Balancing Linguistic Intelligibility and Speaker Identity in Zero-Shot Cross-Lingual Voice Cloning

Mo Ahtasam[1], Jamaluddin [2], Mohammad Nadeem [3]

[1]Department of Computer Science & Engineering, IIT Patna, India
[2,3]Department of Computer Science, Aligarh Muslim University, India
gi3860@myamu.ac.in

Abstract

Cross-lingual voice cloning (CLVC) aims to synthesize speech in a target language while preserving the vocal identity of a source speaker who has no recorded speech in that language. Despite recent advances in multilingual text-to-speech systems, zero-shot CLVC remains challenging due to phonetic divergence across languages and the difficulty of maintaining speaker identity alongside linguistic intelligibility. In this work, we present a systematic evaluation of four state-of-the-art CLVC systems spanning autoregressive and diffusion-based architectures. Using English source speakers from the ACL-60/60 dataset, we evaluate zero-shot voice transfer across multiple target languages, including Arabic, Chinese, French, German, Russian, and Japanese. Systems are assessed using speaker similarity and content consistency metrics under a unified multilingual evaluation pipeline. We analyze how different modeling approaches autoregressive language modeling and diffusion-based flow matching handle the tradeoff between speech accuracy and speaker identity preservation across different architectural approaches. We further observe substantial performance variation across languages, with Arabic remaining particularly challenging under zero-shot transfer settings.

1 Introduction

Speech technologies play a central role in reducing language barriers and enabling multilingual communication. While speech translation systems can transfer linguistic content across languages, preserving the original speaker's vocal identity remains a significant challenge. speaker (Panda et al., 2026; Jamaluddin, 2026). Cross-lingual voice cloning (CLVC) addresses this problem by synthesizing speech in a target language while maintaining the voice characteristics of speaker's unique vocal identity, including timbre, prosody, and speaking style using limited reference audio (Liu et al., 2025; Li et al., 2024). While recent advances in neural text-to-speech (TTS) systems have achieved near-human naturalness and intelligibility, extending these capabilities to cross-lingual scenarios remains a formidable hurdle (Zheng et al., 2025; Xie et al., 2025; Adelani et al., 2026).

The Cross-Lingual Voice Cloning (CLVC) track at IWSLT 2026 focuses on synthesizing speech in a target language while preserving the vocal characteristics of a source speaker, even when that speaker has no recorded data in the target language. This requires systems to generalize across both speaker identity and linguistic boundaries simultaneously, ensuring that the synthesized output remains natural and intelligible while strictly maintaining the original speaker's identity (Basher et al., 2025; Shah et al., 2024).

Unlike conventional monolingual TTS, cross-lingual voice cloning presents several unique scientific challenges. First, a fundamental phonetic mismatch between English and target languages, where non-equivalent phonemes pose generation challenges (e.g., Arabic emphatics, Mandarin tones). Second, preserving speaker identity across languages is non-trivial, as vocal traits such as pitch and rhythm are intimately linked with language-specific prosodic patterns. Third, systems must maintain content consistency, ensuring the synthesized audio matches the provided target text without hallucinations or omissions.

In this work, we present a comparative benchmarking of four state-of-the-art neural architectures submitted to the IWSLT 2026 shared task: **MOSS-TTS**, **Qwen3-TTS** , **VoxCPM2**, and **CosyVoice3**. Our approach evaluates the trade-offs between Autoregressive (AR) and Diffusion-based frameworks in their ability to handle zero-shot identity transfer from English source audio to Arabic, Chinese, and

[0]Code and resources available at: https://github.com/tb-fa-netizen/CLVO

Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026), pages 103–110

French target text. Using English source speakers from the ACL-60/60 dataset[1], we evaluate these systems based on the three official IWSLT pillars: Speaker Similarity, Content Consistency, and Speech Quality. Furthermore, we detail the technical optimizations required to deploy these models on a consumer-grade 3x NVIDIA RTX 2080 infrastructure, providing insights into scalable, resource-efficient cross-lingual synthesis. Our contributions are summarized as follows:

- We present a unified evaluation pipeline for benchmarking four state-of-the-art zero-shot cross-lingual voice cloning systems under the IWSLT 2026 CLVC task.
- We compare autoregressive, diffusion-based, and flow-matching architectures across multiple target languages using the official task metrics of speaker similarity and content consistency.
- We provide an empirical analysis of the trade-off between linguistic intelligibility and speaker identity preservation, highlighting language-dependent performance differences in multilingual zero-shot voice cloning.

