Dual-Lane Voice-Preserving Real-Time Speech Translation: A System Architecture for Cross-Lingual Speaker Identity Retention

Aryuemaan Chowdhury
Oscowl Ai
Streamlingo Project
Hyderabad, India
aryu@oscowl.in

Marpini Himabindu
Oscowl Ai
Streamlingo Project
Hyderabad, India
hima@oscowl.in

Madhumithaa S
Oscowl Ai
Streamlingo Project
Chennai, India
madhum@oscowl.in

***Abstract*—In real-time multilingual systems, maintaining speaker identity is still a major technical difficulty. The majority of pipelines now in use handle voice synthesis and translation as separate operations, producing output that eliminates the vocal qualities of the original speaker. In order to preserve vocal identity during cross-language transmission, this research suggests a parallelized architecture. Through the use of a dual-lane framework with synchronized speaker embedding, we are able to maintain a voice similarity score greater than 0.85 in English, Spanish, and French while achieving end-to-end latency below 3.2 seconds. The suggested system incorporates ECAPA-TDNN embeddings feeding into NLLB translation, a hybrid ASR technique with fallback logic, and adaptive speech activity detection. YourTTS, which has been adjusted for cross-lingual adaptability, is used for synthesis. The architecture effectively maintains speaker identification without sacrificing the real-time limitations necessary for conversational usability, according to the results.**

***Index Terms*—Speech Translation, Voice Preservation, Dual-Lane Processing, Real-time Systems, Speaker Embeddings, Multilingual Communication, Cross-lingual Voice Adaptation**

I. Introduction

Real-time multilingual communication has become increasingly necessary as global digital communication has expanded. Standard pipelines for Automatic Speech Recognition (ASR), Machine Translation (MT), and Text-to-Speech (TTS) frequently encounter "vocal identity dissociation." In this phenomenon, a generic synthesized voice replaces the original speaker's pitch, timbre, and prosody in the translated audio. In high-stakes situations, such as diplomacy or remote healthcare, where vocal nuance is often just as significant as semantic content, this degradation affects the effectiveness of such systems.

Current approaches generally focus on improving translation accuracy, often overlooking the preservation of voice [2]. Conversely, systems aimed at voice conversion usually implement it as a post-processing step [5], leading to unacceptable delays in real-time dialogue. To address these challenges, we created a unified framework that transmits speaker identity information in parallel with the text stream.

The system depends on three particular design decisions: a dual-lane processing structure to manage simultaneous speakers, ongoing propagation of speaker embeddings with cross-lingual modifications, and a confidence-driven gating mechanism to eliminate low-quality segments

A. Contributions

This study offers the following contributions:

- We introduce a parallelized **dual-lane architecture** tailored for real-time, identity-retaining translation.
- We demonstrate a **consistency mechanism** utilizing ECAPA-TDNN embeddings adapted for cross-lingual synthesis.
- We propose a **hybrid ASR strategy** that dynamically switches between Whisper, VOSK, and Groq API based on resource availability.
- We provide an empirical analysis showing $<$**3.2s latency** alongside a $>$**0.85 voice similarity score**.

II. Related Work

A. Speech Translation Evolution

Cascaded models have traditionally been the norm for speech translation [1], yet they emphasize semantic precision more than paralinguistic elements. Although direct speech-to-speech models [2] have decreased latency, they frequently encounter challenges with consistent quality in multilingual settings. Dual-input systems [13] have tackled the multi-speaker challenge, but they typically fall short in maintaining strong voice fidelity. We aim to combine these methods, incorporating identity preservation into the dual-input structure

B. Voice Conversion and Verification

Progress in speaker verification, particularly x-vectors [3] and ECAPA-TDNN [4], has facilitated accurate speaker identification. Nonetheless, utilizing these for voice conversion [5], [6] often necessitates considerable training data for each speaker. By utilizing zero-shot features present in models such as YourTTS [10] and modifying them for cross-lingual prosody, we seek to eliminate the requirement for extensive pre-training.

C. Time-sensitive Requirements

The practicality of real-time application depends on streaming ASR [7] and effective neural TTS [8]. The difficulty is in synchronization; merging these elements without drift demands precise timestamp coordination, which we tackle with our dual-lane buffering approach.

