StellarSpeech - Emotion based real-time voice translator

Amulya Thatha
computer Science
Pace University
Seidenberg School of Computer Science and Information Systems
NewYork, USA
AT65325N@pace.edu

Nikhileswar Dusanapudi
Computer Science
Pace University
Seidenberg School of Computer Science and Information Systems
NewYork, USA
ND06643N@pace.edu

Chandu Mandalapu
Computer Science
Pace University
Seidenberg School of Computer Science and Information Systems
NewYork, USA
CM58880N@pace.edu

Nandini Vadlamudi
Computer Science
Pace University
Seidenberg School of Computer Science and Information Systems
NewYork, USA
NV15014N@pace.edu

FNU Kaleemunnisa
Computer Science
Pace University
Seidenberg School of Computer Science and Information Systems
NewYork, USA
klnu@pace.edu

***Abstract*—The goal of this project is to improve the quality, emotional context, and fluency of local language voice translation in real time. Neural Machine Translation (NMT) and Automatic Speech Recognition (ASR), two deep learning innovations, are used in the technology to incorporate prosody-aware translation models that maintain regional dialects and emotive tones. In order to tackle issues like domain adaptation and noise resilience, the objective is to develop an adaptive system with minimal latency that can provide precise, natural-sounding translations. In order to improve the efficiency and usability of real-time speech translation for local languages, the project will investigate cutting-edge technologies like edge AI and integrate lightweight transformer structures.**

***Index Terms*—Speech-to-Text, Neural Machine Translation, Automatic Speech Recognition(ASR), Text-to-Speech**

I. Introduction

Deep learning-based real-time speech translation has made great strides, but issues still exist, particularly with regional languages. Early approaches struggled with emotional context and fluency. By giving local languages priority, this program seeks to enhance real-time voice translation while maintaining semantic correctness and emotional purity. In order to maintain tone and meaning, we provide an adaptive method that integrates emotion-aware models into Text-to-Speech (TTS), Automatic Speech Recognition (ASR),and Neural Machine Translation (NMT). The objective is to develop an effective, low-latency real-time translation system that can function even in settings with limited resources.

II. Ease of use

Our user-friendly real-time speech translation system makes it possible for non-technical people to communicate between languages with ease. Accurate real-time translations that preserve emotional tone and contextual meaning are produced by combining ASR, NMT,and TTS. The system is accessible and efficient due to its optimization for low-resource devices. Users can overcome language hurdles and communicate naturally with low latency and an intuitive user interface.

III. Literature Review

Real-time voice translation has advanced significantly in recent years, mostly as a result of advancements in deep learning technologies. Initially, the cornerstones of voice translation methods were rule-based approaches and statistical machine translation (SMT). For instance, IBM's early SMT algorithms laid a strong basis for translation studies, but they usually fell short in producing fluid translations and capturing context (Brown et al., 1993).

The introduction of neural machine translation (NMT) models significantly enhanced translation abilities. One particularly important innovation was the sequence-to-sequence (Seq2Seq) model incorporating attention processes proposed by Bahdanau et al. (2015). This model captures speech sequences with subtle links well. Additionally, modern transformer-based approaches like Google's BERT (Devlin et al., 2018) and T5 (Raffel et al., 2020) employ contextual embeddings and self-attention strategies, which greatly increase accuracy and linguistic coherence.

Furthermore, real-time translation systems have benefited from developments in automatic voice recognition (ASR) and text-to-speech (TTS) technologies. In integrated models like Google's Translatotron (Jia et al., 2019) and Meta's SeamlessM4T (Zheng et al., 2023), ASR, NMT, and TTS components can be readily coupled. This end-to-end integration enhances speech naturalness and system responsiveness, enabling more fluid real-time interactions.

However, maintaining emotional nuance in voice translation is still a difficult task. Despite having significant contextual meaning, older systems frequently ignore crucial features of speech, such as pitch, intonation, rhythm, and emotional tone. Emotion-rich speech embeddings and expressive TTS techniques are two new affective computing efforts that have started to tackle these issues. For instance, significant advancements in emotion-rich voice synthesis have been effectively demonstrated by Tacotron (Wang et al., 2017) and WaveNet (van den Oord et al., 2016).

