ARCHITECTURAL INNOVATIONS AND PERFORMANCE OPTIMIZATION IN REAL-TIME SPEECH-TO-SPEECH TRANSLATION SYSTEMS

Anubhav Maurya, Anuj Shubham Arya, Anil Kumar, Abhimanyu Singh
Department of Computer Science and Engineering, Raj Kumar Goel Institute of Technology, Ghaziabad, India

ABSTRACT

The invention of the real-time voice translation (RTVT) systems is a revolutionary step in computational linguistics and digital signal processing, and it addresses the basic human requirement of a fluent cross-lingual communication. The current research paper gives a comprehensive analysis of the technical structures of modern speech-to-speech translator (S2ST), focusing on the difference between modular cascaded pipelines and integrated end-to-end structures. This paper outlines the processes of processing streaming audio, articulatory features disentangling, and latency budgeting, by exploring the state-of-the-art models of SeamlessM4T, RT-VC and SimulTron. The methodology involves a rigorous review of neural machine translation, automatic speech recognition and text-to-speech synthesis, supported by mathematical formulations of measuring accuracy and speed, including Word Error Rate (WER) and Bilingual Evaluation Understudy (BLEU). Results have shown that, unlike cascaded systems, which are more accurate when a many-to-many language pair is used, direct models are much lower in cumulative latency and do not lose emotional prosody by using a text-based intermediary. Hands-on experimentation showed that, under optimized conditions, voice conversion could be achieved in as little as 61.4 milliseconds — with complete translation pipelines finishing in around 2 seconds. This study has implications on critical areas like healthcare, access by the Deaf and Hard of Hearing (DHH), and international business relationships. Looking ahead, on-device processing and hybrid architecture are expected to shape the 2026 technological landscape, striking a balance between linguistic accuracy and the fluid, natural rhythm of human conversation.

INTRODUCTION

The high degree of globalization that the twenty-first century has brought with it has necessitated the creation of superior communication technologies that will have the capacity to break the language barrier in real time. The basic idea of a

Real time voice translator is to take the input of a source speaker as a spoken input and use sophisticated computational layers, translate the input into a synthesized translation in the target language. The development of streaming models is essential to the applications of international diplomacy to emergency medical services, where a second of delay can make a difference.5 The high degree of globalization that the twenty-first century has brought with it has necessitated the creation of superior communication technologies that will have the capacity to break the language barrier in real time. The basic idea of a real time voice translator is to take the input of a source speaker as a spoken input and use sophisticated computational layers, translate the input into a synthesized translation in the target language. The development of streaming models is essential to the applications of international diplomacy to emergency medical services, where a second of delay can make a difference. At the heart of this research are two competing approaches: cascaded pipelines, which break translation into distinct stages, and E2E models, which handle everything in one go. By mapping source audio directly to target audio tokens, E2E systems sidestep the compounding errors and latency that cascaded architectures often struggle with. Yet deploying these systems in everyday settings is far from straightforward, as real-world noise and speaker variability remain persistent obstacles. On a more promising front, voice personalization has gained serious momentum, with RT-VC and Translatotron[2] leading early efforts in zero-shot voice conversion and tailored speech synthesis.

LITERATURE REVIEW

The history of speech-to-speech translation can be described as the shift towards the lack of symbols and the use of statistics, and eventually the modern day of deep neural networks. The initial attempts that were done in the 1980s and 1990s were based on hand-written rules and small vocabularies, which were inappropriate with the variability of natural speech. The next generation of Statistical Machine Translation (SMT) was more flexible, but had difficulty with the computational requirements of real-time processing.8

The current era of speech translation is defined by Large Language Models and specialized transformer designs. Meta, Google, and OpenAI have each played a significant role in pushing the field forward, releasing foundation models trained across hundreds of thousands of hours of multilingual speech — a scale that directly enables the low-latency performance modern applications demand.

