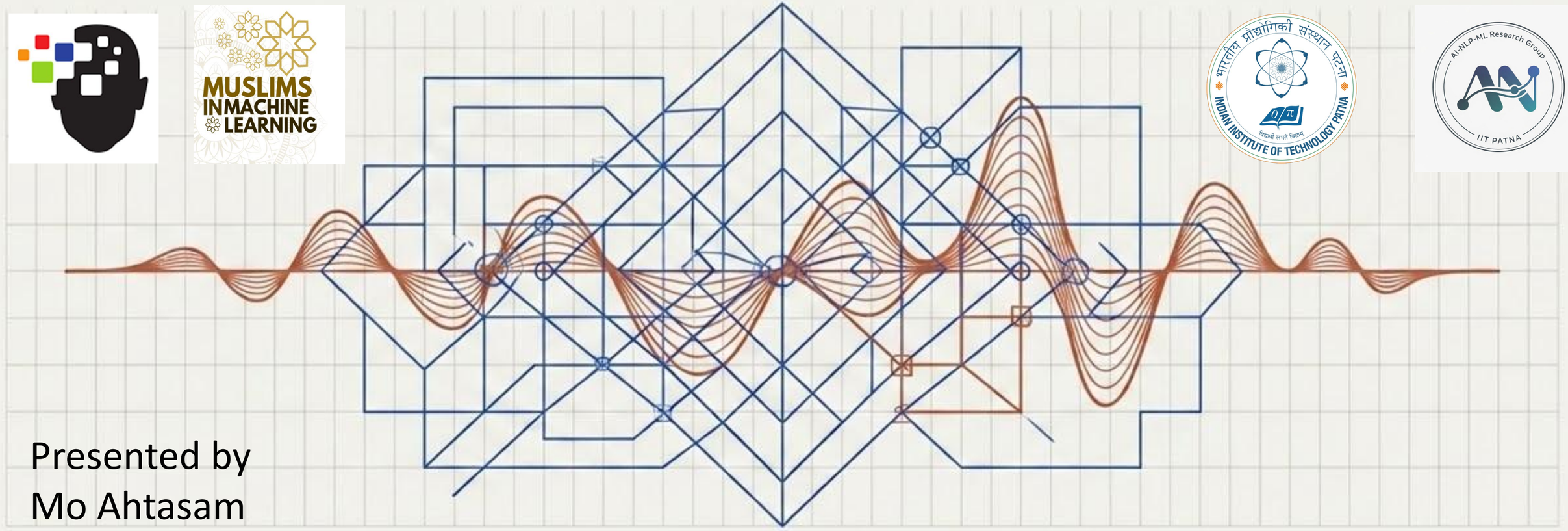
Toward Speaker-Preserving Arabic Speech: A Bilingual Resource and Comparative Evaluation of Cross-Lingual Voice Cloning



Presented by
Mo Ahtasam

Benchmark Analysis & Bilingual Dataset Release (ACL-60/60)

2 Billion

Used in religious practice by Muslims worldwide.