Meanwhile, hybrid approaches that blend deep learning models with language rule-based methods have also attracted attention. These hybrid approaches aim to increase grammar accuracy and language fluency. However, real-time translation applications still face numerous obstacles, particularly in managing domain-specific terminology, processing delays, and noise. Finding a balance between speedy processing and accurate preservation of the speaker's prosodic and affective content is essential for effective translation.

Our study directly tackles these problems by developing emotion-sensitive translation models that consider both prosodic cues and emotional embeddings. With an emphasis on translations into English, German, Spanish, and French, our method ensures the emotional authenticity and semantic integrity of translated speech. We use transformer designs that are lightweight and optimized for real-time speed without compromising translation accuracy or speech quality.

For this study, we collected multilingual speech datasets from public archives like Mozilla's Common Voice, GlobalPhone, TESS, and RAVDESS. To further enhance language representation and emotional diversity, we also produced more synthetic datasets, expanding coverage to include more manifestations of emotional states and regional linguistic variations.

Improved low-latency translation algorithms, advanced prosodic modeling techniques, and better control of dialectal variance are promising directions for future research. Furthermore, there is a lot of promise for significantly reducing computational loads with cutting-edge technologies like edge computing and quantum computing, which might completely transform the real-time voice translation market.

IV. Methodology

A. Speech Processing Pipeline

Developing a well-structured and methodical pipeline to efficiently translate spoken audio into text and then convert the translated text back into speech that sounds natural is the research's methodology. The end result of this thorough procedure is an integrated multilingual system that may improve communication between speakers of different languages and bridge linguistic gaps by utilizing cutting-edge technology in speech recognition, natural language processing, and speech synthesis.

B. Speech-to-Text (Transcription)

First, pre-trained Wav2Vec2 models—which have been specially adjusted to handle the targeted languages like English, Hindi, and Telugu—are used to convert audio inputs into text. These sophisticated models can reliably and effectively transcribe speech because they use the Connectionist Temporal Classification (CTC) approach. By standardizing all audio samples to a 16 kHz sampling rate, audio preprocessing techniques are used to guarantee constant input quality. This greatly improves transcription accuracy and compatibility with other procedures.

C. Emotion Classification (Fine-Tuning):

The pretrained Wav2Vec2 base model (wav2vec2-base) is refined to create a customized emotion classification model that is used for emotion recognition. The training dataset comprises audio recordings in Telugu and English that have been categorized with various emotional categories. Preprocessing standardizes each audio sample at a sampling rate of 16 kHz, allowing for consistent data processing. A LabelEncoder transforms the category emotion labels into numerical values so that they can be used for computational research.

In order to extract detailed acoustic information, the processed audio is input into the Wav2Vec2 model during fine-tuning. An additional linear classification layer is then used to classify the audio into distinct emotions. The model was trained over multiple epochs using the Adam optimizer and CrossEntropy loss function to attain a strong prediction accuracy for emotions. For further use in real-time emotion recognition tasks, the completed model and related label encoder were stored.

D. Neural Machine Translation (Telugu-English, Hindi-English, English-Hindi):

Advanced sequence-to-sequence (Seq2Seq) neural machine translation models based on Long Short-Term Memory (LSTM) architectures were used to translate text between Telugu-English, Hindi-English, and English-Hindi language pairs. Large parallel corpora with thousands of multilingual sentence pairs were used to meticulously train these models. During text preparation, phrases were tokenized, special tokens (and) were added to enable precise decoding, and tokenized sequences were padded to guarantee consistent lengths be-

tween batches. Separate encoder and decoder modules, each combined with embedding layers and LSTM units, made up the Seq2Seq architecture. Input sentences were converted into meaningful latent representations by the encoder, and the decoder used these representations to produce precise translations that were appropriate for the situation. Categorical cross-entropy loss and the RMSprop optimizer were heavily used during the training process, and validation measures were used to carefully track the model's performance. For effective real-time inference, the optimal translation model weights were saved after completion.

