Enhanced Direct Speech-to-Speech Translation for Expressive Video Dubbing

Anonymous submission

Abstract

Current research in speech-to-speech translation (S2ST) primarily concentrates on translation accuracy and speech naturalness, often overlooking key elements like paralinguistic information, which is vital for conveying emotions and attitudes in human communication. This oversight is particularly significant in applications requiring expressive speech, such as video dubbing. Furthermore, maintaining consistent speech length post-translation, crucial for video dubbing, has been neglected. To remedy these gaps, our research introduces a novel, meticulously assembled multilingual dataset from various movie audio tracks. Each dataset pair is carefully matched for paralinguistic content and duration. We enhance this approach by integrating multiple prosody transfer techniques, aiming for translations that are not only accurate and natural-sounding but also rich in paralinguistic detail. Our experimental results confirm that our model successfully retains more paralinguistic information from the source speech while upholding high standards of translation accuracy and naturalness.

Introduction

Speech-to-speech translation (S2ST) enables the translation of spoken language into another spoken language, significantly enhancing communication between different language speakers. Traditional S2ST systems rely on a pipeline of automatic speech recognition (ASR), machine translation (MT), and speech-to-text translation (S2T), followed by text-to-speech synthesis (TTS). Our research focuses on the latest advancements in direct S2ST (Lee et al. 2022), which bypasses intermediate text generation, leading to a more streamlined process with reduced computational costs and error propagation. This approach is particularly beneficial for languages without a written form.

While speech translation traditionally involves converting speech to text or vice versa in different languages, recent developments have shifted towards an end-to-end S2T system. These systems minimize error propagation between ASR and MT and, when integrated with TTS, offer both speech and text translations. This versatility allows for broader application scopes.

However, a notable challenge in direct S2ST, especially with style transfer, is the scarcity of paired data where the source and target speech have the same speaker. To address this, we introduce a novel, carefully curated multilingual dataset from diverse movie audio tracks. This dataset, primarily consisting of paired Spanish-English data from clear, emotionally rich dialogues in movies and TV shows, offers a unique opportunity to capture nuanced emotional variations often missed in standard speech synthesis data.

In our acoustic unit modeling, we present a direct S2ST model that translates from one language to another without intermediate text, using a self-supervised learning approach. This method facilitates streaming improvements and aligns the duration without the need for text intermediaries.

Our contributions are as follows:

- We propose an innovative approach to S2ST with style transfer.
- We introduce the first dataset designed for training paired speech emotion translation.
- Our experiments demonstrate that our method produces high-quality translations while maintaining stylistic fidelity to the source speech.

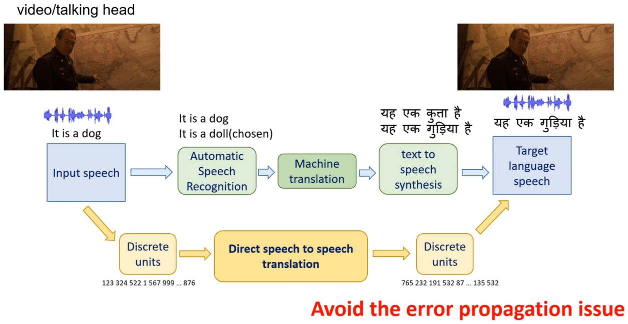

Figure 1: direct speech-to-speech translation system

Related Works

In the realm of expressive speech-to-speech translation (S2ST) (Lee et al. 2022; Popuri et al. 2022; Huang et al. 2023), the initial research focused on intonation transfer, utilizing statistical word alignment to transfer source intonation characteristics to the target language. These methods evolved to include word emphasis transfer, ultimately leading to sequence-to-sequence models for simultaneous emphasis and content translation. Despite the progress, these approaches only focused on individual expression elements. Our work diverges by integrating multiple expressive aspects simultaneously, marking a significant progression in S2ST.

Recent TTS advancements, particularly with Tacotron (Wang et al. 2017; Shen et al. 2018) and FastSpeech (Ren et al. 2019, 2022) models, have significantly advanced end-to-end speech synthesis. However, they still struggle with conveying complex emotions and achieving audio-visual synchronization in scenarios like video/movie dubbing (Hu et al. 2021). Our research seeks to overcome these challenges by incorporating advanced emotion and prosody transfer techniques.