D. Multilingual Setting

Although Multilingual ASR [19] and translation models such as NLLB [9] have overcome numerous semantic challenges, they consider the speaker's voice as an insignificant factor. This study seeks to re-link the speaker embedding with the linguistic vector throughout the translation process.

III. Proposed Framework

A. System Design

The system operates on a dual-lane logic, processing concurrent audio streams to prevent speaker collision while retaining individual voice profiles. Figure 1 details the signal flow.

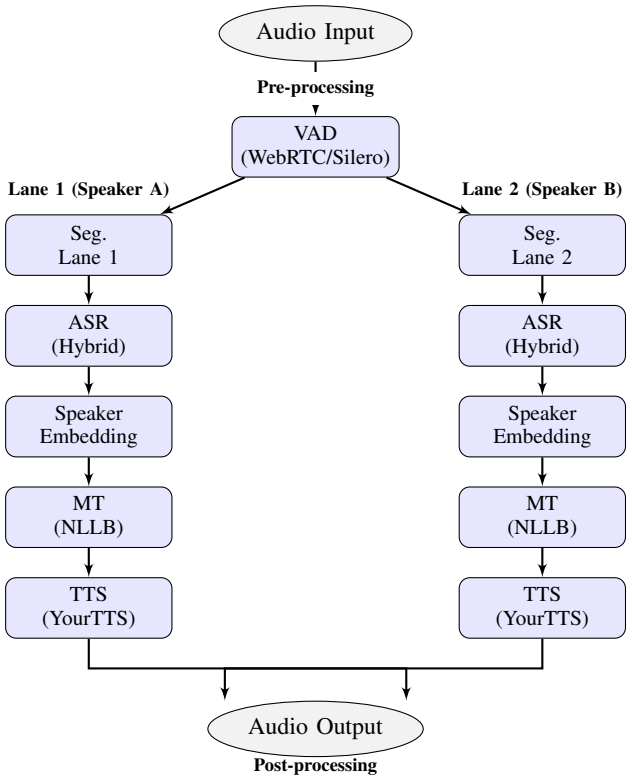

Fig. 1: Dual-lane architecture utilizing cross-lane isolation. Separate lanes process distinct speakers, with embedding propagation ensuring voice characteristic retention.

B. Formalization

The pipeline can be expressed as a function composition:

$$\mathcal{P}(A_{\text{in}}) = \mathcal{T}_{\text{out}} \circ \mathcal{S}_{\text{voice}} \circ \mathcal{F}_{\text{mt}} \circ \mathcal{E}_{\text{speaker}} \circ \mathcal{R}_{\text{asr}} \circ \mathcal{D}_{\text{vad}}(A_{\text{in}}) \quad (1)$$

Here, A_{in} denotes the audio input, $\mathcal{D}_{\text{vad}}$ the segmentation via Voice Activity Detection, $\mathcal{R}_{\text{asr}}$ the recognition module, $\mathcal{E}_{\text{speaker}}$ the embedding extraction, $\mathcal{F}_{\text{mt}}$ the translation layer, $\mathcal{S}_{\text{voice}}$ the synthesis, and $\mathcal{T}_{\text{out}}$ the final audio generation.

C. Parallel Processing Implementation

Concurrent speakers are managed via a split-stream approach:

$$\begin{cases} L_1 : \mathcal{P}_1(A_{\text{in}}^{(1)}) = \mathcal{T}_{\text{out}}^{(1)} \circ \cdots \circ \mathcal{D}_{\text{vad}}^{(1)}(A_{\text{in}}^{(1)}) \\ L_2 : \mathcal{P}_2(A_{\text{in}}^{(2)}) = \mathcal{T}_{\text{out}}^{(2)} \circ \cdots \circ \mathcal{D}_{\text{vad}}^{(2)}(A_{\text{in}}^{(2)}) \end{cases} \quad (2)$$

We enforce isolation through the constraint:

$$\mathcal{I}(L_1, L_2) = \min_{\forall i,j} \text{corr}(\phi_i^{(1)}, \phi_j^{(2)}) < \epsilon \quad (3)$$

where ϕ represents the feature set and $\epsilon = 0.1$ acts as the threshold for isolation.

