Speech to Speech Translation for English and Hindi with Speaker Preservation

Dhruv Prasanna*, Avinash Nithyashree, Namith V Shetty, Praharsha Kosuri, Pavan A C

Computer Science Engineering Dept, PES University, 100 Feet Ring Road, 560085, Bangalore, Karnataka, India

* Corresponding author

doi: https://doi.org/10.21467/proceedings.178.7

ABSTRACT

This paper presents an advanced speech to speech translation system designed to facilitate accurate communication between English and Hindi speakers with near real time responses while preserving the original voice of the speaker. The system uses a cascaded architecture consisting of Automatic Speech Recognition (ASR), Machine Translation (MT), and Text to Speech (TTS) components. The resulting system is able to accurately translate between English speech and Hindi speech and vice versa. The techniques shown attempt to tackle the difficulties brought on by the different language structures and phonetic differences between Hindi and English by making use of transformer based models in each module. The presented system is capable of providing accurate translations and performs on par with state of the art models and services like Google Translate and ChatGPT. HuBERT, a speech representation model is utilized to perform voice cloning on the voice of the target speaker, this allows the system to preserve the speakers voice while translating which helps more effective communication. HuBERT enhances clarity and emotional realism in TTS by leveraging speaker specific attributes extracted from the original speech to synthesize the translated material in a similar voice to the original speaker.

Keywords: Transformers, HuBERT, Hindi, Translation, Voice cloning

1 Introduction

In an increasingly connected world, finding means of effective and accurate communication across language barriers is a significant challenge. This makes it increasingly more important to find solutions and techniques to foster efficient and effective cooperation and understanding between languages. Recent breakthroughs in speech processing and machine translation techniques as well as the emergence of voice cloning methods have helped improve solutions to address these challenges. Speech to speech translation systems are a promising emerging solution to allow for real time multilingual communication. This paper explores a comprehensive approach to speech to speech translation, specifically dealing with the Hindi English language pair. In addition, it also explores the preservation of the identity of the speaker through voice cloning which allows the system to imitate the vocal characteristics of the original speaker, which are leveraged while generating the output translation. A cascaded architecture is presented that incorporates a series of state of the art models to produce accurate natural sounding translations which resemble the original speaker. Integrating Automatic Speech Recognition (ASR), Machine Translation (MT), and Text to Speech (TTS) techniques in a cascaded pipeline to bridge the gap between Hindi and English conversations, while maintaining clarity, fluency, and emotional connection. One important aspect this paper presents is voice cloning, a method of mimicking human speech. This is achieved by a model that has been trained on the distinct vocal traits of a particular speaker. The goal is to imitate and capture the subtleties, intonations, and unique characteristics of a voice to generate an accurate replication. The ability to replicate the voice of speaker lifelike can improve user engagement and interaction. This makes this system able to personalize user experiences, which has applications in a variety of industries, including virtual assistants, entertainment, accessibility services and many more. Thus, the convergence of machine translation, voice cloning, and

Proceedings DOI: 10.21467/proceedings.178; Series: AIJR Proceedings; ISSN: 2582-3922; ISBN: 978-81-984081-8-1

speech recognition is an illustration of the various ways that developments in language processing technology are transforming a range of domains.

2 Related Work

There are generally two types of speech to speech translation systems; on one hand, there are end to end models [1], and on the other, there are systems that are a cascade of three different sub models [2]. Cascading systems consist of three main components: speech recognition, machine translation, and text to speech. The approach taken by the Translatotron models consists of a speech analyzer, a language translator, a speech synthesizer, and a special connection between them. The Translatotron models perform very well; they are able to achieve natural speech and good translation accuracy; they even perform close to more complex cascade systems. Recently, an end to end speech translation system that uses a single neural network was developed. But it requires a huge amount of high quality speech to speech parallel corpus. The proposed system follows the pipeline based approach, which does not require the speech to speech parallel corpus and involves connecting different components in a cascade to form the speech to speech translation pipeline. Cascade approaches tend to perform better, but are more complex and require each component to be trained separately. Another potential challenge of such an approach is that it may not be effective in preserving para linguistic and nonlinguistic information. This problem is tackled by making use of a custom HuBERT to extract nonlinguistic information about the speaker.

