Translatica: A Survey and Implementation Study on Speech-to-Speech Translation and Voice Synthesis with Speaker Preservation

by

Baaz Jhaj

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Bachelor of Science

Approved April 2025 by the
Thesis Committee:

Dr. Steven Osburn, Chair
Dr. Haolin Zhu #2

Arizona State University
May 2025

Abstract

This thesis presents *Translatica*, a modular speech-to-speech translation (S2ST) system that preserves both linguistic meaning and the speaker's vocal identity across languages. Alongside developing a working prototype, this work surveys the landscape of S2ST methods and motivates the choice of a modular architecture over direct approaches, emphasizing flexibility, interpretability, and voice fidelity. The system combines state-of-the-art tools in transcription, translation, and voice synthesis to enable expressive, speaker-preserving dubbing of prerecorded videos. Through implementation and evaluation, the thesis explores the trade-offs between accuracy, latency, and control, demonstrating how modular design enables customization for diverse use cases. Future work includes real-time translation, enhanced speaker tracking, and applications in education and live media.

Contents

1 Introduction

Language is one of the most powerful expressions of our humanity. It allows us to share not only information, but our inner lives: memories, emotions, humor, culture, and identity. Through language, we pass down history, express love and grief, resolve conflict, and create shared meaning. It is how we recognize each other as human [4].

Yet language can also be a boundary. When we cannot understand the words someone speaks, connection becomes harder to establish. Miscommunication, or silence, can make others feel foreign, distant, even threatening. This is not merely a social inconvenience; it has real consequences. Historically, the absence of a shared language has often enabled dehumanization. In times of war, for instance, it becomes easier to harm those we do not understand. Psychological research and military accounts suggest that soldiers are more likely to dehumanize enemies who speak an unfamiliar language or none at all [6, 8]. The linguistic gap creates distance, and that distance can dull empathy.

Conversely, when someone speaks our language, or even tries to, we instinctively view them as more relatable and trustworthy. Language doesn't just carry words; it conveys identity, tone, and emotion. And when spoken in a familiar voice, it bridges not only linguistic divides but emotional ones as well [16].

These insights form the foundation of the Translatica project. In an increasingly global world, the ability to communicate fluidly across languages, without losing the personality or presence of the speaker, has become deeply valuable. Real connection demands more than just translated words; it requires rhythm, tone, and voice. This thesis explores how modern AI technologies can be used to translate not only what is said, but how it is said, capturing the speaker's sound, intent, and style to make cross-linguistic communication feel more human.

The demand for such a system is clear. From global meetings and education to entertainment and media, real-time and emotionally resonant translation is increasingly important. While subtitles offer accessibility, they lack the immediacy and intimacy of hearing speech in one's own language, especially when that voice reflects the original speaker's character. That gap is what Translatica aims to close: a system designed for near real-time speech-to-speech (S2S) translation that preserves both meaning and vocal identity.

But Translatica's goal extends beyond converting speech. It aims to make translation feel personal and natural, as if the speaker is genuinely speaking your language. Realizing this vision requires carefully integrating four core technologies: automatic speech recognition (ASR), speaker diarization (SD), machine translation (MT), and text-to-speech synthesis (TTS). Each presents unique challenges in terms of accuracy, latency, and expressiveness, and combining them into a cohesive pipeline adds further complexity.

Translatica is a modular S2ST system that performs spoken language translation while preserving the speaker's voice and emotional tone. It uses OpenAI's Whisper for robust ASR [15], PyAnnote for speaker diarization and segmentation [2], GPT-based models for fluent, context-sensitive translation [14], and a suite of TTS systems, including F5 TTS [3], Google Cloud TTS [7], and UniAudio [17], for expressive voice generation. Each was selected or adapted to balance performance, latency, and speaker fidelity.

Modularity was a central design principle. It enabled iterative refinement and targeted control at every stage of the pipeline, transcription, translation, and synthesis. While end-to-end systems like SeamlessM4T [1], SeamlessExpressive [13], and UnitY [10] offer tightly integrated architectures with low latency and strong prosody retention, they are harder to interpret, adapt, or debug. In contrast, Translatica's modular structure supports precise interventions, such as using ChatGPT prompts to guide tone or fine-tuning F5 TTS for accent adaptation. This flexibility makes modular pipelines especially suitable for niche domains, speaker-aware dubbing, and experimentation with expressive synthesis.

This thesis is both a technical survey and an implementation study. It examines the trade-offs between direct and modular S2ST systems, evaluates voice synthesis models, and analyzes transcription, translation, and diarization strategies. It presents Translatica, a modular system that translates and dubs pre-recorded videos while preserving the original speaker's voice in a new language. The system's architecture, design decisions, and limitations are discussed in detail.