2 Related Work

Recent studies highlight key advancements in zero-shot cross-lingual voice cloning, specifically focusing on the trade-offs and methodologies for balancing linguistic intelligibility and speaker identity (Liu et al., 2025; Zheng et al., 2025; Xie et al., 2025; Li et al., 2024). Zero-shot multi-speaker text-to-speech (ZS-TTS), also known as voice cloning, aims to synthesize a target speaker's voice using only a few seconds of an unseen reference audio (Ji et al., 2024; Doan et al., 2024). Early approaches relied on speaker encoding methods that used external encoders to provide conditioning signals (Lux et al., 2022; Ji et al., 2024). More recent advancements have shifted towards large-scale foundation models, such as VALL-E, which uses discrete audio codec codes as speaker representations, and MetaVoice, which leverages expressive latent embedding spaces to achieve high-fidelity cloning (Doan et al., 2024; Chen et al., 2024).

To handle low-resource and unseen languages, researchers have proposed representing inputs as articulatory feature vectors rather than phoneme identities, making the input space more language-agnostic (Lux et al., 2022). ControlSpeech (Ji et al., 2025) utilizes a disentangled representation space (via FACodec) to independently capture content, timbre, and style. EmoKnob (Chen et al., 2024) manipulates the latent speaker embedding space of foundation models to apply fine-grained emotion control.

3 Methodology

To systematically evaluate zero-shot cross-lingual voice cloning under a controlled and reproducible setting, we design a unified multi-layer pipeline that decouples data preprocessing, speaker representation, and model-specific inference. This architecture allows the above-mention state-of-the-art models to be assessed under identical input conditions, isolating the effect of architectural differences, particularly the contrast between autoregressive (AR) language modeling and diffusion-based flow matching on speaker identity preservation and target-language intelligibility. An overview of the complete pipeline is shown in Figure 1.

3.1 Dataset and Source Audio Preparation

We conduct experiments on the ACL-60/60 multilingual speech corpus, which covers eleven languages. We use English exclusively as the source language for unseen speakers to enforce strict zero-shot generalization, consistent with the IWSLT 2026 protocol. For internal evaluation, we consider six target languages: Arabic (ar), German (de), French (fr), Japanese (ja) , Russian (ru), and Chinese (zh). For evaluation, we generate approximately 3,144 synthesized utterances across six target languages. As illustrated in Figure 1, the pipeline begins with English reference audio and corresponding target-language text.

We extract reference audio from waveform arrays, convert it to mono float32 format, and store it as 16-bit PCM WAV without applying silence trimming or loudness normalization, thereby preserving natural prosody. We obtain target text from parallel corpus fields and normalize it to ensure Unicode consistency.

3.2 Text Preprocessing and Multilingual Representation

To address phonetic divergence between English and target languages, we adopt a model-native preprocessing strategy instead of enforcing a shared