2.1 Speech Recognition and Machine Translation

The past two and a half decades have witnessed a significant increase in the amount of labeled training data availability for speech recognition across numerous languages. This proliferation of data has enabled and improved deep learning based approaches [3], enabling them to effectively leverage recorded and transcribed speech. Most of the recent speech translation research has focused on the speech to text setup. Research on ASR and MT systems looks into more effective approaches to integrate MT models with ASR output lattices in order to reduce the problem of error propagation between the two. Although the whisper models are capable of performing MT as well as ASR, it is not made use of for translation as it is unidirectional and translates only to English currently. In recent years, Neural Machine Translation (NMT) models have undergone a significant shift in architecture. Traditionally, NMT models used Recurrent Neural Networks (RNN) approaches like LSTM (Long Short Term Memory) for encoding and decoding sentences. More recent models make use of transformer based encoder decoder models [4]. Transformers have the advantage of being able to generate contextual representations, which allow them to outperform previous approaches, such a model is utilized for translating between English and Hindi.

2.2 Text To Speech and Voice Cloning

In the realm of voice cloning, most existing solutions are constrained by the issue of monolingualism. They are capable of cloning the voice of a person and generating an output within a single language that is native to the original speaker, but when it comes to adapting that voice across a different linguistic context they struggle and cannot consistently provide a coherent output. This limitation restricts the application of voice cloning in a wider multilingual setting. It also means it cannot be utilized in speech to speech translations, hindering its potential for cross cultural communication and accessibility. Also, extending the use of voice cloning for multilingual speech translation is largely unexplored; conventional speech to speech translation systems often disregard the vocal identity of the speaker, replacing it with a generic robotic sounding voice in the translated output. This disconnect leads to some emotional detachment and hinders effective communication. Although there have been many recent developments in the fields of speech to speech translation, current systems mainly deal with dominant languages like English. The system presented creates a robust solution for lower resourced languages like Hindi as well as English. Recently, TTS systems have adopted deep learning approaches [5], these models use powerful neural networks trained on very large speech datasets. By learning the complex relationship between text and audio features, they are able to

generate highly natural speech that closely mimics human prosody. While they are impressive, they do have limitations. These models often require large amounts of training data specific to a particular speaker or language and tend to struggle with unseen text. For unseen speakers, generative GPT styled models such as AudioLM [6] can produce syntactically and semantically realistic speech continuations while preserving speaker identity and prosody.

3 System Architecture

This paper proposes a state of the art multi lingual speech to speech translation system that seamlessly integrates many components, based on neural networks. The system architecture follows a cascade of three major components: an Automatic speech recognition (ASR) component, which transcribes the input speech and sets the language codes for downstream modules, a Machine Translation (MT) component, which provides translations in the output language, a Voice Cloning (VC) component, which creates a semantic representation of the speaker that the final Text To Speech (TTS) component can conditionally generate audio from that has the characteristics of the original speaker but in the new output language. As can be seen from the data flow diagram **Figure 1.** voice cloning and speech recognition are performed parallelly, this is because a text transcription is not required for the voice cloning module so it can be performed in tandem to reduce the total inference time of the system. The pipeline is focused on lowering inference times and being computationally lightweight and making some concessions to accommodate this.