Ultimately, Translatica is more than a translation tool, it is a step toward breaking barriers and humanizing communication. The pages that follow review relevant technologies, compare S2ST strategies, and describe the system's development and evaluation. The goal is not just to cross language barriers, but to preserve the human presence behind the words.

2 Background

2.1 The Speech-to-Speech Translation (S2ST) Pipeline

Speech-to-speech translation (S2ST) refers to the task of converting spoken language in one language into spoken output in another, ideally preserving both meaning and vocal expression. Most S2ST systems fall into one of two categories: modular (or cascaded) systems, and direct (end-to-end) models.

Translatica follows the modular pipeline, which is composed of three main components:

Automatic Speech Recognition (ASR)

This component transcribes the input speech into text in the original language. For Translatica, OpenAI's Whisper was used, an advanced ASR model known for its robustness to accents and noisy environments. Whisper generates both a transcript and timestamps, which are essential for later synchronization [15].

Speaker Diarization (SD)

To support multi-speaker content, Translatica incorporates speaker diarization using PyAnnote [2], which segments audio based on speaker identity. This step ensures that translated output retains speaker boundaries, enabling voice-preserving synthesis and better temporal alignment during dubbing. Diarization is especially important for educational, interview, or panel content, where multiple speakers alternate rapidly or overlap.

Machine Translation (MT)

After transcription, the source language text is passed to a translation model. Translatica primarily uses OpenAI's GPT-based models, which offer strong fluency and contextual understanding [14]. Comparisons were also made to APIs like Google Translate or Amazon Translate, with GPT providing superior flexibility for in-context translation.

Text-to-Speech Synthesis (TTS)

The final step in the pipeline is converting translated text back into speech. In building Translatica, I evaluated several TTS systems, including Google Cloud TTS [7] for its speed and multilingual coverage, and F5 TTS [3] for its voice cloning and expressiveness. While each had strengths, F5 TTS was ultimately selected for its balance of low-latency synthesis, natural prosody, and ability to preserve speaker identity, making it the most suitable option for modular, voice-preserving translation.

2.2 Role of Prosody and Speaker Consistency

Prosody, the rhythm, stress, and intonation of speech, plays a vital role in conveying emotion and intent. A translation system that ignores prosody can misrepresent meaning or emotional nuance. Similarly, maintaining speaker identity is essential for trust and coherence in dubbed content.

Translatica addresses these by exploring speaker-specific TTS models and exploring expressive vocoders where feasible. While fine-tuning can add computational cost, it can significancy improve speaker similarity and emotional resonance.

2.3 Direct S2ST Approaches: Vocoder-Based Architectures

Unlike modular systems, direct S2ST models such as Meta's SeamlessM4T and SeamlessExpressive skip intermediate text entirely [1, 13]. These models map source audio into discrete acoustic units that encode both phonetic and prosodic information. A unit vocoder, such as HiFi-GAN [11], then reconstructs target language speech directly from these representations.

Advantages:

- Lower latency, especially with streaming models [13]
- Better prosody retention and natural pacing [1]

- Fewer moving parts reduces cascading errors [12]

Limitations:

- Difficult to interpret or fine-tune [10]
- Less adaptable for voice cloning or accent control
- High data demands for parallel S2S training [1]

Translatica's modular approach allows for experimentation at each stage, making it ideal for prototyping, debugging, and targeted fine-tuning. Beyond flexibility, modularity enables an ensemble-style design, where stacking specialized models for ASR, translation, and synthesis can produce results that exceed the capabilities of any single system. This composition allows strengths to be combined and weaknesses isolated, making the whole greater than the sum of its parts. While direct S2S systems may dominate future commercial deployments, modular approaches are likely to grow increasingly competitive as individual components improve, offering a scalable and adaptable path to high-quality, voice-preserving translation.

2.4 Limitations in Existing Systems

Whether modular or direct, current S2ST systems share several unresolved challenges:

- **Latency:** Real-time translation remains demanding [1].
- **Domain generalization:** Models often fail on informal, technical, or low-resource language.
- **Emotion and expressivity:** Flat intonation remains a barrier for truly human output.
- **Speaker consistency:** Voice cloning still lacks robustness under accent or noisy input.
- **Multilingual reliability:** Underperformance in non-English or underrepresented dialects [1].

Translatica navigates these trade-offs through a modular approach, leveraging open-source models alongside fine-tuning and prompt engineering techniques to enhance expressiveness and speaker realism.

3 Direct vs. Modular Speech-to-Speech Translation

Recent efforts in speech-to-speech translation (S2ST) fall into two broad paradigms: direct end-to-end models and modular cascaded systems. Direct S2ST models aim to translate speech to speech in one unified model without explicit intermediate text, whereas modular approaches pipeline separate automatic speech recognition (ASR), machine translation (MT), and text-to-speech (TTS) components. We survey key examples of each approach, their architectural characteristics, and the strengths and weaknesses that emerge from these design choices.