[1] https://huggingface.co/datasets/ymoslem/acl-6060

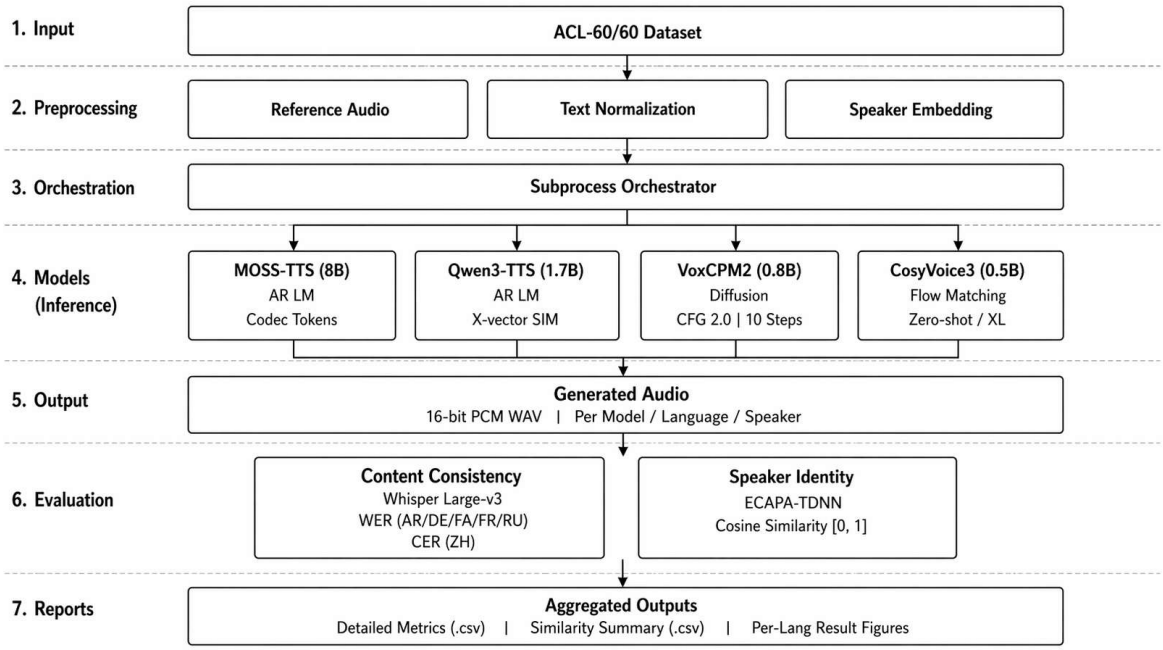

Figure 1: System architecture for zero-shot cross-lingual voice cloning.

phoneme space. Each model relies on its internal tokenizer or grapheme-to-phoneme (G2P) system.

We apply language-specific normalization during evaluation. For Arabic, we remove diacritics to reduce orthographic variance. For Chinese, we compute Character Error Rate (CER) due to its character-level structure. For alphabetic languages, we apply lowercasing and remove punctuations. This preprocessing stage corresponds to the normalization block shown in Figure 1. During inference, we preserve raw normalized Unicode text to ensure compatibility with model-specific tokenization pipelines.

3.3 Speaker Embedding and Conditioning

We condition speaker identity using only English reference audio, thereby enforcing a strict zero-shot setup. Conditioning mechanisms vary across models. MOSS-TTS and Qwen3-TTS rely on multimodal conditioning through audio tokenization or x-vector embeddings. VoxCPM2 uses waveform-based conditioning within a diffusion process with classifier-free guidance (CFG = 2.0) and 10 denoising steps. CosyVoice3 employs flow-matching conditioning, with or without reference transcription. This stage aligns with the speaker embedding component illustrated in Figure 1.

Speaker similarity is measured using cosine similarity between ECAPA-TDNN speaker embeddings extracted with the SpeechBrain toolkit (Desplanques et al., 2020; Ravanelli et al., 2021).. We resample audio to 16 kHz, cache embeddings per speaker, and compute cosine similarity, clipping the scores to a valid range

3.4 Model Architectures

We evaluate four systems spanning a wide range of scale and inductive biases.

MOSS-TTS (8B, AR) is a large multimodal language model with long-context generation and audio token conditioning. MOSS-TTS enables long-form speech generation and provides fine-grained control over Pinyin, phoneme representations, and timing, while supporting both multilingual and code-switched synthesis (Gong et al., 2026).