Model / Framework	Primary Contribution	Training Data (Audio Hours)	Key Technological Pillar
Whisper	Massively Multilingual ASR	680,000	Transformer Encoder-Decoder
SeamlessM4T	Unified Multimodal S2ST	1,000,000	w2v-BERT 2.0 / UnitY
RT-VC	Real-Time Voice Conversion	N/A	Articulatory Coding / DDSP
SimulTron	Streaming S2ST	Time-Synchronized	Conformer / Wait-k Attention
AudioPaLM	Generative S2ST	Large-Scale LLM	Audio Tokens / LLM

Table 1: Major models in the evolution of modern S2ST and RTVT systems.

2 Recent sources highlight the streaming of ability as the motif feature of contemporary RTVT. The streaming models can achieve this by reducing the latency to such an extent that it nearly resembles human simultaneous interpreting.9 Moreover, the advent of zero-shot voice conversion is a significant advance in personalization. Before modern voice synthesis, translation systems made do with one-size-fits-all TTS voices that sounded stiff and impersonal, bearing little resemblance to the original speaker. Alternative algorithms such as Retrieval-based Voice Conversion (RVC) and RT-VC enable so-called voice conversion cloning of the voice of a speaker with little reference material, without losing the emotional prosody that is often lost in written intermediaries.

METHODOLOGY OF REAL-TIME TRANSLATION SYSTEMS

Sender Audio Capture & Preprocessing Receiver

The proposed methodology to build a real-time voice translator consists of the combination of several high-performance neural modules that are designed to be used in a low-latency setting. This section gives an account of the two main architectural solutions and the engineering tricks behind streaming audio.

THE CASCADED PIPELINE ARCHITECTURE

The cascaded architecture has continued to be the standard production grade system because of its modularity and the ability to optimize individual components to specific language pairs or domains.5

1. **Audio Capture and Pre-processing:** The system starts with a capture module, which typically uses WebRTC as a transport protocol, and manages echo cancellation and noise suppression.6 Voice Activity Detection (VAD) is used to divide the audio into chunks or utterances, to ensure that the system only processes active speech.

2. **Streaming Automatic Speech Recognition (ASR):** The audio samples (they may be of 200ms to 400ms duration) are transferred to an ASR engine. The ASR is a source of partial hypotheses, which update in real-time as the speaker proceeds, followed by a final transcript, once a silence or punctuation boundary is detected.
3. **Machine Translation (MT):** The text is sent to an MT engine. In streaming, the MT engine will have to deal with incomplete

sentences or wait until enough context is available. Glossary and domain adaptation is sometimes injected at this point to guarantee accuracy in special jargon.

4. **Streaming Text-to-Speech (TTS):** The text being translated is syllabized into audio. Streaming TTS architectures start playing as the rest of the sentence is being synthesized.5

Processing Stage	Latency Budget (Target)	Technologies Involved
Capture & Transport	60 – 120 ms	WebRTC, Opus, VAD
ASR (Streaming)	150 – 350 ms	Conformer, Whisper, Deepgram
Machine Translation	120 – 350 ms	NLLB, Claude 4.6, DeepL
TTS (Streaming)	75 – 250 ms	ElevenLabs Turbo, Cartesia
Playback	60 – 120 ms	SFU Jitter Buffer

Table 2: Latency budgets of typical high-performance cascaded RTVT system.5

END-TO-END AND DIRECT MODELS

End-to-end models take a more unified approach, collapsing what would otherwise be a chain of separate components into one continuous neural network. An example of this is the SimulTron architecture that employs a Conformer-based streaming encoder to capture both the global and local dependencies within the audio sequence.19 A particular subset of this is Real-Time Voice Conversion (VC). The RT-VC system is an articulatory feature space that separates linguistic content (the "what") of a sentence and the identity of the speaker (the "who"). This is achieved through:

- **EMA Inverter:** A causal model that can predict vocal tract kinematics (articulatory features) given the raw audio.
- **DDSP Vocoder:** A differentiable digital signal processor vocoder that recreates speech based on these articulatory features with ultra-low latency.11

MATHEMATICAL FOUNDATIONS FOR EVALUATION

Measuring the real-world effectiveness of these methods comes down to a handful of standardized mathematical tools that the research community has widely adopted.