3.1 Direct S2ST Models and Architectures

Direct speech-to-speech translation (S2ST) models perform translation without producing text transcripts as an intermediate step. Instead, they often rely on an intermediate acoustic representation. A pioneering example is the Fairseq S2UT model (Lee et al., 2021), which introduced speech-to-unit translation [12]. In this approach, the model predicts a sequence of discrete acoustic units, compact sound-like tokens that represent the translated speech. These units are then passed to a vocoder, a neural speech synthesizer that converts them into waveform audio.

The acoustic units themselves are learned using self-supervised speech encoders, such as HuBERT [9] or EnCodec-based quantizers [5], which are applied to target-language speech to build a vocabulary of meaningful audio tokens [12]. This approach has shown to outperform earlier methods that relied on predicting raw spectrograms, leading to more stable training and better audio quality. The vocoder, typically trained separately on target-language audio, plays a crucial role in reconstructing intelligible and natural-sounding speech from these units.

Modern direct S2ST systems aim to unify speech recognition, translation, and synthesis within a single model. Facebook's S2UT, for example, is trained to predict both translated text and acoustic units simultaneously. This joint modeling strategy combines the strengths of both representations: text provides linguistic structure that guides learning, while acoustic units retain the expressive qualities of speech. Training the model to align language content with speech rhythm in this way leads to more natural and fluent output, especially beneficial in low-resource settings where training data is limited [12].

Other models follow similar strategies. Google's Translatotron 2 and Meta's UnitY use two-stage designs, first predicting either phonemes or translated text, then generating speech [10]. These intermediate steps act as anchors during training, improving both optimization and translation quality by simplifying the learning task.

Training direct S2ST models requires large datasets of parallel speech, where audio in one language is aligned with its spoken translation in another. These models often begin with self-supervised pretraining, using architectures like HuBERT or w2v-BERT to learn general-purpose speech features from unlabeled audio, followed by supervised fine-tuning on speech translation pairs [12]. In cases where real parallel data is scarce, researchers use pseudo-labeling, generating synthetic translations to expand training coverage.

Meta's SeamlessM4T pushes this framework to a global scale. It supports over 100 languages using a single unified architecture trained on millions of hours of real and synthetic multilingual speech [1]. Its specialized variants, SeamlessExpressive, which captures vocal style and emotion, and SeamlessStreaming, which enables low-latency, real-time translation, demonstrate the flexibility and expressive potential of direct S2ST systems [13].

In summary, modeling both text and acoustic units jointly improves training stability and performance, allowing direct S2ST models to generate fluent, expressive speech translations while preserving the speaker's vocal identity and emotional tone.

3.2 Strengths and Weaknesses of Direct Approaches

One of the key strengths of direct speech-to-speech translation (S2ST) models is their ability to preserve the expressive qualities of speech, things like tone, emotion, rhythm, and even the speaker's unique voice. These fentures, known collectively as paralinguistic information, often get lost in traditional systems that convert speech into text before translating. Since direct models skip the text step and map speech directly to translated speech, they can maintain a closer resemblance to how the original speaker sounded. For example, models like Translatotron and SeamlessExpressive have demonstrated the ability to retain vocal style and speaker identity in the translated output, something that's much harder to do in cascaded systems [13].

Direct models aim to reduce latency by unifying transcription, translation, and synthesis into a single end-to-end process. In theory, this eliminates the overhead introduced by cascading separate modules and simplifies real-time execution. However, in practice, these models are often large and computationally intensive, requiring powerful GPUs and substantial memory to operate efficiently. As a result, they are less suited for live translation or deployment on edge devices, where both speed and resource efficiency are critical. Paradoxically, modular systems, despite their complexity, can often achieve better real-time performance by relying on optimized, lightweight components tailored to each stage.

That said, direct models offer unique advantages in certain contexts. One of their most promising features is the ability to operate without relying on written language resources. Traditional modular systems depend heavily on text data to train translation models, but many of the world's languages, especially those that are primarily spoken, lack standardized orthographies or sufficient textual corpora. Direct models, by training directly on speech-to-speech pairs, can support oral and low-resource languages that text-based systems cannot reach [1].

Despite these strengths, direct approaches face several persistent challenges:

- **Lower translation accuracy:** Compared to modular systems, direct models have historically underperformed in formal evaluations. Because they must simultaneously learn transcription, translation, and speech synthesis, the optimization task is more complex. While newer models like SeamlessM4T are closing the gap, modular pipelines still tend to deliver more accurate translations, especially in complex or real-world domains [1].

