Real-Time Speech Translation with Voice Cloning

Design, Implementation, and Evaluation of a Modular Real-Time System using Open-Source Models

Students

David Bürge

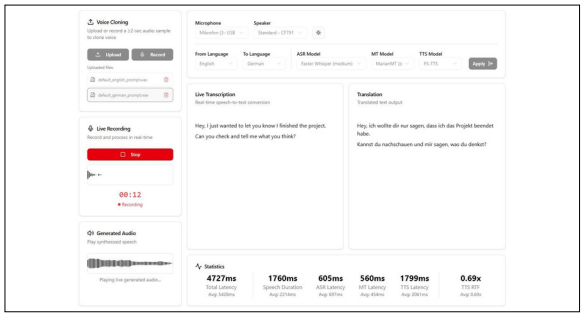
Tareq Kattit

Problem: Advances in neural speech synthesis have made it increasingly easier to clone a person's voice using only a brief audio sample. While this opens new possibilities, it also introduces risks such as impersonation, fraud, misinformation, and privacy violations. At the same time, it remains unclear whether open-source models can produce results that are truly convincing. This work aims to develop and evaluate a real-time system for voice conversion and speech translation based on open-source models running on consumer hardware, with a focus on generating speech that is natural, intelligible, and perceptually convincing.

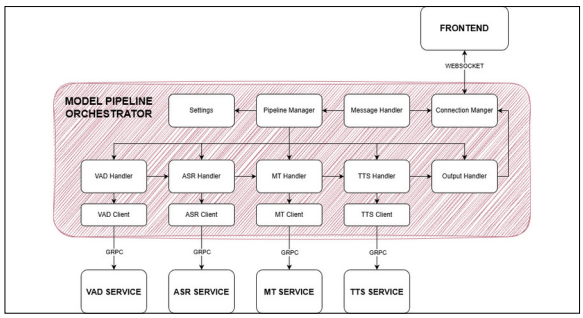
Approach: To achieve this goal, a system is developed integrating and orchestrating open-source models for automatic speech recognition (ASR), machine translation, and text-to-speech synthesis, including Whisper, MarianMT, and text-to-speech models supporting zero-shot voice cloning such as F5-TTS and VoxCPM. The models are selected based on their tradeoff between transcription or synthesis quality and low latency for real-time use. Because the selected models do not support fully streaming inference, the continuous audio input must be segmented before processing. Voice activity detection (VAD) is therefore used to detect short pauses in the incoming speech and split the audio stream into manageable segments. These segments are then processed sequentially by the ASR, machine translation, and speech synthesis components. To reduce transcription errors caused by short noise bursts, the VAD was configured to emit segments with a minimum duration of 300 milliseconds and with reduced sensitivity. The quality and credibility of the generated speech are evaluated through a combination of automatic metrics and human assessments, focusing on naturalness, intelligibility, and voice similarity, while also accounting for end-to-end latency.

Result: The system shows that real-time voice conversion can be achieved on consumer hardware using open-source models. For short German audio samples, the generated speech is perceptually convincing, achieving an average naturalness score of 4.28/5.0 and a voice similarity score of 0.8/1.0, with comparable results for English synthesis. When speech translation is enabled, the end-to-end system achieves an average latency of approximately 1.7 seconds, rising to about 3.9 seconds when speech duration is taken into account, which remains suitable for real-time use. While the overall translation quality is acceptable for short, isolated segments, limitations emerge for longer speech. The machine translation models used do not support providing preceding segments as context, resulting in more literal, sentence-by-sentence translations, while nuances spanning multiple segments are lost. In addition, the modular design allows individual models to be swapped, enabling experimentation with alternative configurations and supporting future extensions. Overall, these results confirm the feasibility of building real-time voice cloning systems with current open-source models and highlight both their accessibility and potential risks they entail.

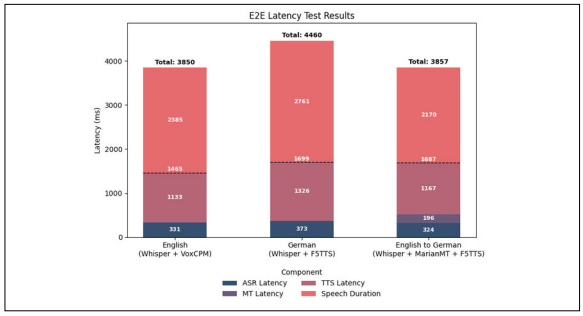
Screenshot of the Frontend
Own presentment



Model Pipeline Orchestrator Component Diagram
Own presentment



Latency Test Results
Own presentment



Advisors
Prof. Dr. Mitra Purandare, Prof. Dr. Markus Stolze

Subject Area
Software Engineering, Artificial Intelligence