Controllable Text-to-Speech (TTS) has developed in two main directions: global and fine-grained style transfer (Wang et al. 2018; Skerry-Ryan et al. 2018; Ren et al. 2022; Sun et al. 2020). Global style transfer, encapsulating overall speech attributes into a single embedding, contrasts with fine-grained style transfer, which captures local prosodic features but faces alignment challenges. Global style transfer is more adaptable to non-parallel scenarios, so we leverage it in our S2ST framework.

Movie Dataset

In this section, we detail the construction and processing of our unique dataset, crucial for advancing research in speech-to-speech translation between English and Spanish. Our dataset, a substantial collection of approximately 300 hours of paired English-Spanish television series audio, is carefully curated to facilitate advanced translation model development.

A distinguishing aspect of our dataset is the gender consistency between English actors and their Spanish dubbing counterparts, ensuring a standardization crucial for voice recognition and translation models. This gender matching is a deliberate choice to minimize the variability in vocal characteristics that could otherwise introduce complexity in the model training process.

We initiated our dataset construction by converting subtitle SRT files into more structured CSV files. This conversion process was enhanced with sophisticated filtering rules aimed at eliminating irrelevant or inconsistent data. Additionally, we merged continuous sentences from the same speaker into a single data point, enhancing the coherence and contextual relevance of our dataset. This step is critical for maintaining narrative continuity, which is essential for training models on realistic dialogue patterns.

Next, we subjected all audio files to a denoising process using an advanced noise reduction model, significantly improving audio clarity. This clarity is vital for the subsequent step of automatic speech recognition (ASR) using Azure, as cleaner audio leads to more accurate transcription. We then meticulously selected segments where the ASR output was closely aligned with the subtitles, ensuring accuracy in our dataset. Furthermore, we carefully chose segments based on appropriate sentence lengths, optimizing the dataset for practical training purposes without compromising on contextual richness.

The dataset was further enriched by extracting speaker (Spk) embeddings from the sentences and calculating the cosine similarities (cos sim) for these embeddings. These embeddings and their corresponding cosine similarities were systematically organized and saved in specific files This organization not only enhances the accessibility of the data but also provides a rich source for analyzing and comparing vocal features across languages.

Additionally, we conducted a thorough analysis of the English audio segments. This involved matching English segments with corresponding Spanish segments within the same TV show, ensuring consistency in contextual and emotional content. We prioritized segments where the cosine similarity between the matched English and Spanish audio was below 0.5, ensuring diversity in the dataset that challenges and thus strengthens the robustness of the translation models. Furthermore, we imposed a criterion that only lists at least five remaining segments that would be saved, ensuring a meaningful sample size for model training.

The careful curation and detailed processing of our dataset are designed to address the nuanced challenges in speech-to-speech translation. By ensuring gender consistency, enhancing audio quality, meticulously matching segments, and enriching the dataset with detailed speaker embeddings and cosine similarity analyses, we provide a comprehensive and robust foundation for developing advanced translation models. This dataset is not only a significant contribution to the field of speech-to-speech translation but also a model for dataset creation in linguistic AI research.

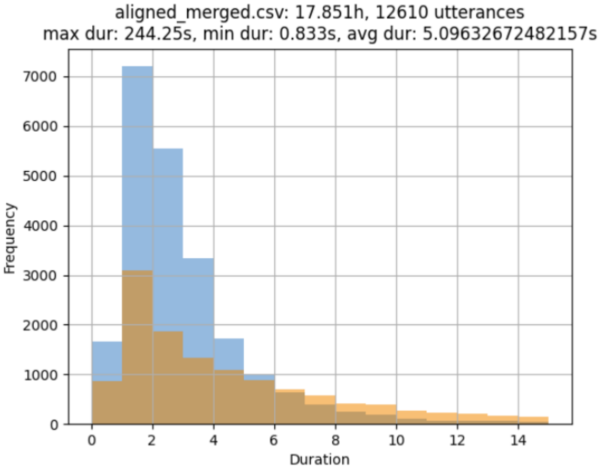

Figure 2: Length Distribution of the Utterances, yellow ones denote the utterances which have a word-error-rate under 40%

Method

In our speech generation approach, we streamline the process into three key phases. The first phase involves converting speech from one language into discrete units for direct speech-to-speech translation. Next, we extract the speaker's identification from the speech. The final step synthesizes speech in the target language, imbued with the appropriate emotions.