22

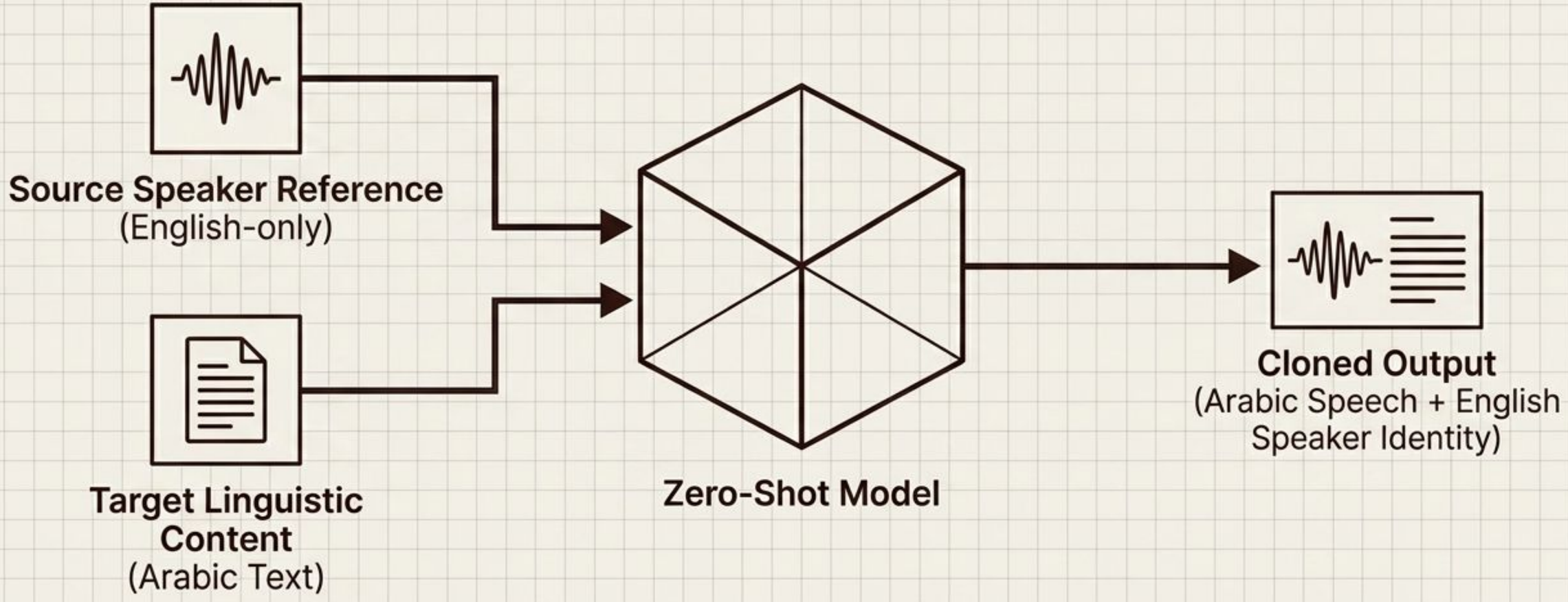
Countries where Arabic is an official or co-official language.

Low-Resource

The paradoxical status of Arabic in modern speech tasks requiring simultaneous control over linguistic content and speaker identity.

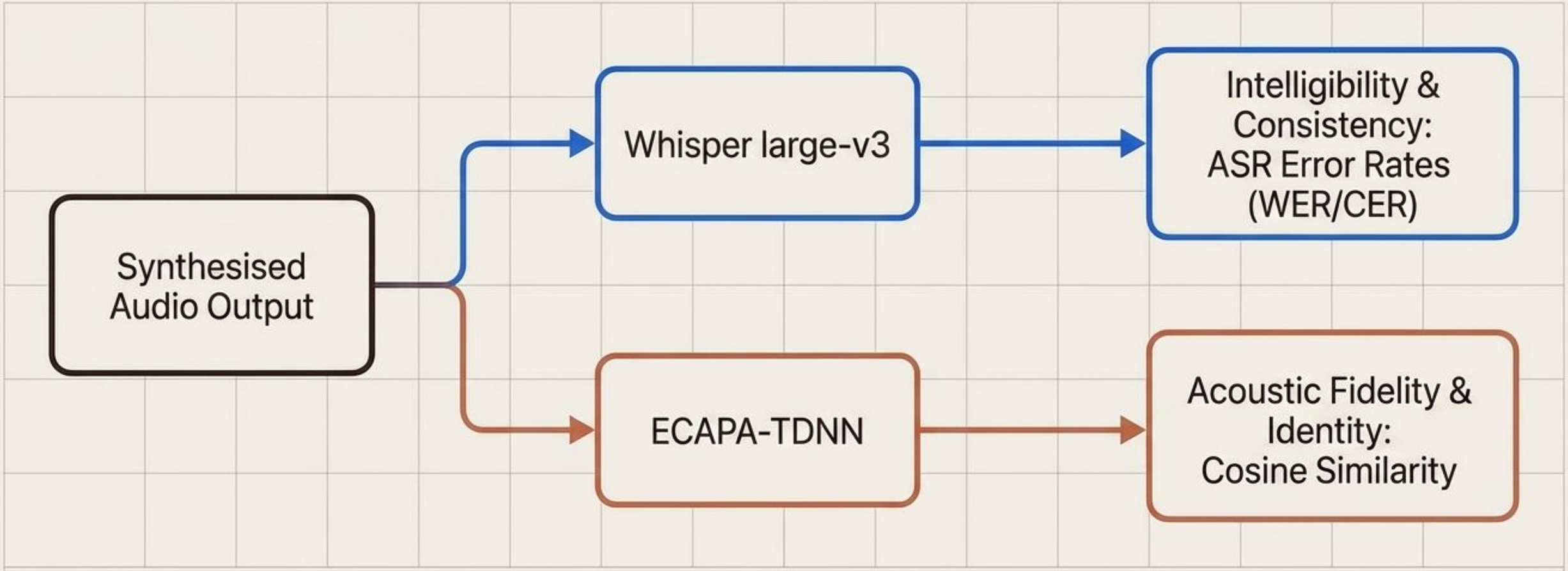
Key Takeaway: Despite rapid progress in high-resource multilingual speech generation, robust cross-lingual voice cloning for Arabic remains severely bottlenecked by a scarcity of high-quality, same-speaker bilingual resources.

