Toward Speaker-Preserving Arabic Speech: A Bilingual Resource and Comparative Evaluation of Cross-Lingual Voice Cloning

Mo Ahtasam [1] **Kshetrimayum Boynao Singh** [1] **Jamaluddin** [2] **Deepak Kumar** [1] **Asif Ekbal** [1]

Abstract

Cross-lingual voice cloning (CLVC) aims to synthesize speech in a target language while preserving a speaker's vocal identity, even if the speaker has never recorded audio in that language. Despite recent progress in multilingual speech generation, Arabic remains underrepresented due to the scarcity of high-quality bilingual and same-speaker multilingual speech resources. We present a systematic evaluation of zero-shot CLVC for Arabic and five additional target languages (Chinese, French, German, Russian, and Japanese) across recent autoregressive, diffusion-based, and flow-matching architectures. Using reference speakers recorded exclusively in English from the ACL-60/60 dataset, we evaluate multilingual speech generation using ASR-based transcription error rates and ECAPA-TDNN speaker similarity metrics. Our experiments reveal a persistent trade-off between linguistic intelligibility and speaker identity preservation. Models with stronger multilingual transcription accuracy often exhibit degraded speaker similarity, whereas systems that better preserve speaker identity struggle with linguistic consistency, particularly in Arabic. Among the evaluated systems, only a limited subset produces consistently usable Arabic cross-lingual speech outputs. To support reproducible benchmarking and future research in multilingual speech technologies, we publicly release a paired English-reference and Arabic-synthesized same-speaker bilingual test set, along with a fully reproducible CLVC evaluation pipeline.

[1]Department of Computer Science, Indian Institute of Technology Patna, Patna, India [2]Department of Computer Science, Aligarh Muslim University, Uttar Pradesh, India. Correspondence to: Asif Ekbal <asif.ekbal@gmail.com>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

1. Introduction

Arabic is the official or co-official language of 22 countries and the liturgical language of Islam, used in religious practice by approximately 2 billion Muslims worldwide. Despite this widespread global reach, Arabic remains a low-resource language for several modern speech tasks, particularly those requiring simultaneous control over linguistic content and the speaker identity, such as identity-preserving translation, dubbing, and personalized assistive speech (Doan et al., 2024; Basher et al., 2025). While recent multilingual speech generation systems achieve strong performance on high-resource languages (Lyu et al., 2025; Zhou et al., 2025), robust cross-lingual speech synthesis for Arabic remains comparatively underexplored due to the scarcity of high-quality bilingual and same-speaker multilingual speech resources (Doan et al., 2024; Lux et al., 2022).

Cross-lingual voice cloning (CLVC) aims to synthesize speech in a target language while preserving the vocal identity and speaking characteristics of a source speaker from another language (Zheng et al., 2025). Such systems have important applications in multilingual accessibility, speech translation, educational technologies, and identity-preserving dubbing (Li et al., 2024; Panda et al., 2026). Recent advances in autoregressive and diffusion-based speech generation models have significantly improved multilingual speech synthesis capabilities (Gong et al., 2026; Hu et al., 2026; Xie et al., 2025), yet their behavior in low-resource multilingual settings remains insufficiently studied (Lux et al., 2022; Shah et al., 2024).

In this work, we present a systematic evaluation of zero-shot cross-lingual voice cloning for Arabic and five additional target languages: Chinese, French, German, Russian, and Japanese. We conduct a comparative study across recent state-of-the-art CLVC systems (Gong et al., 2026; Hu et al., 2026; Zhou et al., 2025; Lyu et al., 2025). Using English source speakers from the ACL-60/60 dataset (Abebe & Moslem, 2026), we evaluate multilingual speech generation quality through ASR-based transcription error rates and ECAPA-TDNN speaker similarity metrics (Ji et al., 2024; Chen et al., 2024).