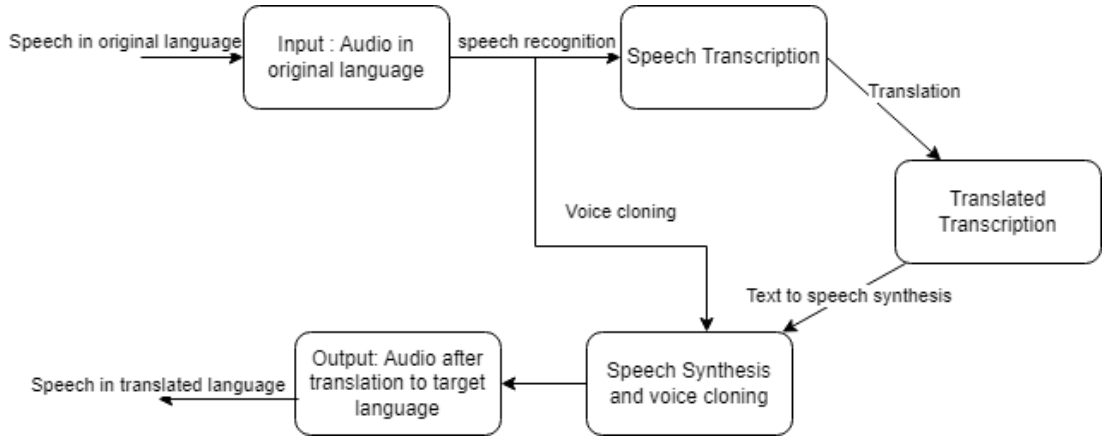

Figure 1: Data Flow Diagram

3.1 Automatic Speech Recognition and Language Detection

Automatic Speech Recognition involves transcribing text from an audio sample. The system uses Whisper [7] as a base model due to its robustness and good performance over many different languages. The robustness of the model allows it to produce accurate transcriptions given a wide range of accents and dialects, it also handles noisy scenarios better than most current models available. The inference times are also quick, these characteristics make it the most suitable for the system where a low inference time robust model is required. The model performs very well when given English speech as input, for Hindi speech although, the performance is not to the standard required and needs further training. The model is further trained on a couple of datasets containing Hindi speech text pairs to improve its performance for Hindi transcriptions. While the current performance of the model is very good, in a cascading pipeline of various models it is important that the data being fed to the other models is as accurate as possible. By fine tuning the model, it is possible to obtain a robust system capable of speech recognition over Hindi and English, performing well over various accents and dialects which is very important for a language like Hindi which is spoken by a diverse population. The model was trained and then fine tuned to reduce word level and character level error metrics for Hindi with the help of multiple datasets, including Hindi speech text pairs

Proceedings DOI: 10.21467/proceedings.178
ISBN: 978-81-984081-8-1

Series: AIJR Proceedings
ISSN: 2582-3922

from common voice, Google Fleurs, and IndicTTS. The audio files obtained for training were resampled to a uniform sampling rate of 16 kHz, so that it follows the expected input format of the Whisper model that has been chosen for the speech recognition and language detection tasks, avoiding potential compatibility issues during training. To increase the diversity of the training data and improve the ability of the model to generalize and perform better in real world scenarios where given audio samples have a lot of noise and are not clear, common data augmentation techniques were employed. The language that is being spoken in the audio file can be determined by the Whisper Language identification model. It uses the audio material to determine which language is most likely and returns the most probable language candidates. After verifying the probabilities are above a certain threshold, the system can assign the appropriate languages for downstream models.

3.2 Machine Translation

Machine translation (MT) has come a long way in recent years, and transformers have revolutionized the field and have become the main choice for most speech processing tasks, achieving state of the art results in translating text between languages. This paper attempts to enhance machine translation systems for languages with lower availability of training data and digital resources. It is important to increase the inclusivity and accessibility of machine translation technology for languages that do not have large amounts of resources and are currently underrepresented like Hindi. Making use of the No Language Left Behind model by Meta, tools, datasets and models are created and shared to speed up the creation of machine translation systems for these languages. NLLB has been utilized due to its transformer based architecture, transformers have become the prevalent architecture for machine translation due to their attention mechanisms. Their ability to generalize relationships between words and leverage attention mechanisms effectively has significantly improved their ability to understand context resulting in higher quality and fluency of the translated text. The NLLB model, trained on massive datasets of text spanning multiple languages, is one of the most capable neural machine translation models available. However, fine tuning the model on language specific datasets gives the model significant improvements in performance for particular translation tasks specifically the Hindi English language pair. The model was fine tuned using the IITB English Hindi dataset [8], improving model robustness and performance.

3.3 Text To Speech and Speaker Preservation

Bark, a GPT style text to audio model, is used for generating speech. It is a transformer based generative model made to produce high quality, humanlike speech from text without the use of phonemes. Without the restriction of using phonemes, Bark can be trained to generate speech in multiple languages. In contrast to earlier methods, the input text prompt is transformed straight to audio without utilizing phonemes in between. Consequently, it can generalize to arbitrary instructions that are not limited to speech, like lyrics in music, sound effects, or other nonspeech noises. Therefore, the same model can be used to generate speech in both English and Hindi, with the scope to further add more languages in the future. The models consist of 3 sub models and a final audio neural codec model: a semantic, coarse, fine, and an audio codec model. The semantic model is an auto regressive transformer that generates semantic tokens from tokenized text. The coarse and fine models iteratively predict codebooks which are decoded into an audio array using the audio codec model. The first three models can be conditioned using the speaker embedding produced by the voice cloning model to generate audio similar to the original speaker. The generative nature of this model makes it expensive to use; therefore, the small version of Bark has been opted for to lower inference times and generate translations quicker. The speaker preservation or voice cloning system consists of a custom quantizer HuBERT model [9]. The model is trained to learn mappings from audio files containing speech to semantic tokens. The semantic tokens generated from the model can be utilized to generate speech that mimics the voice of the original person for unseen speakers as well. This represents a significant advancement for multi speaker and multilingual TTS synthesis, showcasing adaptability to novel and unseen speakers, demonstrating flexibility and delivery of natural sounding and coherent results most of the time.