- **Data requirements:** Training effective direct models requires large amounts of parallel speech data, recordings of the same content in two different languages. This kind of data is much harder to collect than text-based translations. Even with techniques like self-supervised pretraining and data augmentation, many language pairs remain underrepresented.

- **Lack of transparency and control:** Because direct models skip the text stage, there's no easy way to inspect or correct mistakes. If the system makes an error, it's difficult to know whether it misheard the original or mistranslated it. There's also no simple way to adjust the output (for example, fixing a mispronounced name) without retraining the entire model. In contrast, modular systems allow for targeted fixes and easier debugging.

- **Limited control over voice and style:** While direct models can carry over some speaker traits, this behavior is difficult to control precisely. For instance, they may unintentionally imitate the speaker's voice, raising ethical concerns around voice cloning. Some systems, like Translatotron 2, had to implement specific design changes to ensure that only the speaker's own voice is reproduced.

- **Training complexity:** Balancing the goals of transcription, translation, and speech generation requires careful model design. Training can become unstable, especially if the data is imbalanced or varies widely in quality. Additionally, evaluating output is not straightforward, researchers often use ASR to transcribe the generated speech just to calculate accuracy metrics like BLEU scores, which can introduce noise and reduce reliability.

In short, direct S2ST systems are powerful for preserving voice and reducing latency, and they open new possibilities for underserved languages. However, they also come with trade-offs: they are harder to train, less transparent, and currently less accurate than traditional cascaded systems. Future research will likely focus on bridging these gaps while maintaining the advantages of direct, unified translation pipelines.

3.3 Modular Cascaded S2ST Approaches

3.4 Trade-offs: Flexibility, Robustness, and Deployability

- **Translation Quality vs. Prosody Preservation:** Cascaded systems often produce more accurate translations because each component, ASR, translation, and TTS, can be individually optimized or replaced with the best available model. Direct models, while less modular, excel at preserving speech nuances like tone, emotion, and voice identity, making them ideal when expressiveness is key.

- **Flexibility and Control:** Modular systems allow you to swap or fine-tune individual components (like the translation or voice synthesis model), making them easier to adapt to specific domains or use cases. Direct models are harder to modify or debug because everything is handled inside one unified system.

- **Speaker-awareness:** Modular pipelines can integrate speaker diarization to enable multi-voice dubbing, speaker-specific translation strategies, or targeted synthesis styles, something not easily supported in direct models.

- **Latency and Real-Time Performance:** Direct models offer lower latency since they generate translations in a single step. Cascaded systems may introduce more delay but benefit from more efficient and well-optimized individual components.

- **Domain and Style Adaptability:** Modular pipelines make it easy to specialize just one part: for instance, swapping in a different translation model, or updating the voice synthesis engine. In contrast, direct models require retraining on new speech pairs for even small changes, limiting flexibility for niche applications.

- **Multilingual and Low-Resource Coverage:** Cascades can support many language pairs by combining existing ASR, MT, and TTS components, even using English as an intermediate "pivot." Direct models need massive training across many languages (using parallel datasets) but are uniquely suited to oral or unwritten languages where no text exists.

- **Real-World Deployability:** Cascades integrate easily into existing systems and allow for modular testing and logging, making them practical for many companies. Direct models are more self-contained and efficient at runtime, but may require more infrastructure and care to ensure safety, interpretability, and compatibility.

In conclusion, direct vs. modular S2ST is not a one-size-fits-all choice; it depends on priorities. Direct models offer a unified solution that can yield more natural-sounding translations with low latency, at the cost of enormous data requirements and less interpretability. Modular approaches offer proven quality and flexibility, but entail more complexity and may discard useful speech information. Ongoing research (from Translatotron 3 to SeamlessM4T) is rapidly closing the quality gap from direct systems, while retaining benefits like voice preservation and streaming. It is foreseeable that hybrid approaches might emerge – for instance, cascades augmented with prosody transfer modules, or end-to-end models with intermediary supervision – to get the best of both worlds. For now, the choice must consider the specific application: whether one values accuracy and control (favoring a cascade) or naturalness and integration (favoring direct). The exciting progress from 2022–2025 suggests that fully end-to-end speech translators will become increasingly viable for real-world use, but modular systems will remain a strong baseline and often a safer bet when accuracy and customization are paramount. Each approach has its merits, and together they push the frontier toward the long-standing goal of a universal translator, one that can not only cross language barriers, but do so with the voice, tone, and presence of the original speaker intact.