The Zero-Shot Mechanism



Strict Zero-Shot Generalisation: The model must clone the voice using only an English reference. The speaker has never recorded audio in the target language.

The Evaluation Pipeline

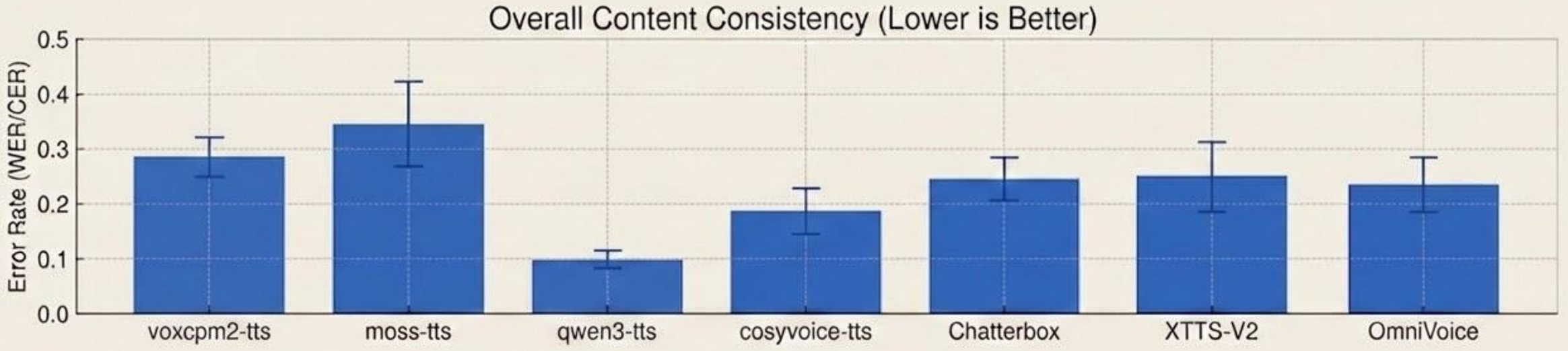


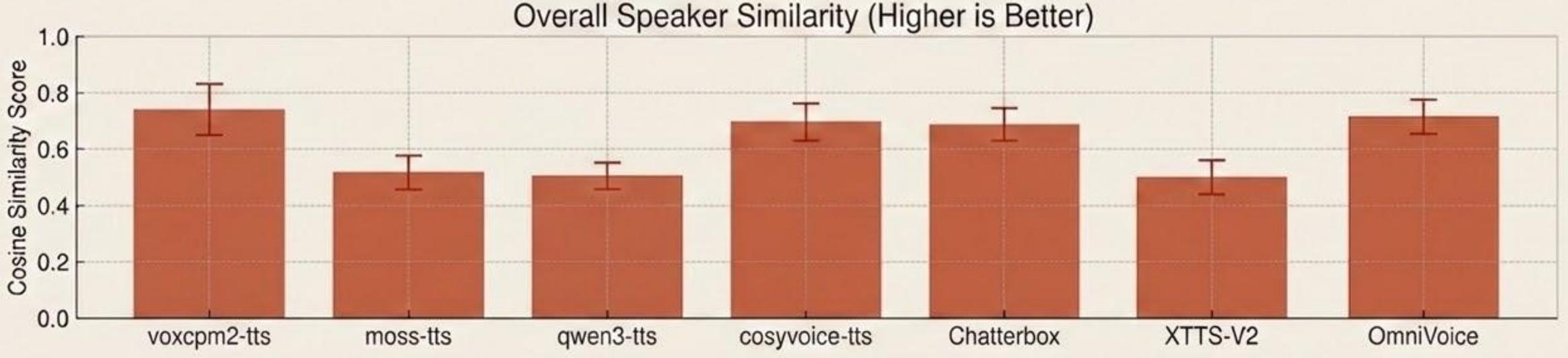
Insight: Assessing cross-lingual voice cloning requires subprocess isolation. We measure how accurately the model speaks the new language versus how flawlessly it preserves the original vocal identity.