It is observed that the best performance from this setup of the custom HuBERT model and Bark model is achieved when input audio clips are around 10 seconds long, producing consistent and coherent speech that sounds very similar to the original speaker. Inputs shorter than this are found not to be as consistent; at times, the model fails to even generate coherent speech. It appears the model begins to hallucinate when there is not enough information given as input from the semantic representation, as when generating speech without a speaker embedding the same issues are not present.

4 Results and Discussion

The model starts by collecting audio recordings from the user using a microphone or taking audio files as input directly. It uses the Automatic Speech Recognition (ASR) module to process and transcribe the recorded audio. Next, the Voice Cloning (VC) model is provided with the sample to generate a speaker representation of the original speaker. The translated text is then transformed back into synthetic speech using the Text to Speech (TTS) model, guaranteeing that the output is in the language selected by the user (Hindi/English). Through its integration, the audio input is processed with low latency and near real time for most cases during the voice cloning and translation stages. Observed inference times are generally equal to the length of the input. It is observed that the model is able to generate natural sounding and coherent speech when given audio samples of length 10 to 20 seconds. For audio samples under 10 seconds in length, it is found that the text to speech model and voice cloning models struggle. The performance of the speech recognition model increases greatly after fine tuning performance with an 8.65 normalized WER for Hindi speech when tested on the Common Voice dataset by Mozilla. This allows the system to have robust performance for both languages. Overall, the translation system gives accurate translations on par with state of the art models and services, for chrF++ it measures 54.2 for English to Hindi and 62.5 for Hindi to English.

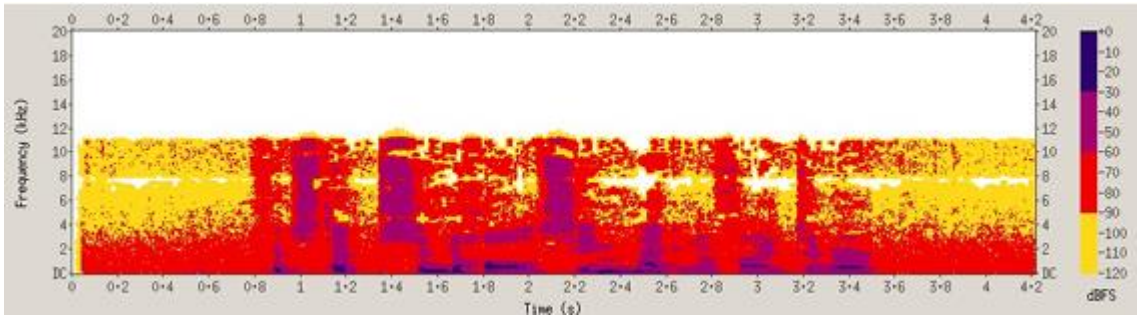

Figure 2: Spectrogram Of Human Speech Input

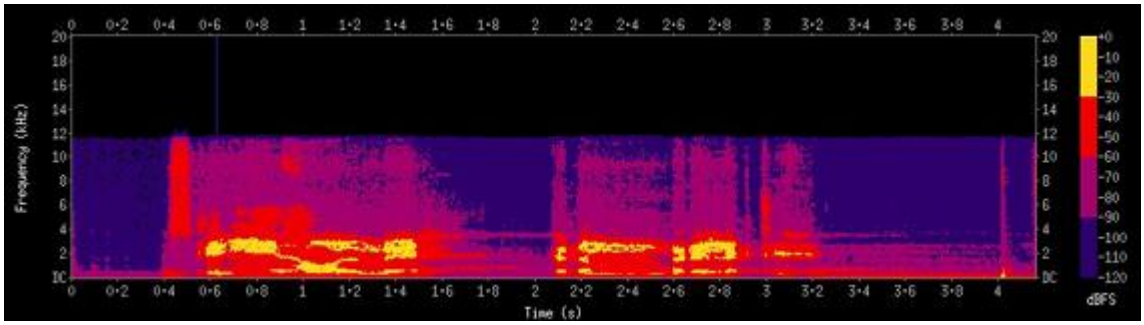

Figure 3: Spectrogram Of Model Generated Output

The range of frequencies of a signal is plotted visually in spectrograms such as **Figure 2** & **Figure 3**. The x and y axes denote time and frequency, respectively, and the colored part of the spectrogram indicates the intensity of the signal at each point in the time frequency plane. Comparing **Figure 2** which measures