4 Neural Voice Synthesis Models for the Translatica Pipeline

In a speech-to-speech translation system like Translatica, the voice synthesis module must generate translated speech that sounds natural and, ideally, preserves the original speaker's vocal identity and expressiveness. While traditional text-to-speech (TTS) models focus on generating speech from written text, our pipeline also explored more flexible approaches to voice synthesis, models that operate on acoustic prompts or enable prosody transfer. Key criteria for evaluating these systems include naturalness (human-like sound and poetic qualities), speaker similarity (maintaining the original speaker's timbre if doing voice cloning), expressiveness (capturing emotion and intonation), latency (fast, real-time synthesis for live use), and cloning capability (whether the system can mimic voices with zero-shot input or requires fine-tuning). We evaluate three approaches, F5 TTS, Google Cloud TTS, and UniAudio, against these criteria, highlighting their strengths, limitations, and suitability for different components within the Translatica pipeline.

4.1 F5 TTS

F5 TTS [3] is a recent open-source TTS model designed for rapid ultra-realistic voice cloning in real time. It stands for ”Fairytaler that Fakes Fluent and Faithful Speech with Flow Matching.” F5 TTS introduces an advanced architecture with a Diffusion Transformer (DiT) and Flow Matching techniques to generate speech without the need for explicit phoneme alignment or duration prediction. A ConvNeXt-based component refines text representations for better speech alignment.

Zero-Shot Voice Cloning: F5 can mimic a speaker's voice from only 10 seconds of reference audio, with no additional fine-tuning required. This enables it to closely match an original speaker's timbre and accent on new text.

High Naturalness and Expressiveness: It produces highly natural, expressive speech with lifelike intonation. The model can convey emotions and control speaking speed, adding richness to the synthesized voice.

Multi-Language Support: F5 TTS is multilingual (demonstrated in English and Chinese) and even allows switching between languages in one utterance. This is valuable for translating mixed-language content.

Real-Time Performance: A sentence takes only a few seconds to synthesize, which is fast enough for live translation use.

Fine-Tuning: While F5 TTS excels at zero-shot voice cloning, it also supports fine-tuning on custom datasets to enhance accent adaptation and speaker similarity. For instance, a Spanish variant was fine-tuned on over 218 hours of Spanish accents, resulting in more natural and expressive Spanish speech synthesis.

Use Cases: F5's blend of speed and quality makes it ideal for interactive systems (e.g. voice assistants or live translators) where both naturalness and low latency are required. Its zero-shot voice cloning enables translating a speaker's speech while keeping their voice, a crucial capability for Translatica when preserving the original speaker's identity. However, F5 TTS's multilingual support is currently limited; while there is an official model for English and Chinese, support for other languages requires specific fine-tuning efforts.

4.2 Google Cloud TTS

Google Cloud Text-to-Speech [7] is a commercial TTS service offering a broad range of pretrained, high-quality voices across 40+ languages and dialects. It uses neural speech synthesis models, such as WaveNet and the newer Neural2 architecture, to generate speech with clear pronunciation and fluid pacing. While technically impressive, the output often lacks emotional depth and can sound robotic or overly neutral, especially in longer or expressive speech scenarios.

Naturalness: Google's Neural2 voices capture basic intonation and rhythm well, producing smooth and intelligible audio. However, their expressiveness is limited, and they rarely convey strong emotion or speaker personality unless a specific voice style is used. This makes them suitable for general-purpose narration but less effective for emotionally rich dialogue or humanlike interaction.

Speaker Variety and Cloning: Users can choose from many predefined voices, but voice cloning is not supported. Google does offer a Custom Voice program that allows businesses to train a new voice using hours of studio-quality recordings, but this process is resource-intensive and inaccessible for most users. As such, all outputs from Google TTS sound like one of its default voices, not the original speaker, which breaks the illusion of seamless dubbing.

Real-Time Performance: One of Google TTS's strengths is its low latency. As a cloud-based service, it can generate short speech segments in under a second, with response times typically between 200 and 1000 ms depending on the voice. This makes it viable for real-time applications, assuming internet access and API integration are available.

Expressiveness and Control: Developers can adjust basic parameters such as pitch, speed, and pause timing using SSML (Speech Synthesis Markup Language), but fine-grained emotional control is limited. Voice style changes, like shifting from formal to excited tone, are only possible with certain predefined voice variants. Unlike research-grade TTS models, it cannot dynamically adjust delivery or mimic a specific person's expressive traits.

Use Cases: Google Cloud TTS is a production-ready solution for rapid, consistent voice output. In Translatica, it serves as a solid fallback when speed and language coverage are more important than voice fidelity. However, for scenarios where preserving the speaker's voice and emotional tone is essential, Google TTS falls short. It could be used in combination with other tools (like a voice conversion model) to personalize output, but by itself it lacks the expressiveness and speaker consistency needed for emotionally faithful translations.