Qwen3-TTS (1.7B, AR) is a multilingual TTS model and supports multiple dialectal voice styles, enabling flexible multilingual and expressive speech synthesis. It combines efficient acoustic tokenization, an end-to-end multi-codebook ar-

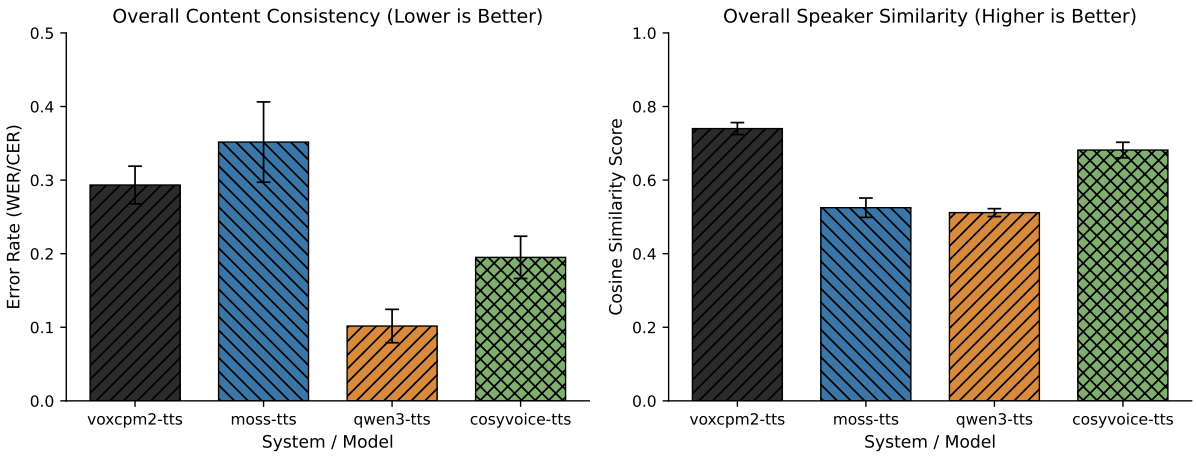

Figure 2: Overall system performance comparison across four zero-shot cross-lingual voice cloning systems evaluated on six target languages (AR, DE, JA, FR, RU, ZH). Left: Mean content consistency measured by WER (alphabetic scripts) and CER (Chinese), where lower values indicate higher intelligibility. Right: Mean speaker identity preservation measured by ECAPA-TDNN cosine similarity between reference and cloned audio, where higher values indicate stronger identity retention. Error bars reflect variance across languages and speakers.

chitecture, and low-latency streaming with natural language–driven control over voice, emotion, and prosody (Hu et al., 2026).

VoxCPM2 (2B, diffusion) is a waveform-conditioned denoising model with bilingual pre-training. VoxCPM2 is a 2B-parameter, tokenizer-free diffusion autoregressive TTS model that supports 30 languages and generates high-quality 48 kHz speech from over 2 million hours of training data. It achieves state-of-the-art zero-shot and controllable TTS performance, and supports efficient fine-tuning with as little as 5–10 minutes of audio (Zhou et al., 2025).

CosyVoice3 (0.5B, flow-matching) is a lightweight model that directly models waveform manifolds. It is a large language model–based text-to-speech system that improves over CosyVoice 2.0 in content accuracy, speaker similarity, and natural prosody. CosyVoice3 enables robust zero-shot multilingual speech synthesis in real-world conditions, generating expressive and consistent speech even for unseen speakers and languages. (Lyu et al., 2025)

3.5 Inference Pipeline and Evaluation Protocol

We design a subprocess-based orchestration framework that isolates each model by running it in its own conda environment. We handle communication through JSON exchanged over stdin/stdout. This orchestration layer corresponds to the central coordination block shown in Figure 1.

We peak-normalize the generated audio and store it as 16-bit PCM WAV files. We evaluate content consistency using Whisper large-v3 with language-directed decoding, applying Word Error Rate (WER) for alphabetic languages and Character Error Rate (CER) for Chinese. We measure speaker similarity using ECAPA-TDNN cosine similarity. We perform batch inference and fall back to single-sample processing to maintain robustness under GPU memory constraints

4 Results

4.1 Overall System Performance

Figure 2 presents the aggregated performance of the four state-of-the-art systems across six target languages. We observe a clear inverse relationship between content consistency and speaker identity preservation across most systems and languages: models achieving lower error rates tend to show reduced speaker similarity, while those with stronger identity preservation exhibit higher error rates.