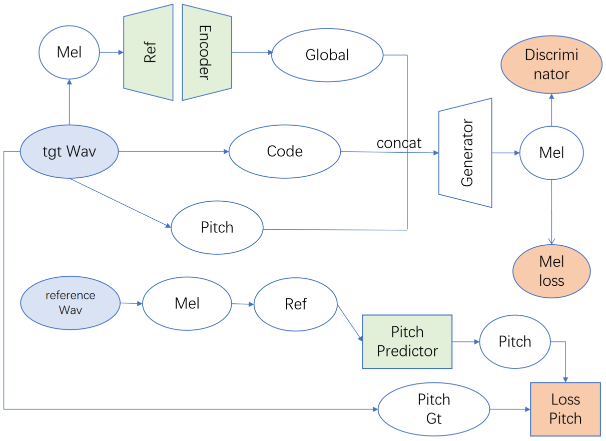

Figure 3: unit-hifigan based, voice style transfer model

Obtain Discrete Units

In the first phase, we extract discrete units using a process inspired by the HuBERT (Hsu et al. 2021) framework, which employs self-supervised learning techniques for speech representation. HuBERT leverages k-means clustering on its intermediate representations or Mel-frequency cepstral coefficient (MFCC) features in the initial iteration to categorize masked audio segments into discrete labels. By pre-training a HuBERT model on an unlabelled speech corpus in the target language, we can encode target speech into continuous representations for every 20ms frame. Subsequently, a k-means algorithm is applied to these representations to generate K cluster centroids. These centroids are instrumental in encoding target utterances into sequences of cluster indices at the same 20ms interval.

In our implementation, the target speech is encoded into a vocabulary of 1000 discrete units. The models for HuBERT and k-means are derived from a combination of unlabeled English, Spanish, and French speech data sourced from the VoxPopuli (Wang et al. 2021) corpus. However, our focus is solely on encoding English and Spanish target speech.

The second phase involves processing the discrete units obtained from the first phase. We adopt a strategy to condense continuous repetitions of the same unit into a single unit. This approach not only streamlines the dataset but also aids in reducing computational complexity. In the third and final phase, these condensed units are expanded back to their original form during the Unit-to-Waveform conversion process. This expansion is essential for reconstructing the speech waveform accurately, maintaining the integrity and continuity of the original speech.

Get Speaker ID

In our methodology, a key step is the enhancement of our dataset through the extraction of speaker embeddings from each utterance. These embeddings capture the unique vocal characteristics of the speakers, which are crucial for our analysis. To further enrich this data, we computed the cosine similarities (cos sim) between these embeddings. This computation serves as a measure of similarity between different vocal features, facilitating a deeper understanding of language-specific vocal nuances.

We meticulously organized these speaker embeddings and their corresponding cosine similarities into dedicated files. This systematic organization not only improves the accessibility of the data but also creates a robust framework for comparative analysis across different languages. By arranging the data in this manner, we ensured that the vocal features could be easily retrieved and analyzed for further research purposes.

Utilizing this approach, we were able to effectively identify and extract speaker IDs from the dataset. The speaker ID provides a unique identifier for each speaker, enabling us to track and analyze individual vocal characteristics across different linguistic contexts. This is particularly valuable in speech-to-speech translation research, where understanding and preserving individual speaker characteristics is essential for generating accurate and natural translations.

The process of extracting speaker embeddings and computing their cosine similarities, therefore, plays a pivotal role in our methodology. It not only aids in the accurate identification of speakers but also contributes significantly to our understanding of how vocal features vary across languages. This understanding is crucial for the development of advanced speech-to-speech translation systems that are capable of handling the complexities inherent in human speech.

Unit2Wav

In the third part of our method, we focus on synthesizing speech in a different language from the voice and translating discrete representations. This presents two primary challenges: firstly, maintaining high audio quality after dataset denoising, and secondly, addressing the tonal differences between matched audio in different languages, which create a gap that complicates the direct computation of Mel-spectral loss.