WORD ERROR RATE (WER)

The accuracy of the ASR component is quantified by the WER, which is the ratio of errors to the total number of words in the reference transcript:

$$WER = \frac{S + D + I}{N}$$

Where S is substitutions, D is deletions, I is insertions, and N is the number of reference words.[25] A lower WER indicates higher precision. In real-time applications, a "Semantic WER" is often used, where only errors that change the meaning (intent) of the sentence are counted.[27]

BILINGUAL EVALUATION UNDERSTUDY (BLEU)

The quality of translation is measured by the BLEU score, which calculates the n-gram overlap between the machine translation and reference human translations. The formula includes a Brevity Penalty

(BP) to ensure the output is of appropriate length:

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

Where BP is calculated as:

$$BP = \begin{cases} 1 & \text{if } |h| > |r| \\ e^{(1-|r|/|h|)} & \text{if } |h| \leq |r| \end{cases}$$

Here, $|h|$ is the length of the hypothesis and $|r|$ is the length of the reference.[28]

REAL-TIME FACTOR (RTF)

Processing speed is measured by the RTF, defined as the time taken to process an audio segment divided by the duration of that segment:

$$RTF = \frac{\text{Processing Time}}{\text{Audio Duration}}$$

For live systems, $RTF < 1.0$ is mandatory. Some researchers prefer the Inverse Real-Time Factor ($RTFx$):

$$RTFx = \frac{\text{Audio Duration}}{\text{Processing Time}}$$

Where a value of $RTFx = 10$ means the system is 10 times faster than real-time.[23]

FINDINGS AND RESULTS

The findings of this research highlight the performance gap between various architectural paradigms and provide empirical data on the latency-accuracy trade-off.

ACCURACY COMPARISONS ACROSS MODELS

Benchmarking using the FLEURS data (many-to-many translation of multilingual speech) shows that cascaded systems currently hold an edge in the quality of translation in many-to-many translation of multilingual speech, whilst unified models are closing the gap in into-English directions.

Model Architecture	Task	FLEURS (BLEU ↑)	CoVoST-2 (BLEU ↑)
Whisper-Large V2 + NLLB	Cascaded	22.7	31.2
SeamlessM4T-Large	End-to-End	24.1	27.1
SeamlessM4T-V2	End-to-End	26.6	29.4
AudioPaLM-2 (8B)	End-to-End	19.7	24.0
XLS-R-2B-S2T	End-to-End	N/A	22.1

Table 3: Comparative accuracy of translation (BLEU scores) of different language pairs translated into the English language.15

Experimental results indicate that knowledge distillation of powerful MT models into E2E ST models can yield translation quality improvements of up to 0.7 BLEU points.33 But cascaded systems involving Whisper and IndicTrans2 have proven to yield up to 51.0 BLEU points of performance in direct translation of low-resource Indian languages.33

LATENCY AND RESPONSIVENESS

Latency is the first limiting factor towards a human-like conversational flow. The results of the research distinguish between system latency (the technical time of processing) and perceived latency (how long a user perceives that he or she has been waiting).

System / Component	Delay Type	Value	Hardware Context
RT-VC	End-to-End	61.4 ms	CPU (Standard)
Deepgram Nova 3	ASR Finalization	247 ms	GPU Accelerated
Google S2ST	Streaming Delay	2000 ms	Cloud Server
SimulTron	Lookahead	Variable (k)	On-device
Cascaded (Basic)	Cumulative	2000 – 4000 ms	Cloud API

Table 4: Latency results with respect to real-time translation systems and components.2

A RT-VC system has achieved a 13.3% reduction in latency compared to the prior state-of-the-art method (StreamVC) with similar

synthesis quality.11 In cascaded pipelines, the switching to offline to streaming ASR reduces latency by a factor of approximately 9x on long inputs.5

ROBUSTNESS TO REAL-WORLD NOISE

The results show that foundation models are much more effective than task-specific models in the presence of noisy conditions. The Conformer-1 architecture of AssemblyAI achieved a 43% error reduction for noisy speech and 49% error reduction for speaker variation compared to existing SOTA models.16

DISCUSSION

Making sense of these results means grappling with a set of competing priorities: how modular a system is, how fast it runs, and how well it handles the subtleties of human language.