4.3 UniAudio

UniAudio [17] is a cutting-edge research model designed for flexible, high-fidelity audio generation across multiple speech tasks. It leverages discrete token-based audio representations and a transformer-based architecture to perform zero-shot voice cloning, prosody transfer, and speech synthesis. In Translatica, we used UniAudio in a style-transfer pipeline inspired by Wang et al. (2023) from Zhejiang University, where the model served as an acoustic language model to perform expressive synthesis without needing speaker-parallel data.

Naturalness: UniAudio produces highly expressive speech with rich prosody, capturing subtle cues such as rhythm, pacing, and emotional tone. It can reflect personality and vocal texture when given appropriate reference prompts. This expressiveness makes it suitable for applications where tone and nuance matter. However, variation between outputs can occur depending on prompt length and acoustic conditioning.

Speaker Variety and Cloning: UniAudio enables zero-shot speaker transfer through prompt-based in-context learning. In our usage, mirroring the referenced paper's three-stage pipeline, we supplied both semantic content and acoustic style prompts to generate translated speech in the original speaker's voice. While the results were high-quality in isolated examples, we encountered significant consistency issues: when splitting a speaker's audio into multiple segments, as is common when processing long videos, UniAudio often produced outputs that sounded like very similar yet distinct voices, disrupting the sense of a continuous, unified speaker throughout the dubbed content.

Real-Time Performance: Due to its autoregressive nature and large model size (760M parameters), UniAudio is not designed for real-time synthesis. Each segment requires significant compute time, making it better suited for batch processing of pre-recorded media rather than live translation or edge deployment. Latency and throughput remain key limitations in practical use cases.

Expressiveness and Control: The model offers excellent control over vocal expressiveness via prompt engineering. It can adapt delivery style, intonation, and pacing, provided the style prompt is well-matched. However, achieving consistent output over long recordings remains challenging. The same speaker prompt may yield noticeably different results across segments, likely due to sensitivity in prompt conditioning and token-level variation in synthesis.

Use Cases: At present, UniAudio is best viewed as a proof-of-concept tool for expressive voice synthesis and style transfer. In Translatica, we used UniAudio because we were interested in the idea of embedding acoustic qualities, such as prosody and vocal style, into our translation models, exploring how expressive features could be preserved across languages. However, due to its speaker drift across segments and relatively slow inference, we ultimately favored F5 TTS for deployment, as it provided faster synthesis and more consistent speaker continuity. Still, an approach similar to UniAudio, or grounded in the same conceptual framework of prompt-based acoustic generation, may represent the future of expressive S2ST. As this architectural paradigm matures, a more stable and efficient version could enable high-fidelity dubbing pipelines that not only preserve vocal identity but also capture the full prosodic nuance of the original speaker, something current models like F5 can only partially achieve.

4.4 Choosing the Right Voice

Each of the three voice synthesis models explored, F5 TTS, Google Cloud TTS, and UniAudio, offered distinct advantages and limitations within the context of speech-to-speech translation. F5 TTS emerged as the most balanced option for Translatica, offering high-quality, low-latency voice cloning with minimal setup, making it well-suited for real-time or interactive applications. Google Cloud TTS provided speed and multilingual coverage at production scale, but lacked the personalization and emotional nuance needed for speaker-preserving dubbing. UniAudio, though not deployable in its current form, introduced a promising conceptual direction for expressive synthesis through prosody transfer and acoustic prompting.

Ultimately, our choice of F5 reflected a practical trade-off between speed, consistency, and speaker similarity. However, the insights gained from UniAudio point toward a future in which voice synthesis systems can more fully capture the richness and individuality of human speech. As research continues to advance, hybrid approaches that combine the efficiency of real-time TTS with the expressive depth of models like UniAudio may define the next generation of multilingual dubbing systems, bringing us closer to natural, voice-faithful cross-lingual communication.

5 Transcription, Translation, and Diarization Strategies

5.1 Transcription via Whisper

The first step in the Translatica pipeline is accurate speech transcription. We used OpenAI's Whisper [15], a powerful automatic speech recognition (ASR) model trained on large-scale, multilingual data. Whisper is particularly well-suited for real-world speech due to its robustness to background noise, varied accents, and spontaneous conversational styles. It also provides word-level timestamps, which are essential for synchronizing translated and synthesized audio with the original video.

By starting with a strong transcription backbone, we ensured that downstream translation and synthesis modules operated on accurate, temporally aligned text inputs.

5.2 Speaker Diarization via Pyannote

For multi-speaker audio, transcription alone is not enough. To preserve speaker identity and enable speaker-specific translation and synthesis, we integrated speaker diarization into the pipeline using the pyannote-audio toolkit. Pyannote [2] uses pre-trained neural models to segment audio by speaker, even in cases of overlapping or rapid turn-taking dialogue.