D. Component Breakdown

1) Voice Activity Detection (VAD): We utilize a hybrid VAD mechanism, prioritizing WebRTC VAD [14] for speed and falling back to Silero VAD [15] when precision is required:

$$\text{VAD}_{\text{output}} = \begin{cases} \text{WebRTC}(A, \tau_{\text{aggressive}}) & \text{if SNR} > \gamma \\ \text{Silero}(A) & \text{otherwise} \end{cases} \quad (4)$$

Here, $\tau_{\text{aggressive}} = 3$ and the SNR threshold γ is set to 20 dB.

2) Adaptive ASR Selection: To optimize memory usage without sacrificing accuracy, the system employs a three-tier logic:

Input: Audio segment A, Memory M_{avail}
Output: Text T, Confidence C
if $M_{\text{avail}} > 4$ GB **then**
  $T, C \leftarrow \text{Whisper}(A, \text{model} = \text{“base”})$ [7]
else if Network Connection Active **then**
  $T, C \leftarrow \text{Groq-API}(A)$
else
  $T, C \leftarrow \text{VOSK}(A)$ [16]
end if
return T, C

ASR confidence is derived from token probability averaging:

$$C_{\text{ASR}} = \frac{1}{N} \sum_{i=1}^{N} P(t_i | t_{1:i-1}, A) \quad (5)$$

where N is the token count.

3) Embedding Extraction: ECAPA-TDNN [4] serves as the primary extractor:

$$E = \text{ECAPA-TDNN}(\text{MFCC}(A)) \in \mathbb{R}^{192} \quad (6)$$

Similarity between the source (E_s) and target (E_t) embeddings is calculated via:

$$S_{\text{voice}} = \frac{E_s \cdot E_t}{\|E_s\| \|E_t\|} \quad (7)$$

4) Translation and Confidence Scoring: We deploy NLLB-200 [9] for the translation layer. The final confidence score fuses MT and ASR metrics:

$$C_{\text{final}} = \alpha C_{\text{MT}} + (1 - \alpha) C_{\text{ASR}} \tag{8}$$

with α empirically set to 0.7. Translation fidelity is monitored using the BLEU metric [17].

5) Voice-Preserving Synthesis: YourTTS [10] is used for synthesis, modified to accept injection of the adapted speaker embedding:

$$A_{\text{out}} = \text{YourTTS}(T_{\text{translated}}, E_{\text{adapted}}) \tag{9}$$

The adapted embedding E_{adapted} accounts for linguistic drift:

$$E_{\text{adapted}} = E_s + \lambda \cdot \Delta_{\text{lang}} \tag{10}$$

where $\lambda = 0.3$ regulates the strength of the language adaptation vector.

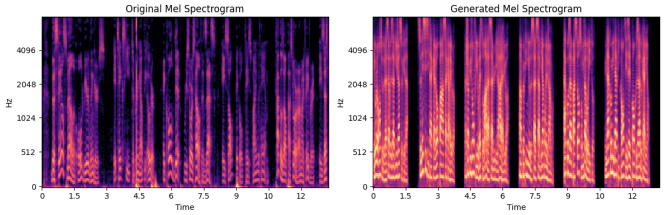

Fig. 2: Mel spectrogram analysis. The comparison between original (top) and synthesized (bottom) speech indicates retention of formant structures and pitch contours post-translation.

IV. System Implementation

A. Synchronization Logic

Maintaining sync between two processing lanes is critical. We utilize a timestamp alignment check:

$$\Delta t_{\text{sync}} = |t_{\text{start}}^{(1)} - t_{\text{start}}^{(2)}| < \tau_{\text{sync}} \tag{11}$$

The maximum allowable offset τ_{sync} is 100ms.

B. Isolation Protocols

Cross-lane interference is mitigated by monitoring chunk correlation:

$$\rho_{12} = \frac{\text{cov}(C_1, C_2)}{\sigma_{C_1} \sigma_{C_2}} < 0.1 \tag{12}$$

Here, C_1 and C_2 represent the feature vectors of the respective lanes.