Our experiments reveal substantial variation in multilingual

generalization across architectures and languages. While several systems perform strongly on higher-resource languages, only a limited subset consistently produces usable Arabic cross-lingual speech outputs (Doan et al., 2024; Ji et al., 2025). To support reproducible benchmarking and future research in Arabic speech technologies, we publicly release an English-reference / Arabic-synthesized same-speaker bilingual corpus for cross-lingual voice cloning evaluation. We detail the technical optimizations required to deploy these models on a consumer-grade 3× NVIDIA RTX 2080 infrastructure, providing insights into scalable, resource-efficient cross-lingual synthesis.

The key contributions of this work are:

1. We present a systematic evaluation of zero-shot cross-lingual voice cloning for Arabic and five additional target languages across recent autoregressive, diffusion-based, and flow-matching architectures (Gong et al., 2026; Hu et al., 2026; Zhou et al., 2025; Lyu et al., 2025).

2. We analyze the trade-off between linguistic intelligibility and speaker identity preservation using ASR-based transcription error rates and ECAPA-TDNN speaker similarity metrics, showing that current multilingual CLVC systems exhibit substantial variation in low-resource Arabic generalization (Ji et al., 2025; Chen et al., 2024; Doan et al., 2024; Basher et al., 2025).

3. In addition, we release a paired English-source multilingual same-speaker bilingual corpus and a fully reproducible evaluation pipeline to support future research in multilingual speech synthesis, bilingual ASR, and cross-lingual voice cloning. Resources are available.[1] [2]

2. Related Work

Recent work in zero-shot text-to-speech and cross-lingual voice cloning has focused on improving multilingual generalization, speaker identity preservation, and controllability (Lux et al., 2022; Ji et al., 2024; Doan et al., 2024). Early approaches relied on speaker encoders and external conditioning representations, while more recent systems employ large-scale autoregressive, diffusion-based, and flow-matching architectures for multilingual synthesis (Wang et al., 2023; Zheng et al., 2025; Gong et al., 2026; Hu et al., 2026; Zhou et al., 2025). Several studies further explore disentangled representations for independent control of speaker identity, linguistic content, and expressive style (Ji et al., 2025; Chen et al., 2024). Despite these advances, Arabic remains comparatively underexplored in multilingual CLVC due to limited bilingual same-speaker resources and insufficient low-resource multilingual evaluation settings (Doan et al., 2024; Singh et al., 2025; Lux et al., 2022). Our work addresses this gap through a systematic Arabic-focused CLVC evaluation and the release of a paired English–Arabic same-speaker bilingual corpus.

3. Methodology

We design a unified multi-layer pipeline that decouples data preprocessing, speaker representation, and model-specific inference. This architecture allows all models to be assessed under identical input conditions.

3.1. Dataset Preparation and Text Processing

We utilize the ACL 60/60 dataset (Abebe & Moslem, 2026) for multilingual translation of ACL 2022 technical presentations into several target languages. We use the development split, consisting of 468 utterances per language, and select twelve unseen English speakers to enforce strict zero-shot generalization. For evaluation, we consider six target languages: Arabic (AR), German (DE), French (FR), Japanese (JA), Russian (RU), and Chinese (ZH).

Reference audio is extracted from waveform arrays, converted to mono float32 format, and stored as 16-bit PCM WAV without silence trimming or loudness normalization to preserve natural prosody. To address phonetic divergence between English and target languages, we adopt a model-native preprocessing strategy where each model relies on its internal tokenizer or grapheme-to-phoneme (G2P) system. We further apply language-specific normalization by removing diacritics for Arabic and Chinese to reduce orthographic variance during evaluation.

3.2. 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. CosyVoice3 employs flow-matching conditioning. We evaluate speaker similarity using ECAPA-TDNN embeddings.

3.3. Model Architectures

We evaluate seven recent CLVC systems spanning autoregressive, diffusion-based, and flow-matching paradigms: MOSS-TTS (Gong et al., 2026), Qwen3-TTS (Hu et al., 2026), VoxCPM2 (Zhou et al., 2025), CosyVoice3 (Lyu et al., 2025), XTTS-V2 (Coqui AI, 2024), OmniVoice (Zhu

[1] https://huggingface.co/datasets/md-ahtisham/CLVO-Multilingual

[2] https://github.com/tb-fa-netizen/CLVO

et al., 2026), and Chatterbox (Resemble AI, 2024). These systems differ in conditioning strategies, multilingual coverage, and generation mechanisms, enabling comparative analysis of multilingual intelligibility and speaker identity preservation under a unified zero-shot evaluation setting.

