

## **Comprehensive Survey on Kannada Language Speech to English Language Translation and Voice Cloning System**

**Sagar Kumar<sup>1</sup>, Sakib Ahamed<sup>2</sup>, Sanjana H. V.<sup>3</sup>, Farhan Khan K. A.<sup>4</sup>, Shilpa M. I.<sup>5</sup>**

<sup>1,2,3,4</sup>Student, Dept. of AIML, PESITM, Shimoga, Karnataka, India

<sup>5</sup>Assistant Professor, Dept. of AIML, PESITM, Shimoga, Karnataka, India

**Corresponding Author**

**E-Mail Id:** shilpami1829@pestrust.edu.in

### **ABSTRACT**

India is a culturally rich country with diverse languages, with over 22 official languages and countless dialects spoken across the country. However, this linguistic diversity often acts as a communication barrier, hindering interactions between individuals who speak different languages. To address this challenge and revolutionize communication, there is an increasing interest in using Artificial Intelligence (AI) for language translation. This research explores the application of AI in language translation, with a specific focus on converting local languages into a universal language. Two AI models, namely VALL-EX and ELLA-V, play an important role in this project. These models are trained on extensive multilingual speech data and are designed to overcome the communication gaps and achieve zero-shot cross-lingual speech synthesis. The proposed approach takes advantage of recent advances in text-to-speech synthesis. With the development of voice cloning techniques and synthesized speech quality approaching human equivalency, the industry has seen huge developments over the years. This research introduces a novel approach to address language barriers, proposing solutions with the help of VALL-EX. This AI models aim to create high-quality zero-shot cross-lingual voice synthesis using data gathered from large multilingual speech samples. By doing this, the study hopes to improve current communication breakdowns and support smooth information transfer across various linguistic contexts.

**Keywords:** Language translation, machine translation, VALL-E X, cross-lingual speech synthesis, language recognition, voice synthesis, voice adaption, voice cloning, T2T, S2S, S2T local language, universal language, Kannada to English

### **INTRODUCTION**

The term “Machine Translation” (MT) refers to computerized system for the translation with or without human assistance. Although the ideal goal of MT system is to produce high quality translation.

The translation or understanding of human being with meaning began in the year 1629, a person named Rene Descartes dreamed of a world where everyone could understand each other no matter what they spoke. Georges Artsrouni applied for the first patents for “Translating Machines” in

the mid 1930s, a dictionary used to translate from one language to other language. We are living in a century replete with discovery and innovation. A recent innovation in technology, IT, and communication has helped many individuals to innovate and improve the quality of their lives. It is important to note that these innovations have solved various language challenges. These innovations have addressed various language challenges, making communication between people who speak different languages but live in the same region much easier. Language transcription software

and universal language dictionaries now allow for the conversion of local languages into more universally understood ones. Businesses and societies have greatly benefited from language software's. Despite these advancements, we have yet to establish a standard universal language for the world. As a matter of fact, English language is the most widely used language for communication purposes. But there is a growing need for systems that can facilitate seamless communication across languages such as Kannada. This is where innovation such as Natural Language Processing (NLP), audio-to-text conversion, language translation, and voice cloning come into play.

In the evolving field of NLP, converting audio into text and translating it, advancements in voice cloning technology have opened new possibilities for more personalized speech synthesis. and these technologies enhance language accessibility by enabling noise-free audio conversion into text, translating the text into a desired language, and finally converting the translated text back into speech using personalized voice models.

## TYPES

### Rule-based Machine Translation

- **Transfer-based machine translation:** To translate between closely related languages, a technique referred to as shallow-transfer machine translation may be used.
- **Interlingual:** Interlingual MT is one instance of rule-based MT approaches. In this approach, the source language, i.e. the text to be translated, is transformed into an interlingual, i.e. source-/target-language independent representation. The target language is then generated out of the interlingua.