Proceedings DOI: 10.21467/proceedings.178
ISBN: 978-81-984081-8-1

Series: AIJR Proceedings
ISSN: 2582-3922

human speech to the output generated by the model in **Figure 3**, it can be seen how well the model is able to generate speech that is consistent with real human speech with a similar amount of texture and variations.

5 Conclusion and Future Work

The speech to speech translation model has gone through many iterations and tested various models and architectures to enhance the synthesis of natural sounding and accurate translations. This model successfully created a system and machine learning pipeline capable of zero shot voice cloning and speech to speech translations between Hindi and English in near real time. The proposed system is able to produce natural sounding coherent speech that sounds very similar to the target speaker for audio clips averaging 10 seconds. The translation system provides a new and innovative way for users who do not share a common language to communicate. The continuous advancements in the field of voice cloning and translation underscore the dynamic nature of this paper, urging further exploration and integration of cutting edge technologies for an enhanced user experience and broader linguistic inclusivity. Despite its notable strengths, there are specific limitations and envision future directions for improvement. To extend its capabilities, possible additions may include increasing the impact of the system by adding support for more regional languages. Improving and optimizing the model to improve the inference times of the models, reducing overall system latency. Furthermore, addressing the situations in which the speech generation model can hallucinate to improve the performance of the model for shorter inputs.

6 Declarations

6.1 Competing Interests

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

6.2 Publisher's Note

AIJR remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

How to Cite

Dhruv Prasanna, Avinash Nithyashree, Namith V Shetty, Praharsha Kosuri, Pavan A C (2025). Speech to Speech Translation for English and Hindi with Speaker Preservation. *AIJR Proceedings*, 58-63. https://doi.org/10.21467/proceedings.178.7

References

[1] Y. Jia, M. T. Ramanovich, T. Remez, and R. Pomerantz, "Translatotron 2: High-quality direct speech-to-speech translation with voice preservation," proceedings.mlr.press, Jun. 28, 2022. https://proceedings.mlr.press/v162/jia22b.html (accessed May 07, 2023).

[2] S. Mhaskar, V. Bhat, A. Batheja, S. Deoghare, P. Choudhary, and P. Bhattacharyya, "VAKTA-SETU: a Speech-to-Speech machine translation service in select Indic languages," arXiv.org, May 21, 2023. https://arxiv.org/abs/2305.12518 (accessed Jan. 07, 2024)

[3] NLLB Team *et al.*, "No language left behind: Scaling Human-Centered Machine Translation," *arXiv.org*, Jul. 11, 2022. https://arxiv.org/abs/2207.04672 (accessed Jan. 10, 2024)

[4] J. Gala et al., "IndicTrans2: Towards High-Quality and Accessible Machine Translation Models for all 22 Scheduled Indian Languages," arXiv.org, May 25, 2023. https://arxiv.org/abs/2305.16307v3 (accessed Dec. 10, 2023)

[5] J. Shen et al., "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 2018, pp. 4779-4783, doi: 10.1109/ICASSP.2018.8461368.

[6] Z. Borsos et al., "AudioLM: A Language Modeling Approach to Audio Generation," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 2523-2533, 2023, doi: 10.1109/TASLP.2023.3288409.

[7] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust speech recognition via Large-Scale Weak Supervision," PMLR, Jul. 03, 2023. https://proceedings.mlr.press/v202/radford23a.html (accessed Dec. 10, 2023)

[8] A. Kunchukuttan, P. Mehta, P. Bhattacharyya. *The IIT Bombay English Hindi Parallel Corpus*. Language Resources and Evaluation Conference. 2018. [Dataset]. Available: https://arxiv.org/pdf/1710.02855.pdf. [Accessed: Nov 04, 2023].

[9] W. -N. Hsu, B. Bolte, Y. -H. H. Tsai, K. Lakhotia, R. Salakhutdinov and A. Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451-3460, 2021, doi: 10.1109/TASLP.2021.3122291.