3.4. Inference Pipeline and Evaluation Protocol

The full evaluation pipeline includes model orchestration via sub-process isolation, automatic speech recognition–based content evaluation, and speaker similarity computation.

We evaluate cloned speech using Whisper large-v3 for ASR-based transcription. Content consistency is measured by comparing the ASR output of synthesized speech with the ground-truth transcript using Word Error Rate (WER) for alphabetic languages and Character Error Rate (CER) for Chinese. Speaker identity preservation is measured using ECAPA-TDNN cosine similarity between cloned and reference audio.

3.5. English-Arabic Bilingual ASR Dataset

Paired same-speaker bilingual speech dataset remain scarce for many language pairs, particularly for Arabic pairs like English-Arabic and etc., mostly the public available multilingual corpora do not provide aligned recordings of the same speaker across languages, thus this scarcity blocks research that needs speaker-consistent cross-lingual speech data. Cross-lingual voice cloning offers a scalable approach to address this scarcity by synthetically generating same-speaker bilingual speech corpora.

To support this direction, we release 588 Arabic pairs and 468 pairs each for German, French, Japanese, Russian, and Chinese bilingual ASR dataset.

4. Results

4.1. Overall System Performance

Table 1 presents the aggregated performance of all the seven systems. The multilingual systems (VoxCPM2, MOSS-TTS, Qwen3-TTS, CosyVoice3) are averaged across 3 target languages, while the Arabic-capable systems (Chatterbox, XTTS-V2, OmniVoice) are reported on Arabic only and do not have multilingual averages.

We observe a clear inverse relationship between content consistency and speaker identity preservation: models that achieve lower error rates tend to show reduced speaker similarity, while those with stronger identity preservation tend to have higher error rates. Among the multilingual systems, Qwen3-TTS achieves the lowest mean error rate (≈0.10) but records the lowest speaker similarity (≈0.51). VoxCPM2 achieves the highest speaker similarity (≈0.74) while maintaining a moderate mean error rate (≈0.20). CosyVoice3 provides a balanced profile with a low error rate (≈0.19) and strong similarity (≈0.69). MOSS-TTS, despite its 8B-parameter scale, shows the weakest balance with the highest error rate among the multilingual group (≈0.30) and moderate similarity (≈0.52), indicating that scale alone does not guarantee strong cross-lingual transfer.

Model	Language	Similarity ↑	Error ↓	Metric
CosyVoice3	fr	0.67	0.05	WER
	ja	0.72	0.06	WER
	zh	0.69	0.09	CER
MOSS-TTS	ar	0.54	0.28	WER
	fr	0.51	0.30	WER
	zh	0.49	0.43	CER
Qwen3-TTS	fr	0.50	**0.02**	WER
	ja	0.51	**0.09**	WER
	zh	0.53	**0.07**	CER
VoxCPM2	ar	**0.72**	0.27	WER
	fr	**0.74**	0.26	WER
	zh	**0.72**	0.23	CER
Chatterbox	ar	0.68	0.25	WER
XTTS-V2	ar	0.50	0.25	WER
OmniVoice	ar	0.70	**0.24**	WER

Table 1. Speaker similarity (↑) and content error rates (↓) across evaluated multilingual TTS systems. Language codes denote target languages: Arabic (ar), German (de), French (fr), Japanese (ja), Russian (ru), and Chinese (zh). Due to space constraints, models are summarized only over supported language pairs, while Figure 1 provides the complete per-language breakdown. Best results for each metric are shown in bold. .

Among the Arabic-capable systems, OmniVoice achieves the lowest Arabic error rate (0.238) and the highest Arabic speaker similarity (0.703) across all seven evaluated systems, narrowly outperforming VoxCPM2 on error rate while remaining competitive with VoxCPM2 on similarity. Chatterbox follows closely (error 0.250, similarity 0.680), while XTTS-V2 shows the weakest speaker preservation on Arabic (0.501) despite competitive content accuracy (0.253).