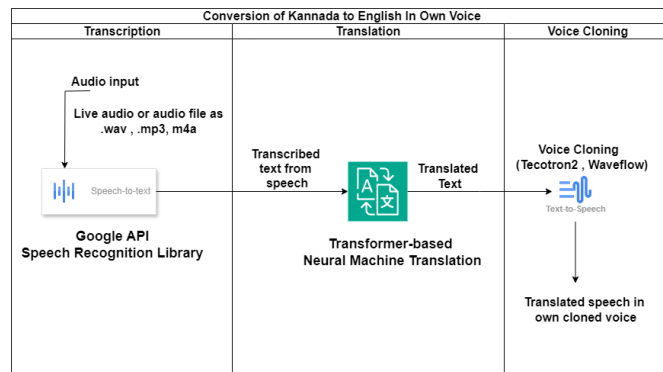
- **Dictionary-based:** MT can use a method based on dictionary entries, which means that the words will be translated as they are by a dictionary.

### Corpus Based Machine Translation

- **Statistical:** Statistical MT tries to generate translations using statistical methods based on bilingual text corpora
- **Example-based:** Example-based MT (EBMT) approach was proposed by Makoto Nagao in 1984. It is often characterized by its use of a bilingual corpus as its main knowledge base, at run-time. It is essentially a translation by analogy and can be viewed as an implementation of case-based reasoning approach of machine learning.
- **Hybrid MT:** Hybrid MT (HMT) leverages the strengths of statistical and rule-based translation methodologies. Several MT organizations (such as Asia Online, LinguaSys, Systran, etc.) claim a hybrid approach that uses both rules and statistics.

This research exemplifies Hybrid Machine Translation, which integrates the strengths of both rule-based and corpus-based approaches. By combining the systematic rules of language with statistical methods derived from extensive bilingual corpora, this study aims to enhance the accuracy and fluency of translations. The process begins with speech recognition, converting spoken Kannada into text, followed by translation into English. Finally, voice cloning technology is employed to generate a natural-sounding output that mimics the original speaker, creating a seamless and intuitive translation experience.

## ARCHITECTURE



**Fig. 1:** Proposed architecture for Kannada transcription, Translation and Voice Cloning.

### Transcription (Speech-to-Text)

**Input:** You provide an audio file (like .wav, .mp3, or .m4a) or even live audio as the starting point.

**Process:** The Google Speech Recognition API is used to convert spoken Kannada in the audio into text form. This means the system listens to the audio and outputs the text in Kannada (e.g., spoken words are transformed into written ones).

### Translation (Text-to-Text)

**Input:** The transcribed text from the first stage (in Kannada).

**Process:** A Transformer-based Neural Machine Translation (NMT) model is employed to convert the Kannada text into English. This is where advanced AI algorithms come into play to ensure accurate and context-aware translation.

**Output:** A grammatically correct and meaningful English translation of the Kannada text.

### Voice Cloning (Text-to-Speech)

**Input:** The translated text in English.

**Process:** You use a Text-to-Speech (TTS) synthesis system that includes models like Tacotron 2 or WaveFlow. These tools generate audio output in your own cloned voice, making it sound as if you're the one speaking the translated English text.

**Output:** The final result is a spoken English version of the input, but in your voice.

## ROBUST AUTOMATIC SPEECH RECOGNITION SYSTEM

Research shows that while these methods provide a measure of Automatic Speech Recognition System (ASR) performance for Kannada Speech Sentences in the Presence of Noise [1], they often underperform in real-world applications where background noise is present. (DNN) shows significant improvement in recognition accuracy, especially for complex acoustic problems. Integration of these technologies with traditional resources can improve performance metrics such as word error reduction (WER). Overall, the data suggest that continued innovation in extraction and distribution technologies is necessary to create effective, noise-free ASR across cultures and ultimately improve user experience and voice technology usage.