E. Text-to-Speech (Synthesis):

In the last step, Google's Text-to-Speech (gTTS) API was used to turn translated texts back into speech, guaranteeing excellent audio output. By preserving naturalness and clarity, the synthesized speech improves the translated content's usability and accessibility. The generated speech files were organized and maintained for assessment, demonstration, or real-world use.

Workflow and System Architecture for Web Deployment and Pipeline Integration A unified web-based multilingual voice processing pipeline is created by the deployment architecture, which combines the different machine learning components. In order to coordinate the data flow through several processing stages while preserving operational consistency, this architecture uses a modular design.

Model Orchestration Structure

The interaction between several AI models inside the processing pipeline is coordinated by the system using a meticulously crafted orchestration framework Sequential Model Loading: To create a solid basis for later processes, the system initializes models in a predetermined order, such as neural machine translation models, emotion classification models, and language-specific speech transcription models.

Centralized Model Registry: All loaded models are kept up to date in a centralized model registry, which makes them accessible at all times during the program lifecycle and for effective resource sharing.

Fallback methods: Even in cases when certain models experience processing issues, the architecture's progressive fallback methods provide system stability by gracefully handling component failures.

Resource Optimization: Depending on the deployment environment, models are handled with consideration for the distribution of computational resources, maximizing processing throughput and memory use.

Processing Pipeline Algorithm:

The main processing algorithm employs a methodical approach intended to convert audio input via several phases of conversion and analysis:

Preprocessing audio: To guarantee constant quality for processing later on, raw audio is normalized and standardized to a 16 kHz sample rate. Language Identification: To identify the input language and route the audio to the proper transcription model, the system uses a script-based heuristic method that analyzes Unicode character distributions.

Parallel Processing Paths:

Audio $\rightarrow$ Emotion Categorization $\rightarrow$ Emotional Context Information Path B: Audio $\rightarrow$ Language-Specific Transcription $\rightarrow$ Text Output Translation Orchestration: The relevant neural machine translation model (English-Hindi, Hindi-English, or Telugu-English) is chosen and applied to the transcribed text based on the recognized source language and the intended target language.

Speech Synthesis: Text that has been translated is sent to text-to-speech services that are appropriate for the target language, producing audio output that sounds natural.

Integration of Results: A complete output package is created by combining translated speech with emotional context metadata.

Flow of Requests and Responses

A organized request-response flow is implemented by the web application to direct user interactions through the multilingual processing system: Handling Multimodal Input: The system offers a versatile entry point for processing by accepting audio uploads in a variety of formats. Processing Orchestration: The system starts a coordinated workflow as soon as input is received, processing the audio using the relevant models according to identified features. Incremental Result Generation: Throughout the pipeline, processing status and incremental results are monitored, allowing for clear user progress reporting.

Comprehensive answer Formation: All processing outputs, including transcription, emotion detection, translation, and synthesized voice, are combined into a single, coherent result package in the final answer.

Flexibility and Adaptability

Adaptive mechanisms are incorporated into the system design to increase its adaptability to various deployment conditions. Dynamic Model Selection: Using input properties, language pairings, and computational resources, the processing algorithm dynamically chooses suitable models. Pluggable Components: A standardized integration interface allows the system to incorporate new language models, emotion classifiers, or translation pairs.

Configurable Processing Paths: Depending on user preferences or computing limitations, the processing pipeline can be set up to omit specific parts.

Optimization of System Performance

Algorithmic optimizations are included in the deployment to improve user experience and performance: Parallel Processing: To reduce overall latency, processing stages are carried out in parallel when dependencies permit.

Adaptive resource allocation maximizes performance during periods of high usage by dynamically allocating computational resources according to processing demands.

Caching Strategy: To avoid duplicating processing for similar inputs, frequently accessible intermediate outcomes are purposefully cached. External Service Integration: To balance system load while preserving quality, resource-intensive tasks like voice synthesis are delegated to specialist external services when necessary.