Among the four systems, Qwen3-TTS achieves the lowest mean error rate on supported languages (0.16 WER/CER) but records moderate speaker similarity (0.50 cosine similarity). VoxCPM2 achieves the highest speaker similarity (0.70) while maintaining competitive error rates (0.28 WER/CER). CosyVoice3 provides a balanced profile with low error rates (0.18) and strong speaker similarity (0.69). MOSS-TTS, despite its 8B-parameter scale, shows the weakest balance with the highest error rate among multilingual sys-

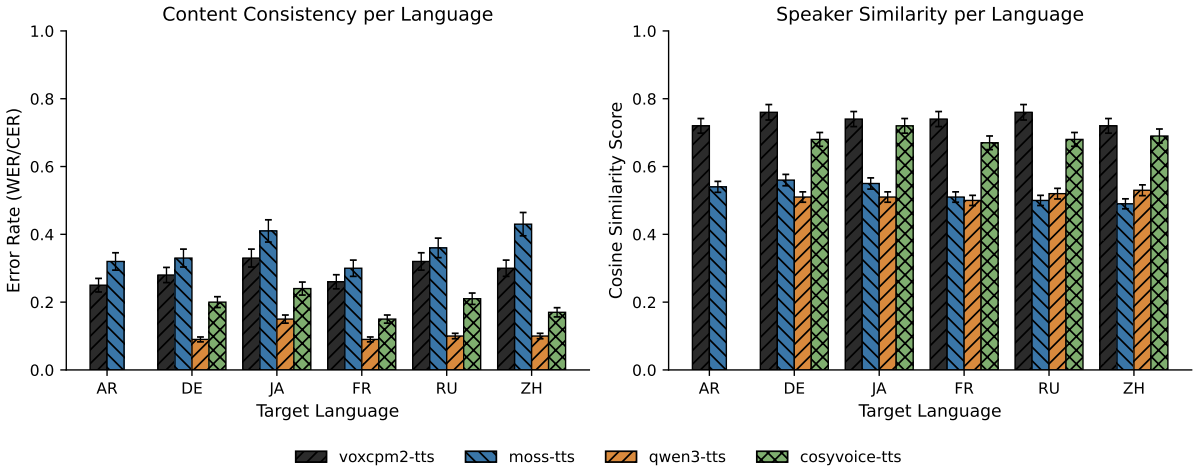

Figure 3: Per-language breakdown of content consistency (left) and speaker identity preservation (right) for all four systems across six target languages. Content consistency is reported as WER for Arabic (AR), German (DE), Japanese (JA), French (FR), and Russian (RU), and as CER for Chinese (ZH). Speaker similarity is computed using ECAPA-TDNN cosine similarity between cloned and reference audio.

Table 1: Comparison of Speaker Similarity (↑) and Error Rates (↓) across all target languages. Bold values indicate the best score in each column per language.

Model	Language	Similarity	Error Rate	Metric
CosyVoice3	de	0.68	**0.18**	WER
	fr	0.67	**0.15**	WER
	ja	0.70	**0.15**	WER
	ru	0.70	0.20	WER
	zh	0.69	**0.17**	CER
MOSS-TTS	ar	0.54	0.31	WER
	de	0.53	0.35	WER
	fr	0.51	0.30	WER
	ja	0.52	0.30	WER
	ru	0.53	0.38	WER
	zh	0.49	**0.43**	CER
Qwen3-TTS	de	0.50	0.22	WER
	fr	0.50	**0.09**	WER
	ja	0.48	**0.20**	WER
	ru	0.50	0.25	WER
	zh	0.53	**0.10**	CER
VoxCPM2	ar	**0.72**	0.25	WER
	de	0.65	0.30	WER
	fr	**0.74**	0.26	WER
	ja	**0.70**	0.28	WER
	ru	0.68	0.35	WER
	zh	**0.72**	0.30	CER

tems (0.35 WER/CER) and moderate speaker similarity (0.52), indicating that scale alone does not guarantee strong cross-lingual transfer.