## ZERO-SHOT CROSS-LINGUAL SPEECH SYNTHESIS

In the domain of zero-shot cross-lingual speech synthesis, advancements are evident in models like VALL-E X [2] and ELLA-V [3], which extend the capabilities of cross-lingual speech synthesis. [2]–[5] These models leverage multi-model approaches to enable foreign language synthesis using one's own voice, inheriting robust context learning capabilities vital for zero-shot cross-

lingual text-to-speech synthesis and speech-to-speech translation tasks. VALL-E X utilizes deep learning abilities to execute high-quality zero-shot cross-lingual synthesis. The process involves obtaining linguistic spelling information from existing or pseudo-input automatic speech recognition (ASR) data, transforming it into sound sequences through legal modifications using tools like G2P conversion, and encoding speech information into sound tokens via an offline neural codec encoder. The landscape of natural language processing (NLP) has been significantly influenced by large-scale generative models such as GPT and DALL-E, revolutionizing capabilities in supporting META AI initiatives. Notably, Voice Box, a non-autoregressive flow-matching model, has been introduced to fill the gap in speech generation, particularly in inferring speech from audio context and text. Operating at scale, Voice Box [6] represents a versatile, text-conditioned speech generative model, having been trained on an extensive dataset comprising over 50,000 hours of unfiltered and unenhanced speech data. However, achieving high-quality zero-shot cross-lingual speech synthesis remains a persistent challenge, prompting the exploration of various methodologies and techniques. Achieving high-quality zero-shot cross-lingual speech synthesis (TTS) remains an ongoing challenge. Here are some methods:

### **Speaker Embedding**

Speaker Embedding [7] provides a unique way for language-independent or neutral representations. Which allow TTS models [6] to synthesize the unseen and unrecognizable languages while keeping the speaker characteristics the same.

### **Multilingual Speech Representation Learning:**

By recognizing multiple speech

representation across languages,[2]. Further speech recognition can be improved by utilizing GAN-based speech synthesis and contrastive unsupervised text selection, a method [3] that combines generative adversarial networks (GAN) [8] and multi-style training (MTR) [8] to enhance acoustic diversity in synthesized data. the models are able to generalize to unobservable languages. Showing improved performance in zero-shot cross-linguistic TTS tasks when models are trained on multilingual material [6].

### **Meta-learning Approaches**

The goal of a meta-learning algorithm is to investigate how fast it can learn from a limited amount of data in a new context. This makes them ideally suited for zero-shot TTS optimization. Studies such as [6], [9] are exploring this direction, showing promising preliminary results.

### **NOVEL EVALUATION METRICS FOR CROSS-LINGUAL TTS**

Speech synthesis models need to be evaluated comprehensively to ensure their effectiveness across diverse linguistic contexts and for evaluating the performance of cross-lingual TTS models requires metrics that evaluate beyond the word accuracy. Here are some following matrices:

#### **Mean Opinion Score (MOS)**

Mean Opinion Score [10] conduct subjective listening test where human listeners rate are naturalness, intelligibility and overall synthesized speech.

#### **Speaker Similarity Score (SSS)**

The SSS Measures [11], [12] the similarity between the synthesized speech and the original voice and metrics considers factors like pitch, timbre and speaking style.

#### **Language-Specific Intelligibility Test (LSIT)**

Developing language specific

intelligibility tests tailored to individual languages that can be provide a more accurate assessment of how well the synthesized speech is understood by native speakers. This approach [19] offers a more nuanced understanding of model performance across diverse languages.

### **CROSS-LINGUAL SPEECH SYNTHESIS SYSTEMS BE DEVELOPED FOR PRACTICAL APPLICATIONS**

Real time cross lingual speech synthesis systems holds immense potential for practical applications to breaking down the language barrier. Developing real-time cross-lingual TTS systems necessitates addressing several key aspects like model efficiency.