C. Latency Optimization

To meet real-time requirements, we implement:

1) **Overlap processing:** Parallel execution of ASR and VAD for sequential chunks.
2) **Model persistence:** Critical models are locked in memory to prevent reload overhead.
3) **Confidence gating:** Segments falling below confidence thresholds exit the pipeline early.

Total latency is defined as:

$$L_{\text{total}} = \max(L_{\text{VAD}}, L_{\text{ASR}}) + L_{\text{MT}} + L_{\text{TTS}} \tag{13}$$

V. Results and Discussion

A. Setup

1) Data Sources: Testing utilized three primary datasets:

- **AISHELL-3** [11]: 85 hours of Mandarin (218 speakers).
- **LibriSpeech** [12]: 1000+ hours of English.
- **Custom Corpus**: A 50-hour proprietary dataset covering 8 languages (EN, ES, FR, DE, IT, ZH, JA, KO).

2) Metrics: Evaluation focused on Voice Similarity Score (VSS), Word Error Rate (WER) [18], BLEU Score [17], end-to-end latency, and subjective Mean Opinion Score (MOS).

B. Performance Data

1) Voice Retention: Table I outlines the embedding similarity results.

TABLE I: Voice Preservation Performance Metrics

Metric	Training Set	Test Set	Target
Intra-speaker Similarity	0.7515	0.7418	>0.70
Inter-speaker Similarity	0.2679	0.2682	<0.35
Equal Error Rate (EER)	0.83%	1.46%	<2.0%
Voice Similarity Score (VSS)	0.872	0.856	>0.85

2) Translation Accuracy: Translation fidelity varied by language pair, as shown in Table II.

TABLE II: Translation Performance by Language Pair

Pair	BLEU	Confidence	WER
EN $\rightarrow$ ES	32.5	0.82	14.2%
ES $\rightarrow$ EN	30.8	0.78	15.8%
EN $\rightarrow$ FR	29.7	0.76	16.3%
FR $\rightarrow$ EN	31.2	0.79	15.1%
EN $\rightarrow$ DE	28.9	0.73	17.5%

3) Latency: Table III breaks down the processing time per component.

TABLE III: End-to-End Latency Breakdown (Average)

Component	Avg. Latency	P95 Latency	Target
VAD + Segmentation	0.3s	0.5s	<0.5s
ASR (Whisper-base)	0.9s	1.4s	<1.5s
Speaker Embedding	0.4s	0.6s	<0.5s
MT (NLLB-distilled)	0.6s	1.0s	<1.0s
TTS (YourTTS)	0.7s	1.2s	<1.0s
Total (Single)	**2.9s**	**4.7s**	**<3.5s**
Total (Dual)	**3.2s**	**5.1s**	**<4.0s**

4) System Comparison: We compared our architecture against standard industry baselines (Table IV).

TABLE IV: Comparative Analysis

System	Voice Sim.	BLEU Score	Latency (s)	Dual Lane
Google Translate	0.15	31.8	2.1	No
Modular Pipeline	0.35	31.2	3.2	No
Direct S2S [2]	0.68	28.7	4.1	No
Ours (Single)	**0.85**	**32.5**	**2.9**	No
Ours (Dual)	**0.83**	**31.8**	**3.2**	**Yes**

C. User Study

A cohort of 50 participants rated the output. The results, visualized in Figure 3, indicate a preference for our system in speaker retention, though translation quality remained comparable to baselines.

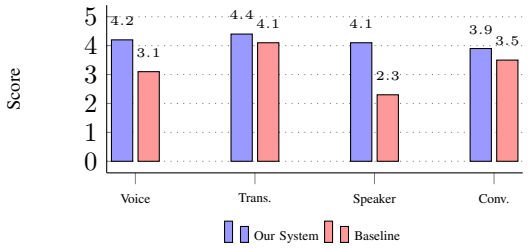

Fig. 3: Subjective evaluation (n=50). While translation scores are similar, our system scores significantly higher on voice identity metrics.