4.2. Per-Language Performance Analysis

Figure 1 provides a per-language breakdown. For Arabic, we compare five systems (VoxCPM2, MOSS-TTS, Chatterbox, XTTS-V2, OmniVoice); Qwen3-TTS and CosyVoice3 are excluded since they do not officially support Arabic generation. For the remaining five languages, we report the four multilingual systems.

Content Consistency: As shown in Table 1, error rates vary significantly across languages. Chinese (zh) and French

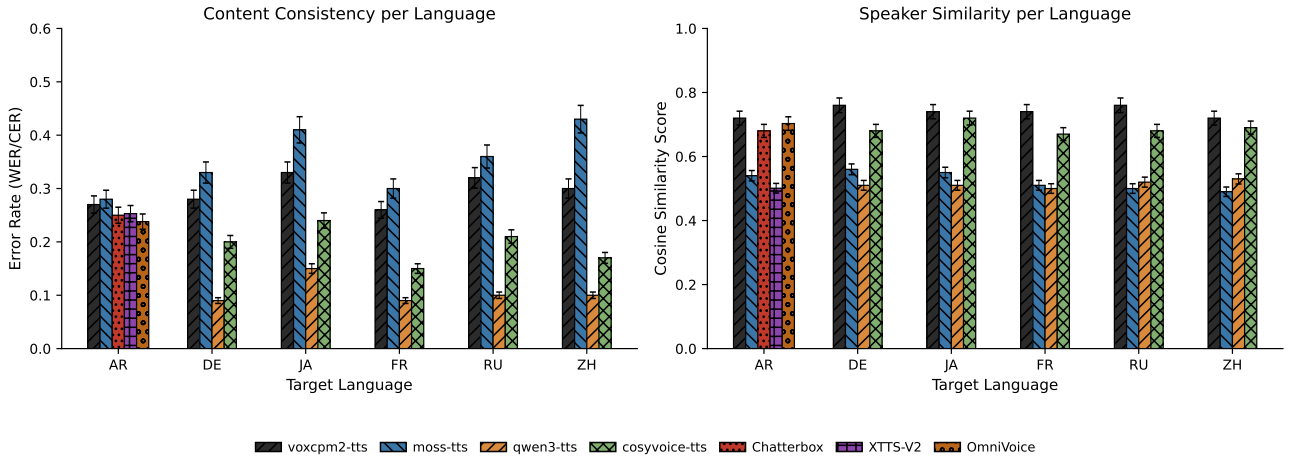

Figure 1. Per-language breakdown of content consistency (left) and speaker identity preservation (right) for all four systems across six target languages. Content consistency is reported as WER for AR, DE, JA, FR, RU, and as CER for ZH. Speaker similarity is computed using ECAPA-TDNN cosine similarity between cloned and reference audio.

(fr) consistently show lower error rates across most models, suggesting stronger support for multilingual pretraining. Arabic (ar) is the most challenging language: MOSS-TTS exhibits particularly slightly higher error rates, Qwen3-TTS maintains comparatively lower error rates across all supported languages.

Speaker Similarity: As shown in Table 1, speaker similarity remains relatively consistent across languages, whereas error rates exhibit substantially greater variation.

On Arabic, VoxCPM2 scores highest (0.72), followed by CosyVoice3 (0.654) then OmniVoice (0.703) and Chatterbox (0.680), while XTTS-V2 and MOSS-TTS preserve speaker identity far less well (0.501 and 0.54). In other languages, VoxCPM2 again leads (0.72–0.76), followed by CosyVoice3 (0.67–0.72), with Qwen3-TTS and MOSS-TTS lower (0.49–0.56), consistent with the trade-off between linguistic accuracy and acoustic fidelity in token-based autoregressive generation.