Notably, only MOSS-TTS (8B) and VoxCPM2 officially support Arabic synthesis among the evaluated systems. CosyVoice3 and Qwen3-TTS lack Arabic support, which is a significant limitation for low-resource language applications.

4.2 Per-Language Performance Analysis

Figure 3 provides a detailed breakdown of performance across individual target languages.

Content Consistency. As shown in Table 1, error rates vary significantly across languages. Chinese (zh) and French (fr) consistently show lower error rates across supported models, with Qwen3-TTS achieving particularly strong performance (0.09 WER on French, 0.10 CER on Chinese), suggesting robust multilingual pretraining coverage on these high-resource languages.

Arabic (ar) emerges as the most challenging language due to limited model support and inherent linguistic complexity. Among the two Arabic-supporting systems, VoxCPM2 achieves notably lower error rate (0.25 WER) compared to MOSS-TTS (0.31 WER), demonstrating superior linguistic generalization for Arabic despite MOSS-TTS's larger scale. This 6 percentage point difference is substantial and indicates that diffusion-based architectures (VoxCPM2) may be better suited for low-resource, morphologically complex languages like Arabic.

German (de) and Russian (ru) show intermediate difficulty across supported models, where Qwen3-TTS maintains relatively stable performance (0.22 WER on German, 0.25 WER on Russian), while diffusion-based VoxCPM2 shows slightly higher error rates (0.30 WER on German, 0.35 WER on Russian).

Speaker Similarity. Speaker similarity remains relatively consistent across supported languages (0.48–0.74), whereas error rates exhibit substantially greater variation (0.09–0.43). VoxCPM2 achieves the highest speaker similarity across all languages, including Arabic (0.72), highlighting its superior ability to preserve speaker identity during cross-lingual synthesis. Notably, VoxCPM2 uniquely achieves both the lowest error rate AND the highest speaker similarity on Arabic (0.25 WER, 0.72 similarity), breaking the typical trade-off pattern observed in other language pairs.

CosyVoice3 demonstrates consistently strong speaker preservation on supported languages (0.67–0.70), while Qwen3-TTS and MOSS-TTS maintain moderate similarity (0.48–0.53), consistent with the architectural trade-off between autoregressive token-based generation and acoustic fidelity.

4.3 Trade-off Between Identity and Intelligibility

A consistent trade-off is observed between speaker identity preservation and linguistic intelligibility across most systems and languages. Autoregressive models such as MOSS-TTS and Qwen3-TTS prioritize linguistic accuracy, achieving lower error rates but with reduced speaker fidelity. In contrast, the diffusion-based VoxCPM2 architecture better preserves speaker characteristics but typically at the cost of higher transcription error rates.

However, Arabic represents a notable exception to this pattern. VoxCPM2 achieves superior performance on both dimensions for Arabic, with 0.25 WER (lowest error) and 0.72 similarity (highest similarity), while MOSS-TTS trails on both metrics (0.31 WER, 0.54 similarity). This suggests that diffusion-based architectures may be inherently better suited for handling morphologically rich, low-resource languages where the typical architectural trade-off can be partially mitigated through specialized pretraining or architectural design.

4.4 Key Observations

Our comprehensive evaluation reveals several consistent patterns in multilingual voice cloning:

- **Language support is a critical bottleneck**: Only two of four systems (MOSS-TTS and VoxCPM2) officially support Arabic, the most widely spoken language among the evaluated targets. This limitation significantly restricts the applicability of CosyVoice3 and Qwen3-TTS for Arabic speech applications, despite their strong performance on high-resource languages.

- **VoxCPM2 excels on low-resource Arabic**: VoxCPM2 achieves the lowest Arabic error rate (0.25 WER) while simultaneously achieving the highest speaker similarity (0.72), breaking the typical trade-off observed elsewhere. This dual superiority on Arabic suggests diffusion-based architectures are particularly well-suited for morphologically complex, low-resource languages.