### **Cross-Lingual Neural Network Speech Synthesis Based on Multiple Embeddings**

Speech synthesis models [13] with neural network-based prediction of speech segment durations and acoustic properties are studied in this research, which covers models for both monolingual and multilingual settings. The paper illustrates the effectiveness of dynamic communication and conversation transformation strategies using the Merlin tools and Tensor-Flow framework. Moreover, neural vocoders such as WORLD and WaveRNN are used for speech comparison in order to assess speech synthesis outputs. The results emphasise how well the speaker can communicate meaning even when words are not delivered, demonstrating the value of the neural vocoder-supported communication strategy. Improved communication modes and more complex human-computer interactions are made possible by this research's contribution to the study and application of speech synthesis technologies in a variety of linguistic circumstances.

### **Practical Study of Deep Learning Models for Speech Synthesis**

Addressing voice cloning for speech synthesis using deep learning, emphasizing laptop usability and learning time. In order to demonstrate TTS models with low resources [14], it uses a range of learning methodologies to explore how learning for English and French language education has progressed from speaker identification to text-to-speech (SV2TTS) architecture. Among these is the recently launched Deep Reader smartphone app, which interprets text using deep learning algorithms. The speaker's voice is represented via embedding vectors created by Siamese networks. The evaluation and training of the SV2TTS architecture for speech synthesis have improved, according to the results. It also covers how DeepReader can help language learners or the illiterate.

### **Multilingual Emotional Text-to-Speech by Cross-speaker and Cross-lingual Emotion Transfer**

METTS (Multilingual Emotional Text-to-Speech by Cross-speaker and Cross-lingual Emotion Transfer) [15] approach was evaluated to be superior among other techniques in terms of its high degree of efficiency, tight semantics, and emotional proximity that specifically contribute to the emotional nature of communication. Reorganizing the perceptual point might appear harder because voice leakage, as well as multi sensory perception, may result in unnecessary bias. METTS is based on cognitive processes to build speaking abilities and enhance emotional intelligence. To evaluate METTS's efficacy, a sentiment matcher module is used to generate labels through content analysis from sentiment-free data embedding. Moreover, this new approach illustrates how the model supports emotional responsiveness by learning and adapting seamlessly in human interaction contexts – even in the absence of written



language cues that other models have historically focused on.

### **TACOTRON2 AND WAVEGLOW**

In the research paper [17], textual data analysis addresses the drawback of traditional methods in natural language generation, while demonstrating the evolution of text-to-speech (TTS) by moving from sequential methods to statistical parametric synthesis. It demonstrates the impact of deep learning, particularly Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN), which improve the quality of audio connections. The review discusses the introduction of WaveNet, a deep noise model that creates a new model for TTS but is an expensive implementation, and the development of Tacotron, a network model tool that generates mel spectrograms of text-to-speech transformations. It also identified challenges such as the need for large datasets and the difficulty of capturing sounds and emotions, leading to the creation of models such as Tacotron 2 to solve the problem. This problem lays the foundation for progress in TTS synthesis. After the conversion of mel-spectral by the Tacotron which preserve the prosody, specific feature then with the help of Waveglow it generate the waveform and the output is as cloned voice similar to given input voice.

### **CONTRIBUTION TOWARDS DATA CENTRIC RESEARCH**

Data plays a crucial role in research particularly in areas like language translation that research contributes towards the data centric research

### **Develop own Datasets with Proper Ethical Considerations**

If expertise and resources allow, consider creating your own datasets specifically tailored to under-represented languages or domains. This process requires careful

planning and adherence to ethical guidelines regarding data collection, consent, and privacy.