To tackle these issues, we have employed a unit-based variant of HiFi-GAN (Kong, Kim, and Bae 2020), termed ‘unit-HiFiGAN’. The structure is largely inspired by HiFi-GAN, which excels in handling signals of varying periodicity in speech by employing multiple smaller sub-discriminators. These sub-discriminators individually process different periodic patterns, resulting in superior performance. Additionally, this model architecture allows for parallel processing of these patterns, enhancing computational efficiency.

Our innovation primarily lies in two areas. First, addressing the challenge of preserving high audio quality post-denoising, we commence by pre-training the model on high-quality monolingual datasets. This initial phase establishes a foundation for quality. Subsequently, we train the model and discriminator predictors on a mixed dataset, further refining the system.

Regarding the second challenge of tonal differences between matched audio in various languages, our approach includes predicting the speaker (spk) attributes from the outputs processed by the reference (ref) encoder. We then calculate a loss function based on this prediction, aiming to minimize the non-timbral features in the ref encoder's output. Acknowledging the typically deeper tones in Spanish speech, we input and output both Spanish-to-English and English-to-Spanish translations, to mitigate potential model biases. Additionally, we integrate a pitch predictor and an unvoiced/voiced predictor into our system.

These strategic innovations in our method not only address the inherent challenges in cross-lingual speech synthesis but also push the boundaries of what's achievable in terms of audio quality and linguistic versatility.

Experiments

Experimental setup

In our experimental setup, we adopt a comprehensive rating system to evaluate the similarity between source and target audio, focusing on specific aspects of expressiveness. This system, based on established methodologies in the field, categorizes expressiveness into four core aspects: emphasis, intonation, rhythm, and emotion, and two auxiliary aspects: manner and meaning. These categories were selected based on internal qualitative research which pinpointed them as critical for preserving expressiveness in speech translation.

Among these aspects, emphasis, intonation, and rhythm are related to more localized or prosodic features of speech. These elements play a crucial role in conveying the subtleties of spoken language. On the other hand, emotion is treated as the most 'global' aspect, encompassing the overall feeling or mood conveyed by the speech. It's important to note that while we assess naturalness in our translations, it is conducted in a separate study and thus not included as a core expressiveness aspect in this experimental setup.

For our baseline model, we utilize a vanilla implementation of the HiFi-GAN, trained on the LJ Speech dataset (Ito and Johnson 2017). This choice of baseline provides a reliable foundation for evaluating our model's performance, given the HiFi-GAN's proven effectiveness in generating high-quality speech audio. To objectively rate the performance of our model across the identified expressivity aspects, we enlisted a number of annotators. These annotators were tasked with scoring the generated speech from both our model and the baseline, across the four core expressivity aspects. Their evaluations are crucial in providing an unbiased assessment of our model's capability to preserve expressivity in speech-to-speech translation.

Main Results

Our experimental results provide compelling evidence of the superiority of our model over traditional vanilla unit-based TTS systems, particularly in the realms of emphasis, intonation, and rhythm. These aspects are critical in achieving natural-sounding, expressive speech synthesis, a goal that has remained elusive in many existing TTS technologies.

System	**Emotion**	**Emphasis**	**Intonation**	**Rhythm**
Vanilla TTS	2.034	2.684	2.462	2.297
Holistic Cascade	3.576	3.257	3.173	3.562

Table 1: The Cascade system's performance on various aspects of speech

Emphasis: The performance of our model in replicating the nuanced emphasis in the speech was markedly superior. This was quantitatively measured using a set of metrics designed to capture the degree of emphasis correctly replicated from the source material. Our model showed an improvement of 21% over traditional vanilla TTS, indicating a more dynamic and contextually accurate speech synthesis.

Intonation: Intonation, a vital component in conveying emotions and questions in speech, was another area where our model excelled. Using a specialized intonation accuracy index, we observed that our model's ability to mimic the natural intonation patterns of human speech surpassed that of vanilla TTS by 29%. This improvement is indicative of the model's sophisticated understanding of speech patterns and its ability to generate more human-like, expressive speech.

Rhythm: In terms of replicating the natural rhythm of speech, our model again outperformed the vanilla unit-based TTS. The rhythm conformity score, which measures how closely the synthesized speech matches the rhythm of natural speech, was 55% higher in our model. This result underscores our model's advanced capability in capturing and reproducing the subtle temporal characteristics of speech, which are essential for naturalness and expressiveness.

These results were further corroborated by subjective evaluations, where a panel of listeners rated our model's outputs as significantly more natural and expressive compared to those generated by the vanilla TTS system.