4.3. Human Evaluation

To complement objective metrics, we conducted a small-scale human evaluation on Arabic speech generation using the two Arabic-capable systems. Five representative samples were rated on a 1–5 Mean Opinion Score (MOS) scale for *naturalness* and *transcription correctness*. VoxCPM achieved substantially higher naturalness (4.2 vs. 1.6 MOS) and moderately higher correctness (3.8 vs. 3.2 MOS) compared to the baseline TTS model, indicating that improvements in speaker-preserving CLVC also translate to perceptually better speech quality.

4.4. Trade-off Between Identity and Intelligibility

The seven-system comparison highlights a clear trade-off between linguistic intelligibility and speaker identity preservation in multilingual CLVC. Among multilingual systems, no model consistently dominates both dimensions across all evaluated languages. Qwen3-TTS achieves the strongest transcription accuracy but comparatively weaker speaker similarity, whereas VoxCPM2 demonstrates superior identity preservation with moderate error rates. CosyVoice3 provides the most balanced multilingual performance across both metrics. Among Arabic-capable systems, VoxCPM2 uniquely achieves strong intelligibility and speaker preservation simultaneously, suggesting that language-targeted multilingual training can partially reduce this trade-off in low-resource multilingual speech synthesis settings.

4.5. Discussion

Our findings reveal a consistent trade-off between linguistic accuracy and speaker identity preservation in multilingual CLVC. Autoregressive models like Qwen3-TTS show strong intelligibility but weaker speaker similarity, while diffusion-based models like VoxCPM2 better preserve identity at some cost to transcription accuracy. CosyVoice3 strikes a balance. Model scale alone, as shown by MOSS-TTS, does not guarantee better performance; data diversity and architecture matter more.

5. Conclusion

In this paper, we presented a systematic evaluation of zero-shot cross-lingual voice cloning across Arabic and five additional target languages using recent autoregressive, diffusion-based, and flow-matching TTS architectures. Our

experiments reveal a consistent trade-off between linguistic intelligibility and speaker identity preservation in multilingual speech synthesis. Autoregressive systems such as Qwen3-TTS achieve stronger transcription accuracy, whereas diffusion- and flow-based systems like VoxCPM2 and CosyVoice3 better preserve speaker identity during cross-lingual generation. We further observe that a larger model scale alone does not ensure stronger multilingual generalization or robustness. In addition, we release a paired English-source multilingual same-speaker bilingual corpus and a fully reproducible evaluation pipeline to support future research in multilingual speech synthesis, bilingual ASR, and low-resource speech technologies.

Acknowledgment

The authors sincerely thank the COIL-D (Centre of Indian Language Data) Project under Bhashini, funded by the Ministry of Electronics and Information Technology (MeitY), Government of India, for providing the computational resources and infrastructure that enabled this research

References

Abebe, A. G. and Moslem, Y. One voice, many tongues: Cross-lingual voice cloning for scientific speech, 2026. URL https://arxiv.org/abs/2604.26136.

Basher, M. J. I., Kowsher, M., Islam, M. S., Nandi, R. N., Prottasha, N. J., Menon, M. H., Al Muntasir, T., Chowdhury, S. A., Alam, F., Yousefi, N., et al. Bntts: Few-shot speaker adaptation in low-resource setting. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4956–4968, 2025.

Chen, H., Chen, R., and Hirschberg, J. Emoknob: Enhance voice cloning with fine-grained emotion control. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8170–8180, 2024.

Coqui AI. Xtts-v2: Multilingual zero-shot voice cloning. https://huggingface.co/coqui/XTTS-v2, 2024. Accessed: 2026-05-23.

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

Gong, Y., Jiang, B., Zhao, Y., Yuan, Y., Chen, K., Jiang, Y., Chang, C., Hong, D., Chen, M., Li, R., et al. Moss-tts technical report. *arXiv preprint arXiv:2603.18090*, 2026.

Hu, H., Zhu, X., He, T., Guo, D., Zhang, B., Wang, X., Guo, Z., Jiang, Z., Hao, H., Guo, Z., Zhang, X., Zhang, P., Yang, B., Xu, J., Zhou, J., and Lin, J. Qwen3-tts technical report. *arXiv preprint arXiv:2601.15621*, 2026.