This allowed us to tag each segment of transcribed speech with a speaker label, enabling voice-preserving synthesis and the potential for speaker-adaptive translation (e.g., adjusting tone or formality based on speaker identity). Diarization also allowed for better organization of longer transcripts and supported clearer temporal alignment during dubbing

5.3 Translation via ChatGPT

One of the most important aspects of speech-to-speech translation (S2ST) is not just the literal accuracy of the translation, but the ability to preserve the tone, style, and intent of the original speaker. For Translatica, we integrated OpenAI's ChatGPT [14] to perform text-based translation from the source language transcript to the target language output. While not a dedicated machine translation (MT) model, ChatGPT offers several compelling advantages for high-quality, flexible translation in a modular S2ST system.

Advantages for Fluency and Tone Matching: Unlike traditional MT systems which prioritize lexical or grammatical correctness, ChatGPT can adapt to context and imitate tone, style, and social nuance. This makes it especially useful for conversational or emotionally expressive speech, where a more literal translation may sound robotic or tone-deaf. Through few-shot prompting and dialogue context, it was possible to produce translations that sounded more natural, personable, or culturally appropriate.

Latency Considerations: Since ChatGPT is accessed via an API and uses large-scale generative models, it introduces non-trivial latency to the pipeline. On average, a sentence-level translation using the API takes 1 to 3 seconds depending on prompt length and system load. While not prohibitive for batch processing or dubbed video generation, this delay could be a bottleneck for real-time applications. Still, the trade-off between latency and quality was often acceptable given the improvements in fluency.

Prompting Strategies and Quality Observations: To improve output quality, we experimented with a variety of prompting techniques, such as asking ChatGPT to match the emotional tone of the original sentence or to keep sentence length close to the original for better timing alignment. In many cases, explicitly guiding the model with phrases like ”Translate this while keeping a casual tone” produced better results than default usage. The model also handled idiomatic expressions and informal speech better than traditional MT systems, especially when given examples or brief instruction. However, for highly technical or domain-specific content, its performance was more variable.

Another strength of using ChatGPT was its ability to utilize broader context. In our pipeline, we provided ChatGPT with full transcripts in addition to the specific sentence being translated. This allowed it to correct minor ASR errors in surrounding text and produce more coherent, contextually accurate translations, especially in dialogue where speaker intention unfolds across multiple lines. By retaining the full transcript context, ChatGPT was better able to disambiguate meaning, maintain consistent tone, and resolve pronouns and named entities more reliably.

In summary, ChatGPT served as a powerful component in the Translatica pipeline for generating expressive and context-sensitive translations. While not suitable for ultra-low-latency needs, its flexibility and stylistic control made it a strong choice for dubbed media and other high-quality translation outputs.

6 Future Work

While Translatica functions effectively as a modular system for translating pre-recorded video content, several areas remain open for enhancement, particularly in terms of speed, speaker tracking, and expanding real-world use cases. These improvements aim to make the system more scalable, expressive, and responsive across domains.

Multiprocessing for Static Video Translation: One of the most immediate optimizations involves implementing multiprocessing for static (non-real-time) video translation. The current sequential pipeline, transcription, translation, and synthesis, can be slow for longer videos. By dividing the video into segments (e.g., by speaker or sentence boundary) and processing them in parallel across multiple CPU or GPU threads, we can significantly accelerate the total processing time. For instance, in internal tests, breaking a 15-minute lecture into five parts and running them simultaneously reduced total translation time by over 60%. This improvement would be particularly impactful for content creators, educators, and localization teams working with large media libraries.

Improved Diarization and ASR via Video Understanding: While Translatica currently uses PyAnnote [2] for audio-based speaker diarization, this approach can struggle in environments with overlapping speech, similar-sounding voices, or background noise. Incorporating visual features, such as face detection, lip motion tracking, and speaker gaze, could significantly enhance both diarization and ASR performance. In multi-speaker settings like roundtable discussions or interviews, visual cues can help disambiguate who is speaking, assign speech segments more accurately, and ensure smoother speaker continuity in dubbing. This is especially valuable in educational content or interviews, where maintaining consistent speaker identity is critical for clarity and listener engagement.

Real-Time Applications and Live Streaming: A major direction for future development is enabling real-time use cases. This includes integrating Translatica into platforms like Zoom, Google Meet, or Microsoft Teams, where participants could hear translated speech in near real time with retained vocal identity. Other applications include customer support centers and multilingual event broadcasts. Live streaming is also a promising target, using a modified, streaming-friendly version of Translatica, it would be possible to dub a livestream into multiple languages simultaneously. While this requires minimizing latency at every stage of the pipeline, the potential impact is substantial: global accessibility for live broadcasts without losing the presence and tone of the original speaker.