The CLVC Contenders Matrix

Model Name	Architecture Paradigm	Language Scope	Defining Trait
Qwen3-TTS	Autoregressive	Multilingual	Multi-codebook tokenisation
MOSS-TTS	Autoregressive (8B)	Multilingual	Multimodal conditioning
VoxCPM2	Diffusion Autoregressive	Multilingual	Waveform conditioning
CosyVoice3	Flow-Matching	Multilingual	Efficient zero-shot flow
OmniVoice	Diffusion	Arabic-Capable	Omnilingual generation
Chatterbox	Undisclosed	Arabic-Capable	English-source transfer
XTTS-V2	Undisclosed	Arabic-Capable	Short-reference transfer

Note: All models evaluated using 12 unseen English lish speakers from the ACL-60/60 development split split to ensure identical input conditions.

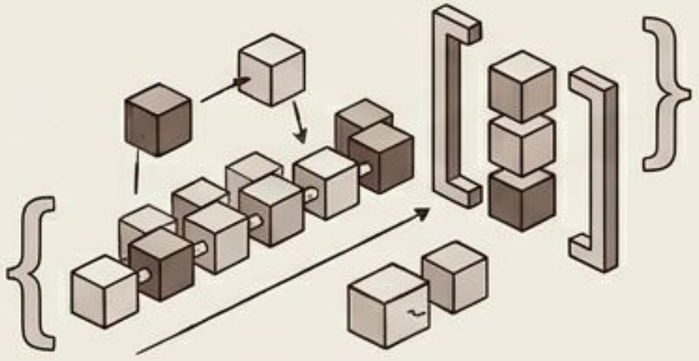




Insight: While models like qwen3-tts and voxcpm2-tts exhibit a strict trade-off between transcription accuracy and vocal identity, the data shows this is not an absolute rule; models like cosyvoice-tts prove that achieving a strong balance across both dimensions is possible.

The Architectural Divide

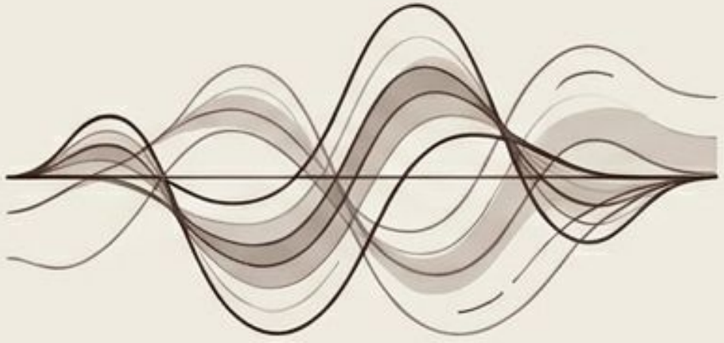
Autoregressive Focus



Highlighted Model: `Qwen3–TTS`

- **Mechanism:** Token-based generation strictly enforces linguistic structure.
- **Outcome:** Text-Dominant (High Intelligibility, Degraded Similarity).