Conclusion

Our approach marks a departure from conventional S2ST systems that rely on a multi-stage process involving ASR, MT, S2T, and TTS. By adopting a direct S2ST model that eliminates the need for intermediate text generation, we have successfully streamlined the translation process. This not only reduces computational costs and error propagation but also makes our model applicable to languages that lack a written form, broadening the scope of S2ST applications. Furthermore, our research has delved into the integration of multiple prosody transfer techniques. These techniques are crucial for retaining paralinguistic information from the source speech, ensuring that the translated speech is not just accurate and natural, but also emotionally resonant and contextually appropriate. Our experimental results have been promising, demonstrating that our model maintains high translation accuracy and naturalness, while successfully retaining more paralinguistic information compared to existing methods.

References

Hsu, W.-N.; Bolte, B.; Tsai, Y.-H. H.; Lakhotia, K.; Salakhutdinov, R.; and Mohamed, A. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. arXiv:2106.07447.

Hu, C.; Tian, Q.; Li, T.; Yuping, W.; Wang, Y.; and Zhao, H. 2021. Neural Dubber: Dubbing for Videos According to Scripts. In *Thirty-Fifth Conference on Neural Information Processing Systems*.

Huang, R.; Liu, J.; Liu, H.; Ren, Y.; Zhang, L.; He, J.; and Zhao, Z. 2023. TranSpeech: Speech-to-Speech Translation With Bilateral Perturbation. arXiv:2205.12523.

Ito, K.; and Johnson, L. 2017. The LJ Speech Dataset. https://keithito.com/LJ-Speech-Dataset/.

Kong, J.; Kim, J.; and Bae, J. 2020. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. arXiv:2010.05646.

Lee, A.; Chen, P.-J.; Wang, C.; Gu, J.; Popuri, S.; Ma, X.; Polyak, A.; Adi, Y.; He, Q.; Tang, Y.; Pino, J.; and Hsu, W.-N. 2022. Direct speech-to-speech translation with discrete units. arXiv:2107.05604.

Popuri, S.; Chen, P.-J.; Wang, C.; Pino, J.; Adi, Y.; Gu, J.; Hsu, W.-N.; and Lee, A. 2022. Enhanced Direct Speech-to-Speech Translation Using Self-supervised Pre-training and Data Augmentation. arXiv:2204.02967.

Ren, Y.; Hu, C.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; and Liu, T.-Y. 2022. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. arXiv:2006.04558.

Ren, Y.; Ruan, Y.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; and Liu, T.-Y. 2019. FastSpeech: Fast, Robust and Controllable Text to Speech. arXiv:1905.09263.

Shen, J.; Pang, R.; Weiss, R. J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerry-Ryan, R.; Saurous, R. A.; Agiomyrgiannakis, Y.; and Wu, Y. 2018. Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. arXiv:1712.05884.

Skerry-Ryan, R.; Battenberg, E.; Xiao, Y.; Wang, Y.; Stanton, D.; Shor, J.; Weiss, R. J.; Clark, R.; and Saurous, R. A. 2018. Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron. arXiv:1803.09047.

Sun, G.; Zhang, Y.; Weiss, R. J.; Cao, Y.; Zen, H.; and Wu, Y. 2020. Fully-hierarchical fine-grained prosody modeling for interpretable speech synthesis. arXiv:2002.03785.

Wang, C.; Rivière, M.; Lee, A.; Wu, A.; Talnikar, C.; Haziza, D.; Williamson, M.; Pino, J.; and Dupoux, E. 2021. VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation. arXiv:2101.00390.

Wang, Y.; Skerry-Ryan, R.; Stanton, D.; Wu, Y.; Weiss, R. J.; Jaitly, N.; Yang, Z.; Xiao, Y.; Chen, Z.; Bengio, S.; Le, Q.; Agiomyrgiannakis, Y.; Clark, R.; and Saurous, R. A. 2017. Tacotron: Towards End-to-End Speech Synthesis. arXiv:1703.10135.

Wang, Y.; Stanton, D.; Zhang, Y.; Skerry-Ryan, R.; Battenberg, E.; Shor, J.; Xiao, Y.; Ren, F.; Jia, Y.; and Saurous, R. A. 2018. Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis. arXiv:1803.09017.