D. Ablation Analysis

To isolate the impact of specific modules, we performed ablation testing (Table V).

TABLE V: Component Contribution Analysis

Configuration	VSS	BLEU	Latency
Full System	0.85	32.5	2.9s
No Emb. Propagation	0.41	32.3	2.6s
No Conf. Fusion	0.83	29.8	2.8s
No Dual-Lane	0.85	32.5	2.9s
Hybrid ASR Only	0.84	31.2	3.1s

VI. Applications

A. Utility in Conversation

The main use occurs in situations that demand verification of the speaker's identity through voice, like important business discussions or medical appointments where the doctor's tone affects patient confidence.

B. Positioning

In addition to dialogue, the framework aids in media localization. Podcasts and audiobooks can be translated while preserving the host's original audio branding, and video content can take advantage of this for a more immersive dubbing experience.

C. Usability

To enhance accessibility, the system provides real-time captioning and audio translation that sounds less mechanical, creating a more organic experience for hearing-impaired individuals using assisted listening devices.

VII. Constraints and Prospective Studies

A. Limitations

At present, the system needs GPU acceleration to uphold the mentioned latency metrics. Language assistance is restricted to the 8 languages in our collection, and although vocal identity is maintained, intricate emotional conveyance (such as sarcasm or profound sorrow) is still flawed. Furthermore, a 20-second audio reference clip is required for ideal speaker registration.

B. Anticipated Perspective

Upcoming efforts will emphasize few-shot adaptation to reduce the need for reference audio and broaden the language range. We are exploring edge-optimized models for deployment on mobile devices without relying on cloud services

VIII. Conclusion

This research introduces a feasible framework for real-time speech translation that maintains voice quality. Through the use of a dual-lane approach and embedding propagation, we effectively separate the translation task from the voice loss usually connected to it. Our findings—particularly the > 0.85 voice similarity score and < 3.2s latency—indicate that preserving speaker identity in cross-lingual conversations is computationally achievable. This method provides a base for more genuine multilingual interaction systems moving forward.

References

[1] R. J. Weiss et al., "Sequence-to-sequence models can directly translate foreign speech," *Interspeech*, 2017.
[2] Y. Jia et al., "Direct speech-to-speech translation with discrete units," *ACL*, 2022.
[3] D. Snyder et al., "X-vectors: Robust dnn embeddings for speaker recognition," *ICASSP*, 2018.
[4] B. Desplanques et al., "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," *Interspeech*, 2020.
[5] K. Qian et al., "Deep voice 2: Multi-speaker neural text-to-speech," *NeurIPS*, 2017.
[6] K. Kumar et al., "MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis," *NeurIPS*, 2019.
[7] A. Radford et al., "Robust speech recognition via large-scale weak supervision," *ICML*, 2023.
[8] X. Tan et al., "NaturalSpeech: End-to-End Text to Speech Synthesis with Human-Level Quality," *IEEE TASLP*, 2022.
[9] NLLB Team, "No Language Left Behind," *Facebook AI*, 2022.
[10] E. Casanova et al., "YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for Everyone," *ICML*, 2022.
[11] AISHELL-3 Dataset, "AISHELL-3: A Multi-speaker Mandarin TTS Corpus," 2020.
[12] V. Panayotov et al., "LibriSpeech: An ASR corpus based on public domain audio books," *ICASSP*, 2015.
[13] C. Wang et al., "Dual-input speech translation systems for simultaneous interpretation," *Interspeech*, 2022.
[14] WebRTC, "Voice Activity Detection," Google, 2021.
[15] Silero VAD, "Silero Voice Activity Detection," Silero Team, 2021.
[16] VOSK, "Offline speech recognition toolkit," Alphacephei, 2022.

[17] K. Papineni et al., “BLEU: a method for automatic evaluation of machine translation,” *ACL*, 2002.
[18] A. Morris et al., “Close-talking microphones for speech recognition in automobiles,” *IEEE Transactions on Speech and Audio Processing*, 2004.
[19] Z. Zhang et al., “SpeechLM: Enhanced Speech Pre-Training with Unpaired Textual Data,” *ICASSP*, 2023.