Diffusion / Flow Focus

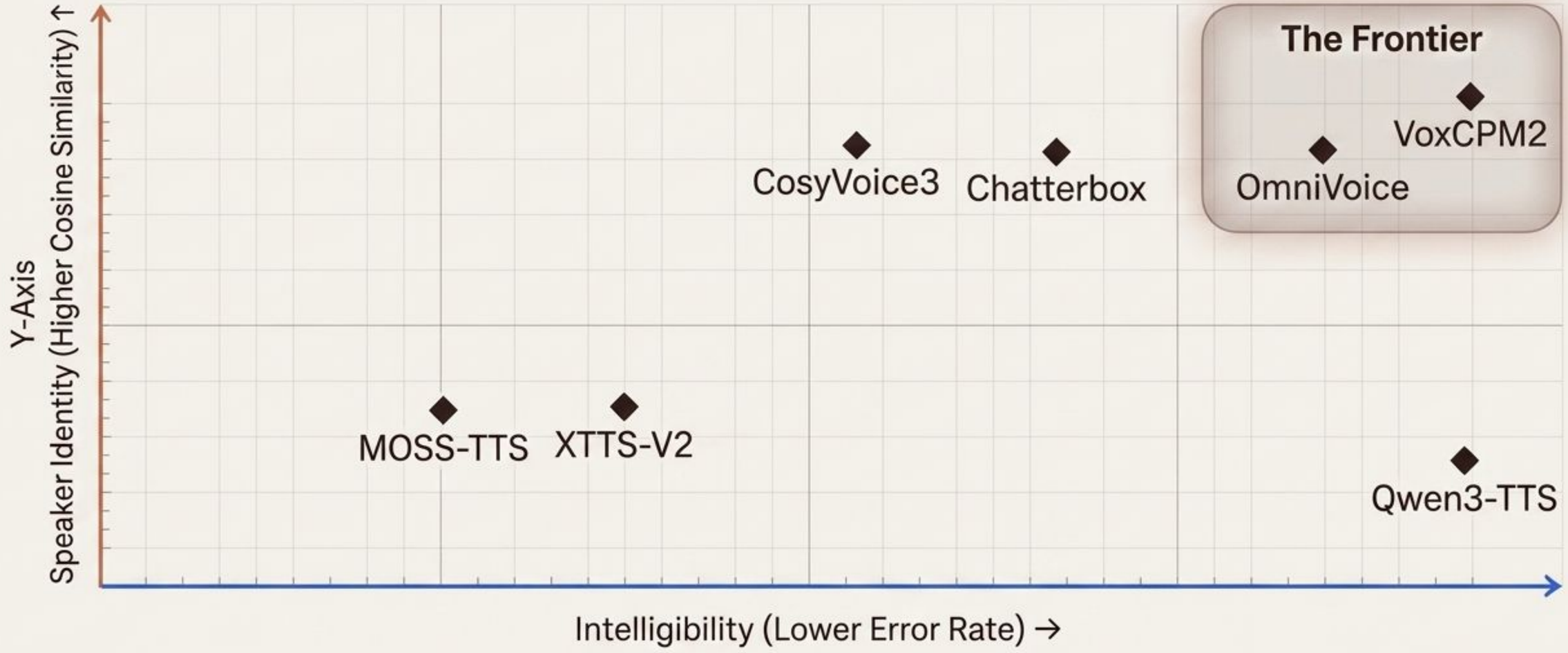


Highlighted Models: `VoxCPM2, CosyVoice3`

- **Mechanism:** Waveform and flow-matching generation prioritise continuous acoustic fidelity.
- **Outcome:** Acoustic-Dominant (High Identity, Reduced Intelligibility).

Synthesis: Architecture is destiny. Tokenisation intrinsically favours linguistic accuracy, while continuous waveform generation inherently protects vocal timbre. Scale alone (e.g., the 8B parameter MOSS-TTS) does not break this divide.

The Trade-off Quadrant



Key Takeaway: OmniVoice and VoxCPM2 approach the ideal Top Right quadrant for Arabic, suggesting that targeted multilingual training can bend, if not entirely break, the trade-off curve.

Releasing the English-Arabic Bilingual ASR Dataset

Dataset Specifications

- **Source:** ACL-60/60 English dev split
- **Architecture:** Paired English reference + Arabic cloned audio (via VoxCPM2)
- **Volume:** 12 strict zero-shot speakers, 49 utterances per speaker (588 pairs)
- **Annotation:** Ground-truth Arabic transcripts + Whisper large-v3 English transcripts

Purpose: The first fully reproducible evaluation pipeline for simultaneous ASR-grounded assessment across both source and target languages.

The Structural Trade-off & the Arabic Frontier

The Trade-off is Structural

Choose autoregressive models for transcription precision; choose diffusion/flow-matching for vocal preservation.

Scale ≠ Scale ≠ Generalisation

Massive parameter counts do not guarantee low-resource robustness (evidenced by MOSS-TTS).

The Arabic Frontier

VoxCPM2 and OmniVoice currently lead, but breaking the intelligibility-identity compromise requires paired bilingual corpora.

github.com/tb-fa-netizen/CLVO | huggingface.co/datasets/md-ahtisham/CLVO_en_ar

THANK YOU

For your time and dedication to advancing acoustic models.

ACCESS REPOSITORY

GET BILINGUAL DATASET

CONNECT ON LINKEDIN

READ THE PAPER

RESOURCES