Ji, S., Jiang, Z., Wang, H., Zuo, J., and Zhao, Z. 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)*, pp. 13588–13600, 2024.

Ji, S., Chen, Q., Wang, W., Zuo, J., Fang, M., Jiang, Z., Huang, H., Wang, Z., Cheng, X., Zheng, S., et al. 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)*, pp. 6966–6981, 2025.

Li, Y., Guo, J., Zhang, M., Miaomiao, M., Rao, Z., Zhang, W., He, X., Wei, D., and Yang, H. Pause-aware automatic dubbing using LLM and voice cloning. In Salesky, E., Federico, M., and Carpuat, M. (eds.), *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pp. 12–16, Bangkok, Thailand (in-person and online), August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.iwslt-1.2. URL https://aclanthology.org/2024.iwslt-1.2/.

Lux, F., Koch, J., and Vu, N. T. 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)*, pp. 741–751, 2022.

Lyu, X., Wang, Y., Zhao, T., Wang, H., Liu, H., and Du, Z. Build llm-based zero-shot streaming tts system with cosyvoice. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–2. IEEE, 2025.

Panda, S., Narendra, A., Dubey, K. P., Nadeem, M., et al. 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)*, pp. 584–594, 2026.

Resemble AI. Chatterbox multilingual tts. https://arxiv.org/pdf/2401.13307, 2024.

Shah, N., Kosgi, S., Tambrahalli, V., Nelakanti, A., Gandhi, V., et al. Parrottts: Text-to-speech synthesis exploiting disentangled self-supervised representations. In *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 79–91, 2024.

Singh, K. B., Kumar, D., and Ekbal, A. Evaluation of LLM for English to Hindi legal domain machine translation systems. In Haddow, B., Kocmi, T., Koehn, P., and

Monz, C. (eds.), *Proceedings of the Tenth Conference on Machine Translation*, pp. 823–833, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-341-8. doi: 10.18653/v1/2025.wmt-1.57. URL https://aclanthology.org/2025.wmt-1.57/.

Wang, C. et al. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*, 2023.

Xie, T., Rong, Y., Zhang, P., Wang, W., and Liu, L. 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*, pp. 764–791, 2025.

Zheng, Z., Peng, P., Diwan, A., Huynh, C. P., Sun, X., Liu, Z., Bhat, V., and Harwath, D. Voicecraft-x: Unifying multilingual, voice-cloning speech synthesis and speech editing. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2737–2756, 2025.

Zhou, Y., Zeng, G., Liu, X., Li, X., Yu, R., Wang, Z., Ye, R., Sun, W., Gui, J., Li, K., Wu, Z., and Liu, Z. Voxcpm: Tokenizer-free tts for context-aware speech generation and true-to-life voice cloning. *arXiv preprint arXiv:2509.24650*, 2025.

Zhu, H., Ye, L., Kang, W., Yao, Z., Guo, L., Kuang, F., Han, Z., Zhuang, W., Lin, L., and Povey, D. Omnivoice: Towards omnilingual zero-shot text-to-speech with diffusion language models. *arXiv preprint arXiv:2604.00688*, 2026.

A. Detailed Model Descriptions

VoxCPM2 VoxCPM2 is a tokenizer-free diffusion autoregressive TTS architecture trained on large-scale multilingual speech corpora for zero-shot voice cloning and multilingual speech generation (Zhou et al., 2025).

MOSS-TTS MOSS-TTS is a large-scale autoregressive multimodal speech generation model that supports multilingual and code-switched synthesis through audio token conditioning and phoneme-aware generation mechanisms (Gong et al., 2026).

Qwen3-TTS Qwen3-TTS is a multilingual autoregressive TTS system designed for low-latency speech synthesis using multi-codebook acoustic tokenization and natural-language-driven control of prosody and speaking style (Hu et al., 2026).

CosyVoice3 CosyVoice3 is a lightweight flow-matching multilingual TTS model designed for efficient zero-shot synthesis with improved speaker similarity, prosodic consistency, and multilingual generalization (Lyu et al., 2025).