### **Data Quality and Labeling**

Make advantage of active learning strategies, in which the system gives top priority to gathering user input points that will enhance model performance. This can support maintaining the accuracy and applicability of the data gathered [21]. Permit users to comment on the accuracy of the translation and make correction suggestions. The translation models can be improved and the data quality raised with the help of these comments. Investigate semi-supervised learning [22] strategies that effectively enhance model performance by utilizing both labeled and unlabeled data (user voice input without explicit translation

### **No Language Left Behind: Scaling Human-Centered Machine Translation**

A multiple Architectural and training of the task using the human translated Benchmark as Human Translated Datasets [23]. FLORES 200: Evaluation datasets of 204 languages, NLLB-SEED: Seed training data contains 39 languages, NLLB-MD: Seed training data which contains 6 languages, Toxicity 200: Word list detect toxicity in 200 languages.

### **ADDRESSING CULTURAL AND LINGUISTIC DIVERSITY IN MACHINE TRANSLATION**

Addressing the challenges of cultural and linguistic diversity in machine translation to ensure accurate and contextually relevant translations and Ensuring the accurate and culturally appropriate translations which help out to from the challenges like Domain adaptations, MT for the Low resources (local) languages.

### **The Challenges of Teaching and Assessing Technical Translation in an**

### **Era of Neural Machine Translation**

In order to improve speech synthesis through deep learning- driven voice cloning, the article investigates laptop accessi- bility and faster learning times. It facilitates the shift from speaker identification to text-to-speech (SV2TTS) architecture[16] in the context of teaching French and English. A variety of learning strategies are examined to illustrate TTS models in resource-constrained environments. Notable is the novel audiobook software Deep Reader, which makes advantage of deep learning to make reading easier. In order to rep- resent speaker voice, Siamese networks' embedding vectors are necessary. The results demonstrate enhanced training and evaluation of the SV2TTS architecture, with implications for helping illiteracy or language learners through DeepReader.

### **Neural Machine Translation for Low-Resource Languages**

Neural Machine Translation (NMT) [18] with respect to Low-Resource Languages (LRLs) and offers suggestions for improving NMT in this field. It looks at a number of methods as possible ways to get better, such as back-translation, transfer learning, multi-NMT models, and unsupervised NMT. Specifically, it emphasizes how important it is to have accessible datasets and how adding languages to common issues like WMT helps further NMT research. The paper also emphasises how developments in sentence ranking algorithms and multilingual embedding generation are important fields driving progress in the discipline. The paper advances NMT capabilities and accessibility, promoting greater inclusion and linguistic variety in the field of machine translation by addressing the special opportunities and problems posed by LRLs.

### **A Novel Approach to Machine Translation: Example- Based System**

To enhance the translation process, speech conventions are included into example based machine translation (EBMT) systems [20] tailored for Indian languages like Marathi. To guarantee correct translation, particularly for complicated languages like Marathi, it makes use of bilingual corpora and n-gram based text classification. Developing cutting-edge teaching resources and resolving issues with language processing that have never been encountered before are challenges. To improve machine translation for communication disparities and boost the efficacy of EBMT, more research leaders are invited. EBMT will effectively promote efficient communication between languages in India and outside by resolving these problems and expediting the translation process.

### **PROPOSED METHODOLOGY**

#### **NLP Stages for Preprocessing and Transcription**

This research aims to achieve three objectives, of which the first is transcribing spoken kannada language to written text since transcribing is the initial necessary procedure for translating. To this end, we employ speech recognition solutions and feature modern solutions like Google Cloud Speech-to- Text. This API tool and Library are meant to provide solutions for transcribing audio to text with the precision needed for more natural speech.