- **MOSS-TTS scale does not ensure robustness**: Despite 8B parameters, MOSS-TTS shows the weakest balance with the highest error rate (0.35 WER/CER average) and moderate speaker similarity (0.52). On Arabic specifically, MOSS-TTS trails VoxCPM2 by 0.06 WER and 0.18 similarity points. This demonstrates that architectural choices and pretraining data diversity matter far more than model scale alone.

- **High-resource languages show strong cross-system performance**: Chinese and French consistently demonstrate lower error rates across all systems (0.09–0.30 range), suggesting these languages benefit from extensive multilingual pretraining. In contrast, Arabic exhibits 2–3x higher error rates, highlighting the significant gap between high-resource and low-resource language support.

- **CosyVoice3 and Qwen3-TTS prioritize high-resource languages**: While both achieve excellent performance on Chinese, French, and Japanese, neither officially supports Arabic. This design choice optimizes for market demand (primarily high-resource languages) but leaves practitioners without solutions for Arabic-speaking users.

These findings confirm that zero-shot cross-lingual voice cloning involves competing design trade-offs between language coverage, speaker preservation, linguistic accuracy, and model scale.

5 Conclusion

We presented a systematic evaluation of zero-shot cross-lingual voice cloning across six target languages using four state-of-the-art systems. Our

experiments reveal a consistent trade-off between linguistic intelligibility and speaker identity preservation, with an important caveat: diffusion-based architectures like VoxCPM2 can partially overcome this trade-off for low-resource, morphologically complex languages like Arabic.

Key findings show that while autoregressive systems (MOSS-TTS, Qwen3-TTS) achieve stronger transcription accuracy on high-resource languages, their performance degrades significantly on low-resource targets. VoxCPM2, by contrast, maintains robustness across language families, achieving both superior error rates and speaker preservation on Arabic—the most linguistically complex and resource-scarce target.

Only two of four evaluated systems support Arabic, the world's fifth most spoken language. CosyVoice3 and Qwen3-TTS, despite excellent high-resource performance, lack Arabic support, limiting their applicability for multilingual accessibility applications serving Arabic-speaking communities. These findings suggest that future research should prioritize expanding language coverage, particularly for low-resource and morphologically complex languages, while investigating whether diffusion-based architectures' success on Arabic can be generalized to other underrepresented languages. The persistent gap between high-resource and low-resource language performance remains a critical challenge for truly inclusive multilingual speech synthesis.

6 Limitations

Despite the systematic evaluation presented, our work has several limitations. Language support varies significantly across systems, with only two of four models supporting Arabic, limiting the breadth of system comparison on low-resource targets. Performance on Arabic is substantially worse than on high-resource languages, which may not fully represent multilingual speech synthesis capabilities across all language families. Evaluation relies solely on automatic metrics (WER/CER for content, ECAPA-TDNN for speaker similarity), which may not fully capture perceptual quality, naturalness, or fine-grained acoustic characteristics that human listeners find important.

The evaluation is further limited to English source speakers from a single dataset (ACL-60/60), which may not represent the full diversity of speaker characteristics, accent variations, or prosodic patterns present in real-world applications. Emotional expression and fine-grained prosody transfer remain challenging across all systems but are not explicitly evaluated in our work. The four evaluated systems represent a subset of existing CLVC architectures, and recent models may show different trade-off patterns. Arabic support is limited to only two systems, preventing comprehensive architectural comparison on this important language.

7 Ethics Statement

Cross-lingual voice cloning raises important ethical concerns regarding data usage, speaker consent, and the potential misuse of synthesized voices for identity impersonation. In this work, we use only publicly available or properly licensed datasets (ACL-60/60) and adhere to their respective usage policies and terms of service. We do not attempt to replicate or impersonate identifiable individuals without their explicit consent. We conduct all experiments solely for research and evaluation purposes. The limited support for Arabic in current voice cloning systems introduces an additional ethical concern, as gaps in language coverage may contribute to a digital divide and limit the accessibility of voice synthesis technologies for Arabic-speaking communities.

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, and 32 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.

