takes the input reference segment as a strong constraint of speaker timbre features and captures core voiceprint information such as vocal texture and speaking habits of the original speaker. With the text of the corresponding language as the semantic constraint, the model generates cross-lingual cloned speech with high timbre similarity.

This overall solution realizes large-scale batch generation that maps a single reference slice to multiple text utterances, greatly enriching the diversity and coverage of candidate samples. Massive differentiated cloning results provide a sufficient and high-quality candidate set for the subsequent optimal selection based on voiceprint embedding similarity, and effectively improve the timbre restoration performance and overall task robustness of the final selected audio.

3.3 Optimal Result Search

A large number of cloned speech samples are generated after multi-slice batch generation. To screen out the audio with the highest voiceprint consistency to the original speaker from candidate results, we conduct voiceprint recognition based on the Enhanced Context-Dependent Adversarial Time Delay Neural Network (ECAPA-TDNN) (Desplanques et al., 2020). Taking cosine similarity as the core quantitative evaluation metric, we complete the automatic selection of optimal results relying on objective indicators.

First, for the 12 original reference audio clips officially provided, three 10-second segments are randomly cropped from each audio file as the baseline timbre samples of the original speaker, which are used to uniformly evaluate the timbre restoration performance of all cloned audio.

A standardized optimal matching pipeline is implemented for each cloned speech sample:

1) Voiceprint feature extraction. The ECAPA-TDNN speaker verification model is adopted to extract fixed-dimensional voiceprint embedding vectors from the three baseline segments of the original audio and all cloned audio generated by different reference slices under the same text. This step converts timbre characteristics into quantifiable high-dimensional feature vectors.

2) Feature similarity calculation. Cosine similarity is utilized to measure the matching degree between the voiceprint vectors of cloned audio and the original baseline voiceprint vectors. It accurately describes the distribution difference in the feature space; a higher value indicates stronger consistency in timbre style and vocal characteristics with the original speaker.

3) Global optimal selection. For all generated audio corresponding to the same text and original speaker timbre, all similarity scores are compared. The audio with the peak cosine similarity is selected as the best cloning result for the current text content.

Through voiceprint feature extraction and quantitative similarity calculation, we construct an objective and reproducible timbre evaluation system. It efficiently realizes the refined filtering of massive generated samples and finally retains the speech results with the best timbre restoration, so as to guarantee the naturalness and voiceprint fidelity of the final output audio of the model.

4 Experiments and Results

We evaluate the BlindData from two perspectives: Content Consistency and Speaker Similarity.

Audio	CER(%)	Cos Sim
2023.acl-long.3	1.51	0.6405
2023.acl-long.6	1.93	0.6470
2023.acl-long.12	2.03	0.6197
2023.acl-long.14	1.93	0.4236
2023.acl-long.23	1.65	0.6143
2023.acl-long.67	2.19	0.6488
2023.acl-long.193	2.10	0.7052
2023.acl-long.289	2.17	0.5317
2023.acl-long.290	2.25	0.4430
2023.acl-long.293	2.10	0.5721
2023.acl-long.809	1.82	0.5389
2023.acl-long.810	1.39	0.6359

Table 1: Quantitative results of content consistency and speaker similarity for the chinese track.

Audio	WER(%)	Cos Sim
2023.acl-long.3	4.40	0.6242
2023.acl-long.6	5.01	0.5579
2023.acl-long.12	5.01	0.5985
2023.acl-long.14	5.87	0.5082
2023.acl-long.23	5.00	0.7173
2023.acl-long.67	4.44	0.5496
2023.acl-long.193	5.12	0.6674
2023.acl-long.289	8.52	0.6695
2023.acl-long.290	4.56	0.4420
2023.acl-long.293	5.09	0.5297
2023.acl-long.809	4.69	0.4758
2023.acl-long.810	4.56	0.6335

Table 2: Quantitative results of content consistency and speaker similarity for the french track.

4.1 Content Consistency

To evaluate content consistency, the generated audio is transcribed into text using the Qwen3-ASR-0.6B model (Shi et al., 2026), and the transcribed text is then compared with the ground-truth text. We calculate the Character Error Rate (CER) for Chinese results and the Word Error Rate (WER) for French results (Levenshtein et al., 1966).

As critical objective evaluation metrics, CER and WER effectively quantify the content fidelity of cross-lingual voice cloning outputs. A lower error rate indicates that the synthesized speech maintains high semantic integrity and linguistic accuracy, with fewer omission, substitution and insertion errors in text content.

As shown in Table 1 and Table 2, the CER values of Chinese cloned samples remain at a low level overall, indicating that the model steadily restores the original semantic information in Chinese cross-lingual generation scenarios. The WER of French synthesized audio is slightly higher than the CER of Chinese samples. This difference mainly arises from the inherent distinctions in pronunciation rules, lexical complexity, and prosodic characteristics between different languages.

Even so, all WER results are within a reasonable range, which verifies that our multi-slice batch generation and optimal screening strategy will not cause serious content distortion while ensuring timbre similarity. The stable and low error rates across bilingual datasets fully prove the strong cross-lingual content retention capability of the Qwen3-TTS model in voice cloning tasks.

4.2 Speaker Similarity

Speaker similarity measures the matching degree of voice features between generated outputs and the source speaker, calculated based on the cosine similarity of speaker embedding vectors extracted by the ECAPA-TDNN model.