This architectural strategy democratizes access to cutting-edge language technology for a variety of user groups by developing an integrated, flexible, and effective multilingual speech processing system that can be accessed via a web interface.

CONCLUSION

This study's research and execution mark a major breakthrough in multilingual speech processing technology, with a special emphasis on bridging Telugu, Hindi, and English communication gaps. This work addresses the growing need for seamless cross-language communication in increasingly globalized and digitally connected environments by strategically integrating state-of-the-art speech recognition, emotion classification, neural machine translation, and speech synthesis.

FUTURE WORK

Even though the current system provides multilingual voice translation with emotion identification in real time, there are still a number of interesting directions that could be pursued in the future.

First, improving language support beyond Hindi-English and Telugu-English couples is vital. Due to limited publicly available parallel corpora for many regional and minority languages, future research should focus on developing low-resource or unsupervised machine translation models and exploring techniques such as zero-shot learning or multilingual transformer architectures to improve coverage.

Second, the emotional fidelity of the output remains a challenge. Although the system can classify emotions in the input speech, the final synthesized speech does not yet reflect the same emotional tone. Future work could integrate emotion-conditioned neural text-to-speech models to preserve or even amplify emotional nuance across translations, enhancing the naturalness and expressiveness of the output. Additionally, improving system performance for low-latency and on-device deployment is an essential next step.

Optimizing model size through quantization, pruning, or efficient architectures (e.g., lightweight transformers) could allow the system to operate efficiently on edge devices, improving accessibility in bandwidth-limited environments.

Finally, incorporating user feedback mechanisms for continuous learning, boosting robustness to noisy inputs, and enhancing domain-specific adaption (such as medical or legal terms) are significant areas to examine. By addressing these future areas, the system can move closer to enabling fully seamless, emotionally sensitive, multilingual communication at scale.

REFERENCES

[1] Tokuda, K., Wester, M., Dines, J., Gibson, M., Liang, H., Wu, Y.-J., Saheer, L., King, S., Oura, K., Garner, P. N., Byrne, W., Guan, Y., Hirsimäki, T., Karhila, R., Kurimo, M., Shannon, M., Shiota, S., Tian, J., Yamagishi, J. (2013). Speaker adaptation and the evaluation of speaker similarity in the EMIME speech-to-speech translation project. Computer Speech Language, 27(2), 420–437.

[2] Dines, J., Liang, H., Saheer, L., Gibson, M., Byrne, W., Oura, K., Yamagishi, J., King, S., Wester, M., Hirsimäki, T., Karhila, R., Kurimo, M. (2013). Personalising speech-to-speech translation: Unsupervised cross-lingual speaker adaptation for HMM-based speech synthesis. Computer Speech Language, 27(2), 420–437.

[3] Sawada, K., Tokuda, K., King, S., Black, A. W. (2017). The Blizzard Machine Learning Challenge 2017. In 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (pp. 331–337). IEEE.

[4] Rugchatjaroen, A., Saychum, S., Oura, K., Tokuda, K. (2017). Generalization of Thai Tone Contour in HMM-Based Speech Synthesis. In 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) (pp. 1102–1105). IEEE.

[5] Sun, S., Luo, C., Chen, J. (2017). A review of natural language processing techniques for opinion mining systems. Information Fusion, 36, 10–25.

[6] Yadollahi, A., Shahraki, A. G., Zaiane, O. R. (2017). Current state of text sentiment analysis from opinion to emotion mining. ACM Computing Surveys, 50(2), 25.

[7] Pérez-Rosas, V., Mihalcea, R., Morency, L.-P. (2013). Multimodal sentiment analysis of Spanish online videos. IEEE Intelligent Systems, 28(3), 38–45.

[8] Poria, S., Cambria, E., Hussain, A., Huang, G.-B. (2015). Towards an intelligent framework for multimodal affective data analysis. Neural Networks, 63, 104–116.