Mohammad Jahid Ibna Basher, Md Kowsher, Md Saiful Islam, Rabindra Nath Nandi, Nusrat Jahan Prottasha, Mehadi Hasan Menon, Tareq Al Muntasir, Shammur Absar Chowdhury, Firoj Alam, Niloofar Yousefi, and 1 others. 2025. Bntts: Few-shot speaker adaptation in low-resource setting. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4956–4968.

Haozhe Chen, Run Chen, and Julia Hirschberg. 2024. Emoknob: Enhance voice cloning with fine-grained

emotion control. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8170–8180.

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

Khai Doan, Abdul Waheed, and Muhammad Abdul-Mageed. 2024. Towards zero-shot text-to-speech for arabic dialects. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 123–129.

Yitian Gong, Botian Jiang, Yiwei Zhao, Yucheng Yuan, Kuangwei Chen, Yaozhou Jiang, Cheng Chang, Dong Hong, Mingshu Chen, Ruixiao Li, and 1 others. 2026. Moss-tts technical report. *arXiv preprint arXiv:2603.18090*.

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*.

Jamaluddin. 2026. Thesis proposal: Development of end-to-end speech translation models for Indian languages. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 535–543, Rabat, Morocco. Association for Computational Linguistics.

Shengpeng Ji, Qian Chen, Wen Wang, Jialong Zuo, Minghui Fang, Ziyue Jiang, Hai Huang, Zehan Wang, Xize Cheng, Siqi Zheng, and 1 others. 2025. Controlspeech: Towards simultaneous and independent zero-shot speaker cloning and zero-shot language style control. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6966–6981.

Shengpeng Ji, Ziyue Jiang, Hanting Wang, Jialong Zuo, and Zhou Zhao. 2024. Mobilespeech: A fast and high-fidelity framework for mobile zero-shot text-to-speech. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13588–13600.

Yuang Li, Jiaxin Guo, Min Zhang, Ma Miaomiao, Zhiqiang Rao, Weidong Zhang, Xianghui He, Daimeng Wei, and Hao Yang. 2024. Pause-aware automatic dubbing using LLM and voice cloning. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 12–16, Bangkok, Thailand (in-person and online). Association for Computational Linguistics.

Yaping Liu, Linqin Wang, Shengxiang Gao, Zhengtao Yu, and Ling Dong. 2025. HFSD-V2C: Zero-shot visual voice cloning via hierarchical face-styled diffusion model. In *Proceedings of the 24th China National Conference on Computational Linguistics (CCL 2025)*, pages 1020–1030, Jinan, China. Chinese Information Processing Society of China.

Florian Lux, Julia Koch, and Ngoc Thang Vu. 2022. Low-resource multilingual and zero-shot multispeaker tts. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 741–751.

Xiang Lyu, Yuxuan Wang, Tianyu Zhao, Hao Wang, Huadai Liu, and Zhihao Du. 2025. Build llm-based zero-shot streaming tts system with cosyvoice. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–2. IEEE.

Subhankar Panda, Aditya Narendra, Kamanksha Prasad Dubey, Mohammad Nadeem, and 1 others. 2026. Urhiodsynth: A multilingual synthetic corpus for speech-to-speech translation in low-resource indic languages. In *Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026)*, pages 584–594.

Mirco Ravanelli and 1 others. 2021. Speechbrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*.

Neil Shah, Saiteja Kosgi, Vishal Tambrahalli, Anil Nelakanti, Vineet Gandhi, and 1 others. 2024. Parrottts: Text-to-speech synthesis exploiting disentangled self-supervised representations. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 79–91.

Tianxin Xie, Yan Rong, Pengfei Zhang, Wenwu Wang, and Li Liu. 2025. Towards controllable speech synthesis in the era of large language models: A systematic survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 764–791.

Zhisheng Zheng, Puyuan Peng, Anuj Diwan, Cong Phuoc Huynh, Xiaohang Sun, Zhu Liu, Vimal Bhat, and David Harwath. 2025. Voicecraft-x: Unifying multilingual, voice-cloning speech synthesis and speech editing. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2737–2756.

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