As shown in Table 1 and Table 2, cosine similarity values (Cos Sim) effectively reflect the matching degree between generated audio and the original speaker's voice (Salton et al., 1975). For the Chinese track (Table 1), cosine similarity ranges from 0.4236 to 0.7052. Sample 2023.acl-long.193 achieves the highest similarity (0.7052), indicating the model effectively captured the speaker's core voice features, while sample 2023.acl-long.290 has the lowest (0.4430), possibly due to prosodic rhythm deviations during generation.

For the French track (Table 2), similarity ranges from 0.4420 to 0.7173, with a similar distribution to the Chinese track. Sample 2023.acl-long.23 reaches the highest (0.7173), slightly higher than the Chinese maximum, demonstrating the model's good adaptability in cross-lingual cloning. Notably, sample 2023.acl-long.289 has a high similarity (0.6695) despite its highest WER, showing the model maintains speaker consistency while ensuring content accuracy.

We further conduct an in-depth analysis on low-similarity failure cases. In the French track, sample 2023.acl-long.290 yields the lowest cosine similarity at 0.4420, and sample 2023.acl-long.809 also shows a below-average score of 0.4758. Different from general content errors reflected by WER, the performance degradation of these cases is mainly

attributed to cross-lingual prosodic mismatches. Since the reference audio and target generated audio belong to different languages, there are inherent differences in intonation, speaking rhythm, stress distribution and phonetic rhythm between languages. When converting the speaker's vocal characteristics across languages, the model cannot fully replicate the personalized prosodic habits of the original speaker, which distorts the extracted speaker embeddings and eventually leads to reduced similarity.

Overall, most samples in both tracks have a cosine similarity above 0.5, proving the method effectively realizes cross-lingual voice cloning. Slight similarity fluctuations, affected by audio prosody and reference voice clarity, are normal and do not affect overall performance.

5 Conclusion

In this work, we build a complete timbre cloning pipeline for the IWSLT 2026 voice challenge, including audio preprocessing, TTS-based bilingual voice cloning and optimal result selection.

We adopt a sliding window overlapping segmentation method to split long original reference audio into standardized short clips, which fully preserves the speaker's timbre and prosodic features and enriches reference sample diversity. Using the Qwen3-TTS-12Hz-1.7B-Base model, we conduct Chinese and French voice cloning in batch. With reference audio for timbre constraint and target text for content constraint, the one-to-many generation mode produces abundant differentiated candidate audio.

To select high-quality outputs, we adopt ECAPA-TDNN to extract speaker embeddings and calculate cosine similarity. Taking randomly cropped original audio as the timbre benchmark, we quantitatively evaluate cloning results and select the optimal audio for each text, eliminating subjective evaluation bias and improving timbre consistency.

Our proposed pipeline realizes stable and high-fidelity cross-lingual voice cloning, balancing generation efficiency and audio quality for competitive scenarios. Future work will optimize reference selection and timbre matching algorithms to further strengthen zero-shot voice cloning robustness.

References

Hussam Azzuni and Abdulmotaleb El Saddik. 2025. Voice cloning: Comprehensive survey. *arXiv preprint arXiv:2505.00579*.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. ECAPA-TDNN: Emphasized Channel Attention, propagation and aggregation in TDNN based speaker verification. In *Interspeech 2020*, pages 3830–3834.

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*.

Kexin Huang, Liwei Fan, Botian Jiang, Yaozhou Jiang, Qian Tu, Jie Zhu, Yuqian Zhang, Yiwei Zhao, Chenchen Yang, Zhaoye Fei, and 1 others. 2026. Moss-voicegenerator: Create realistic voices with natural language descriptions. *arXiv preprint arXiv:2603.28086*.

Vladimir I Levenshtein and 1 others. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Pandula Pallewatta, Samantha Mathara Arachchi, Adrian D Cheok, Sanam Rizvan, Kasun Karunanayaka, Trapp Kayuni, and Tinmei Aleksandr. 2025. Human voice synthesis and cloning using generative ai models: A comprehensive review of recent advances and applications. In *2025 7th International Conference on Electrical, Control and Instrumentation Engineering (ICECIE)*, pages 378–385. IEEE.

Zengyi Qin, Wenliang Zhao, Xumin Yu, and Xin Sun. 2023. Openvoice: Versatile instant voice cloning. *arXiv preprint arXiv:2312.01479*.

Gerard Salton, Anita Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, and 1 others. 2026. Qwen3-asr technical report. *arXiv preprint arXiv:2601.21337*.

Chengzhe Sun, Tianle Yang, and Siwei Lyu. 2026. A survey of ai-generated voices and their detection. *APSIPA Transactions on Signal and Information Processing*, 15(1):76–109.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, and 1 others. 2025. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.

Siyi Zhou, Yiquan Zhou, Yi He, Xun Zhou, Jinchao Wang, Wei Deng, and Jingchen Shu. 2026. Indextts2: A breakthrough in emotionally expressive and duration-controlled auto-regressive zero-shot text-to-speech. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 35139–35148.

Yixuan Zhou, Guoyang Zeng, Xin Liu, Xiang Li, Renjie Yu, Ziyang Wang, Runchuan Ye, Weiyue Sun, Jiancheng Gui, Kehan Li, and 1 others. 2025. Voxcpm: Tokenizer-free tts for context-aware speech generation and true-to-life voice cloning. *arXiv preprint arXiv:2509.24650*.