Applications in Educational Platforms: Translatica is already well-suited to traditional lecture videos where a presenter speaks in front of slides. In fact, many of our early use cases involved dubbing university lectures for multilingual audiences. Looking ahead, we are actively exploring how to adapt not only the audio but also the accompanying visual content. For example, by detecting and translating on-screen text in presentation slides (e.g., via OCR and language models), we could automatically generate fully translated versions of lecture recordings, complete with dubbed speech and edited slides in the target language. This has significant potential for MOOCs, open education platforms, and international online courses.

Together, these future directions point toward a more expressive, scalable, and real-time Translatica, one that brings multilingual voice translation closer to the seamlessness of in-person communication, and expands access to knowledge, events, and conversations across languages and borders.

7 Conclusion

This thesis presented the design, implementation, and evaluation of *Translatica*, a modular speech-to-speech translation (S2ST) system built to preserve both linguistic meaning and vocal identity across languages. In an era where language remains a key barrier to empathy and communication, Translatica aims to make translation feel not only accurate, but human.

Through a combination of state-of-the-art models in automatic speech recognition (ASR), machine translation (MT), and voice synthesis, we constructed a pipeline capable of translating pre-recorded videos while maintaining speaker similarity, tone, and emotional nuance. Our evaluation of both modular and direct S2ST models helped clarify key trade-offs, between accuracy and expressiveness, control and latency, and informed our decision to adopt a modular architecture enhanced with prompt engineering and zero-shot voice cloning.

We also explored a range of voice synthesis methods, ultimately choosing F5 TTS for its strong balance between speed, naturalness, and speaker fidelity. Tools like ChatGPT enabled context-aware, emotionally fluent translations, while future-facing models like UniAudio suggested new possibilities in expressive dubbing. This thesis makes two key contributions: first, it surveys emerging research at the intersection of translation and voice identity; second, it presents a working prototype capable of translating speech across languages while preserving the speaker's personality and vocal presence. As global communication increasingly relies on richer, more immediate forms of interaction, systems like Translatica will be vital for creating inclusive, accessible, and emotionally resonant experiences.

Future work will extend Translatica's capabilities toward real-time performance, live streaming, educational slide integration, and more scalable deployment. With continued development, this system, and others like it, could one day enable seamless voice-preserving communication across any language, helping bridge not only linguistic divides, but cultural and emotional ones as well.

References

[1] Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al. Seamlessm4t: Massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*, 2023.

[2] Hervé Bredin. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proceedings of Interspeech 2023*, pages 1983–1987, 2023.

[3] Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*, 2024.

[4] David Crystal. *Language Death*. Cambridge University Press, Cambridge, UK, 2000.

[5] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, Yossi Adi, and Morgane Riviere. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.

[6] Erik H. Erikson. Pseudospeciation in the nuclear age. *Political Psychology*, 6(2):213–217, 1985.

[7] Google Cloud. Cloud text-to-speech api, n.d. Accessed: April 17, 2025.

[8] Nick Haslam. Dehumanization: An integrative review. *Personality and Social Psychology Review*, 10(3):252–264, 2006.

[9] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

[10] Hirofumi Inaguma, Sravya Popuri, Ilia Kulikov, Peng-Jen Chen, Changhan Wang, Yu-An Chung, Yun Tang, Ann Lee, Shinji Watanabe, and Juan Pino. Unity: Two-pass direct speech-to-speech translation with discrete units. *arXiv preprint arXiv:2212.08055*, 2022.

[11] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *arXiv preprint arXiv:2010.05646*, 2020.

[12] Ann Lee, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Sravya Popuri, Xutai Ma, Adam Polyak, Yossi Adi, Qing He, Yun Tang, Juan Pino, and Wei-Ning Hsu. Direct speech-to-speech translation with discrete units. *arXiv preprint arXiv:2107.05604*, 2021.

[13] Meta AI. Seamlessexpressive and seamlessstreaming models for low-latency voice translation, 2023. Accessed: April 17, 2025.

[14] OpenAI. Chatgpt and gpt api for natural language tasks, 2023. Accessed: April 17, 2025.

[15] OpenAI. Whisper: Robust speech recognition via large-scale weak supervision, 2023. Accessed: April 17, 2025.

[16] Diana Sidtis and Jody Kreiman. In the beginning was the familiar voice: Personally familiar voices in the evolutionary and contemporary biology of communication. *Integrative Psychological and Behavioral Science*, 46(2):146–159, 2012.

[17] Dongchao Yang, Jinchuan Tian, Xu Tan, Rongjie Huang, Songxiang Liu, Xuankai Chang, Jiatong Shi, Sheng Zhao, Jiang Bian, Xixin Wu, Zhou Zhao, and Helen Meng. Uniaudio: An audio foundation model toward universal audio generation. *arXiv preprint arXiv:2310.00704*, 2023.