CASCADED VS. END-TO-END DOMINANCE

The data attests to the further supremacy of cascaded pipelines in enterprise settings where control and debuggability are the most important factors. In an E2E system, where audio is synthesized directly, such interventions are far more difficult to implement without introducing a lot of latency.22 Nonetheless, the linguistic latency of cascaded systems, i.e., the delay caused by the need to wait until a sentence boundary has been reached, is a bottleneck. Although E2E streaming systems such as SimulTron may wait as little as a few hundred milliseconds, most commercial cascaded systems wait as long as several seconds to wait until grammatical correctness is achieved.9

THE "VOICE CLONING" REVOLUTION

The field has undergone many changes, but few have felt as personal or as impactful as the shift toward preserving the speaker's own voice and identity in the translated output. E2E systems such as Translatotron 2 and special VC systems like RT-VC have shown that content and style can be disentangled.11 This enables the translated output to sound as though it was spoken by the same person, which is critical in maintaining emotional connection in personal communication or high-stakes negotiations.21

Feature	Cascaded (Traditional)	End-to-End (Modern)
Prosody	Robotic / Generalized	Natural / Preserved
Identity	Lost in text layer	Retained (Voice cloning)
Emotion	Stripped	Captured via audio tokens
Context	Sentence-based	Continuous / Streaming

Table 5: Qualitative variations in user experience in translation paradigms.

AVAILABILITY AND CUSTOMIZED WORK PROCESSES

A special workflow, Communication Access Real-time Translation (CART), is highlighted by the use of RTVT to the DHH community. Studies indicate that even though the accuracy of ASR is on the rise, the professional transcribers are still preferred as the rates of errors are too high in the conditions of noisy group operation of fully automated systems.36 A hybrid solution, where non-professional transcribers correct ASR output in real-time, has been shown to be promising in terms of the ability to increase the accuracy levels to the point where DHH users will rate it positively.36

DIFFICULTIES WITH LOW-RESOURCE LINGUALISTIC

A pattern that can be found throughout findings is the performance disparity between high-resource languages (e.g., English, Spanish, German) and the low-resource ones (e.g., Kazakh, Basque, many Indian languages). In these situations, E2E models fail because they do not have pairs of aligned triplets (speech-text-translation) that have been trained on different datasets, allowing them to outperform simple models by large margins in such situations.

LIMITATIONS AND FUTURE SCOPE

Even with the progress made there are still a number of technical challenges. Computer-based models of speaker identification in group discussions (diarization) remain computationally intensive, and introduce latency.123Current models have a hard time with the so-called code-switching, in which a speaker switches among two languages mid-sentence. It is possible that future research will focus on the so-called "Voice Observability" and the creation of metrics that would correlate with human judgment to a greater degree than BLEU or WER.13

CONCLUSION

Real-time voice translation has quietly crossed a threshold — no longer a laboratory curiosity, it has matured into a practical tool, powered by the convergence of streaming neural architectures and vast multilingual datasets. This study has shown that although cascaded pipelines can offer the modularity and accuracy needed to perform complex enterprise and low-resource language tasks, end-to-end models can provide a more natural, low-latency conversational experience, and the capacity to maintain speaker identity.5 The technical performance standard of RTVT is now at a perceived latency of about 500ms to 2 seconds, which still falls within the range of natural human response times.2 With the reduction of CPU latency to 61.4ms in specialized voice conversion tasks, it is likely that in the future, the focus will not be on raw accuracy, but on preserving paralinguistic nuance and the adherence to global ethical standards, which will ensure that technology is a bridge more than a barrier to global understanding.