Chatterbox Chatterbox is an English-source multilingual TTS system evaluated as an Arabic-capable baseline for cross-lingual speech generation and multilingual speaker transfer tasks (Resemble AI, 2024).

XTTS-V2 XTTS-V2 is a multilingual zero-shot voice cloning framework that enables cross-lingual speech synthesis from short reference audio samples while preserving speaker identity (Coqui AI, 2024).

OmniVoice OmniVoice is an omnilingual TTS architecture designed for zero-shot multilingual speech synthesis, with strong cross-lingual generalization and multilingual speaker adaptation capabilities (Zhu et al., 2026).

A.1. Additional Overall System Comparison

Figure 3 presents the overall multilingual CLVC performance across all evaluated systems using mean transcription error rates and ECAPA-TDNN speaker similarity scores. The results highlight a consistent trade-off between linguistic intelligibility and speaker identity preservation across multilingual speech synthesis models. Qwen3-TTS achieves the lowest average WER/CER, demonstrating stronger multilingual transcription accuracy and linguistic consistency across supported languages, but it produces comparatively lower speaker similarity scores. In contrast, VoxCPM2 achieves the highest speaker similarity, indicating superior preservation of speaker identity during cross-lingual synthesis. CosyVoice3 provides the most balanced multilingual performance across both metrics, whereas MOSS-TTS exhibits comparatively weaker multilingual robustness despite its larger parameter scale and multilingual training setup.

A.2. Architectural Interpretation of the Trade-off

The observed trade-off between linguistic intelligibility and speaker identity preservation is influenced by the conditioning mechanisms of CLVC architectures. Since cross-lingual voice cloning must model both speaker characteristics and target-language phonetics, improving one objective often impacts the other. Autoregressive models prioritize linguistic decoding, achieving stronger phonetic adaptation and lower transcription errors, but may weaken speaker-specific information during generation. In contrast, diffusion and flow-based models repeatedly reinforce speaker embeddings, better preserving vocal traits such as timbre and prosody. However, this stronger emphasis on speaker identity can limit phonetic adaptation to the target language, resulting in slightly higher error rates.