REFERENCES

1. Real-Time Voice Translation Project | PDF - Scribd, accessed on April 29, 2026, https://www.scribd.com/presentation/813569219/real-time-voice-translator
2. Real-time speech-to-speech translation - Google Research, accessed on April 29, 2026, https://research.google/blog/real-time-speech-to-speech-translation/
3. Real-Time Voice Translation SDK for Customer Experience - Krisp, accessed on April 29, 2026, https://krisp.ai/blog/real-time-voice-translation-sdk/
4. What are the challenges of real-time speech recognition? - Milvus, accessed on April 29, 2026, https://milvus.io/ai-quick-reference/what-are-the-challenges-of-realtime-speech-recognition
5. Real-Time Speech-to-Speech Translation: Architecture Guide - Deepgram, accessed on April 29, 2026, https://deepgram.com/learn/real-time-speech-to-speech-translation
6. AI Interpretation Platform Development in 2026: A Buyer's and Builder's Guide - Fora Soft, accessed on April 29, 2026, https://www.forasoft.com/blog/article/ai-interpretation-platform-2026
7. Voice AI Architecture Guide: Cascaded vs Speech - TeamDay.ai, accessed on April 29, 2026, https://www.teamday.ai/blog/voice-ai-architecture-guide-2026
8. An Empirical Comparison of Cascade and Direct End-to-End Speech Translation for Low-Resource Language Pair - MDPI, accessed on April 29, 2026, https://www.mdpi.com/2073-431X/15/4/222
9. Real End-to-End Speech-to-Speech Translation is among us - Dr. Claudio Fantinuoli, accessed on April 29, 2026, https://www.claudiofantinuoli.org/2025/11/28/real-end-to-end-speech-to-speech-translation-is-among-us/
10. Speech Benchmarks - CodeSOTA, accessed on April 29, 2026, https://www.codesota.com/browse/speech
11. RT-VC: Real-Time Zero-Shot Voice Conversion with Speech Articulatory Coding - arXiv, accessed on April 29, 2026, https://arxiv.org/html/2506.10289v1
12. Direct Speech to Speech Translation: A Review - arXiv, accessed on April 29, 2026, https://arxiv.org/html/2503.04799v1
13. The voice AI stack for building agents in 2026 - AssemblyAI, accessed on April 29, 2026, https://www.assemblyai.com/blog/the-voice-ai-stack-for-building-agents
14. Best open source speech-to-text (STT) model in 2026 (with benchmarks) | Blog - Northflank, accessed on April 29, 2026, https://northflank.com/blog/best-open-source-speech-to-text-stt-model-in-2026-benchmarks
15. Joint speech and text machine translation for up to 100 languages - PMC - NIH, accessed on April 29, 2026, https://pmc.ncbi.nlm.nih.gov/articles/PMC11735396/
16. SeamlessM4T—Massively Multilingual & Multimodal Machine Translation - Meta AI, accessed on April 29, 2026, https://ai.meta.com/research/publications/seamlessm4t-massively-multilingual-multimodal-machine-translation/
17. RT-VC: Real-Time Zero-Shot Voice Conversion with Speech Articulatory Coding - arXiv, accessed on April 29, 2026,

https://arxiv.org/abs/2506.10289
18. SimulTron: On-Device Simultaneous Speech to Speech Translation - arXiv, accessed on April 29, 2026, https://arxiv.org/html/2406.02133v1
19. [Literature Review] SimulTron: On-Device Simultaneous Speech to Speech Translation, accessed on April 29, 2026, https://www.themoonlight.io/en/review/simultron-on-device-simultaneous-speech-to-speech-translation
20. RT-VC: Real-Time Zero-Shot Voice Conversion with Speech Articulatory Coding, accessed on April 29, 2026, https://www.researchgate.net/publication/394270729_RT-VC_Real-Time_Zero-Shot_Voice_Conversion_with_Speech_Articulatory_Coding
21. Retrieval-based Voice Conversion - Wikipedia, accessed on April 29, 2026, https://en.wikipedia.org/wiki/Retrieval-based_Voice_Conversion
22. Speech-to-Speech vs Cascaded Voice AI: Which Architecture Should You Deploy? - Coval, accessed on April 29, 2026, https://www.coval.ai/blog/speech-to-speech-vs-cascaded-voice-ai-which-architecture-should-you-deploy
23. Best Speech to Text Models 2025: Real-Time AI Voice Agent Comparison - NextLevel.AI, accessed on April 29, 2026, https://nextlevel.ai/best-speech-to-text-models/
24. RT-VC: Real-Time Zero-Shot Voice Conversion with Speech Articulatory Coding - ACL Anthology, accessed on April 29, 2026, https://aclanthology.org/2025.acl-demo.37.pdf
25. Word error rate - Wikipedia, accessed on April 29, 2026, https://en.wikipedia.org/wiki/Word_error_rate
26. Understanding Word Error Rate for Speech Recognition Systems - LlamaIndex, accessed on April 29, 2026, https://www.llamaindex.ai/glossary/what-is-word-error-rate
27. Benchmarking STT for Voice Agents - Daily.co, accessed on April 29, 2026, https://www.daily.co/blog/benchmarking-stt-for-voice-agents/
28. A Survey on Evaluation Metrics for Machine Translation - MDPI, accessed on April 29, 2026, https://www.mdpi.com/2227-7390/11/4/1006
29. (PDF) 449U Evolution of Performance Metrics for Accurate Evaluation of Speech-to-Speech Translation Models: A Literature Review - ResearchGate, accessed on April 29, 2026, https://www.researchgate.net/publication/397602128_449U_Evolution_of_Performance_Metrics_for_Accurate_Evaluation_of_Speech-to-Speech_Translation_Models_A_Literature_Review
30. Evaluation metrics for ASR - Hugging Face, accessed on April 29, 2026, https://huggingface.co/learn/audio-course/chapter5/evaluation
31. Reproducing Speech to Text Translation Results from two prominent foundation models: Whisper and SeamlessM4T - ACM REP, accessed on April 29, 2026, https://acm-rep.github.io/2024/posters/ACM_REP24_paper_31.pdf
32. GenTranslate: Large Language Models are Generative Multilingual Speech and Machine Translators - arXiv, accessed on April 29, 2026, https://arxiv.org/html/2402.06894v2
33. NICT's Cascaded and End-To-End Speech Translation Systems using Whisper and IndicTrans2 for the Indic Task - ACL Anthology, accessed on April 29, 2026, https://aclanthology.org/2024.iwslt-1.3/
34. NICT's Cascaded and End-To-End Speech Translation Systems using Whisper and IndicTrans2 for the Indic Task, accessed on April 29, 2026, https://aclanthology.org/2024.iwslt-1.3.pdf
35. (PDF) SeamlessM4T-Massively Multilingual & Multimodal Machine Translation, accessed on April 29, 2026, https://www.researchgate.net/publication/373297650_SeamlessM4T-Massively_Multilingual_Multimodal_Machine_Translation
36. Communication Access Real-Time Translation Through Collaborative Correction of Automatic Speech Recognition - arXiv, accessed on April 29, 2026, https://arxiv.org/html/2503.15120v1
37. Evolution of Performance Metrics for Accurate Evaluation of Speech-to-Speech Translation Models: A Literature Review - Great Britain Journals Press, accessed on April 29, 2026, https://journalspress.com/LJER_Volume25/Evolution-of-Performance-Metrics-for-Accurate-Evaluation-of-Speech-to-Speech-Translation-Models-A-Literature-Review.pdf
38. Real-Time Multilingual Speech Translation for Peer Communication, accessed on April 29, 2026, https://irjaeh.com/index.php/journal/article/view/944