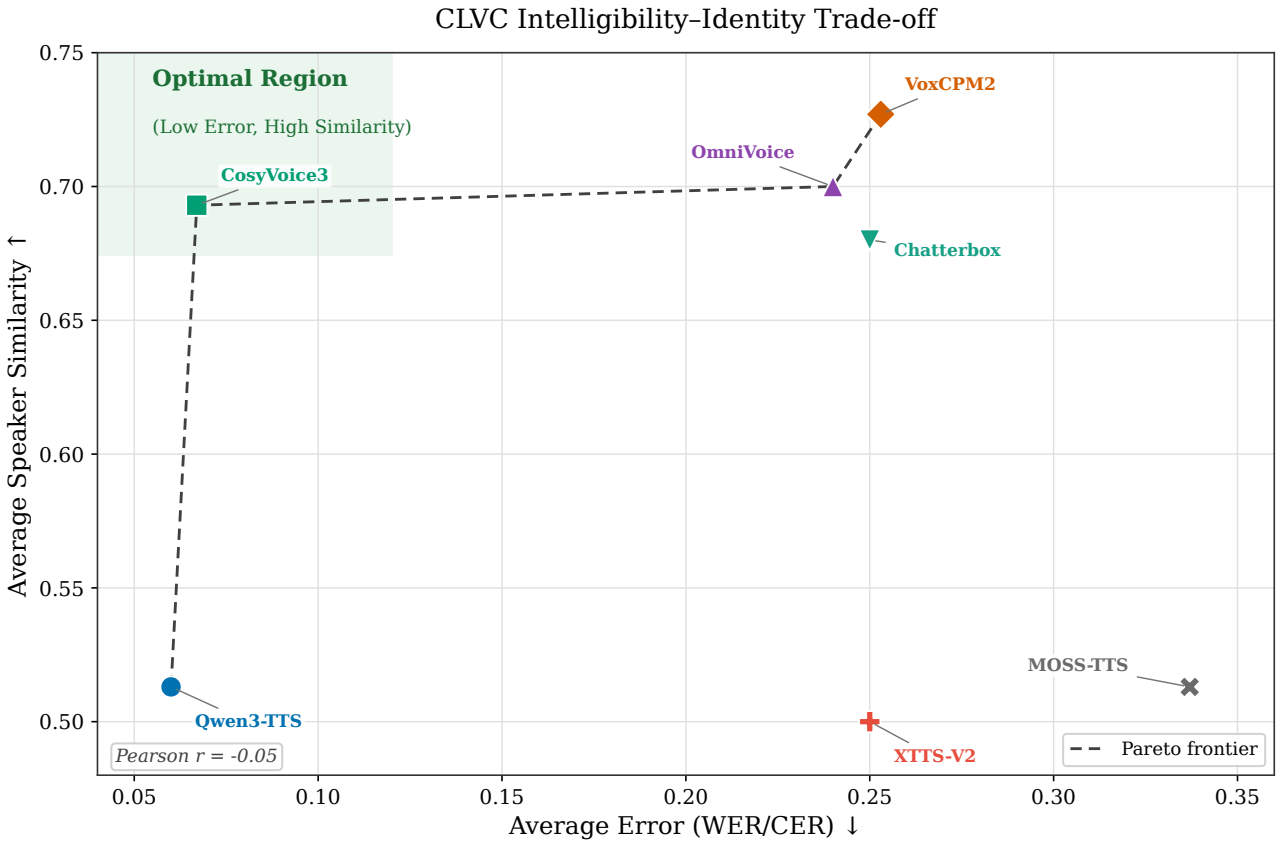

Figure 2. Quantitative trade-off between linguistic intelligibility (average WER/CER, lower is better) and speaker identity preservation (average speaker similarity, higher is better) across evaluated CLVC systems. Values are averaged over all available target languages for each model. The Pareto frontier identifies non-dominated systems, showing that no single model can simultaneously optimize both objectives.

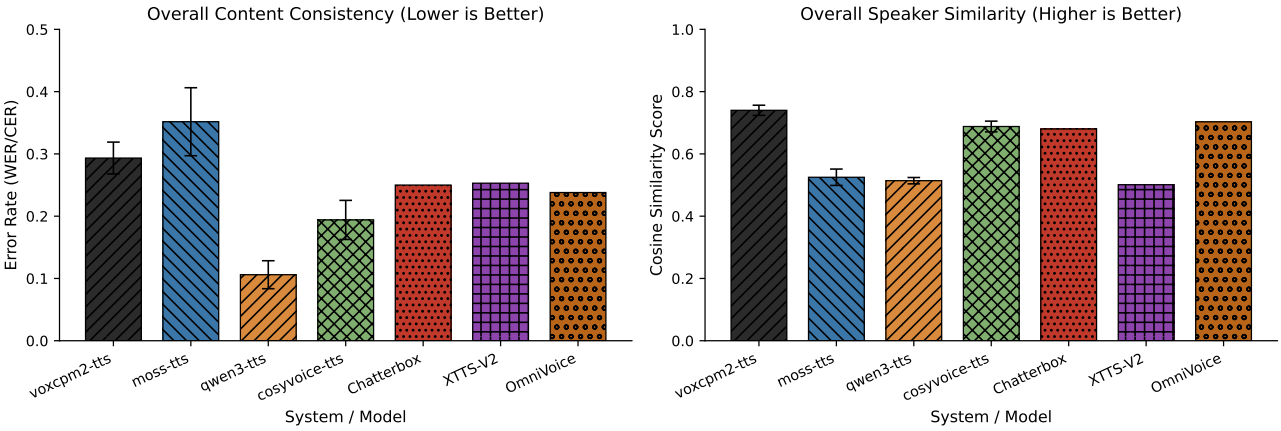

Figure 3. Overall multilingual CLVC performance across evaluated systems. Top: mean content consistency measured using WER/CER, where lower values indicate better linguistic accuracy. Bottom: mean speaker similarity measured using ECAPA-TDNN cosine similarity, where higher values indicate stronger speaker identity preservation across synthesized multilingual speech.