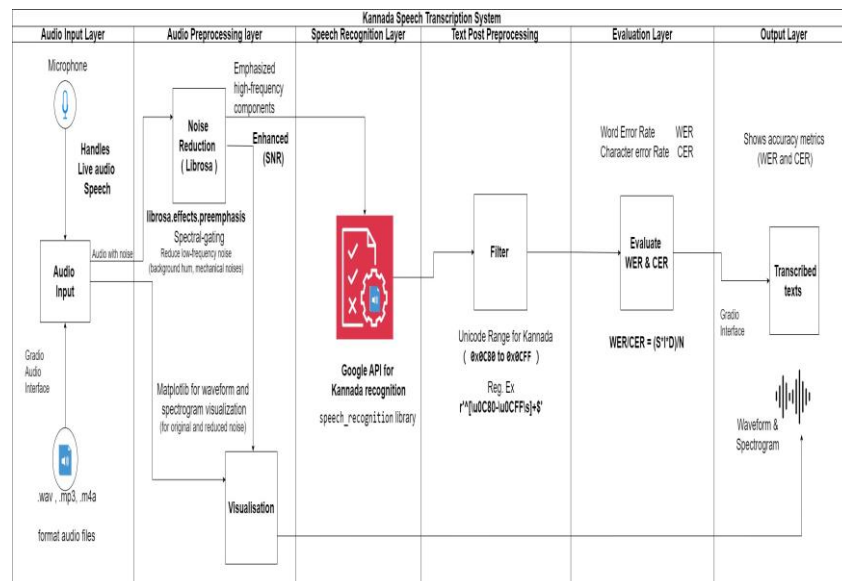
Before proceeding to the transcription, we pre-process the audio data using methods meant to remove noise. This means with the help of librosa effects pre-emphasis function that enhances high frequencies and suppresses noise at low frequencies. Thus, high frequency components are amplified and background interferences, such as hums or environmental noises are attenuated, as well as the noise of the speech signal is

improved. By passing the audio through this pre- emphasis filter we can make the speech recognition models work properly.

After that, with the help of filtering the audio, we pass to the step of converting speech to text. This way, with the help of Google Cloud Speech-to-Text we can leverage cloud computing to transcribe the

audio quickly.

After doing the text post preprocessing, it uses the Unicode Range for Kannada (0x0C80 to 0x0CFF). and Regular expression for same as a filter for Kannada transcribe from speech to text



**Fig. 2: Proposed Architecture for Kannada transcription.**

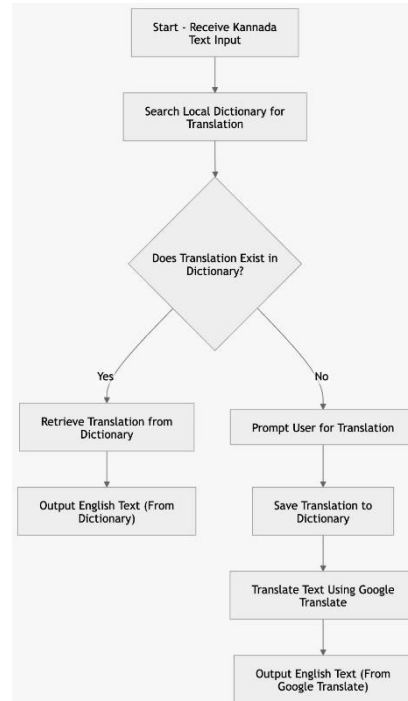
Ultimately, this speech-to-text conversion not only serves as a foundation for our translation work but also sets the stage for subsequent processing, including the integration of Tacatron and Waveglow for text-to-speech synthesis. This pipeline aims to create a seamless flow from spoken language to written text, paving the way for accurate and natural-sounding translations.

### Translation of Kannada to English Language

The major purpose of this research is to overcome the obstacles to translation from Kannada speech and dialect into

English with considering the regional and contextual differences. To obtain accurate as well as reasonable results the study incorporates various sophisticated technologies such as speech recognition systems, the machine translation technique, and the regional dialect analysis. In the case of transcribe the Google cloud speech to text API is used which gives accurate and scalable means of transcribing Kannada speech. accuracy of transcription, and translation, a dictionary containing more than 2000 words and phrases particular to the regions is incorporated in the system, which covers Mysuru Kannada, Bengaluru Kannada, and Mangaluru Kannada.





**Fig. 3:** *Proposed Architecture for Kannada Translation with Regional Context.*

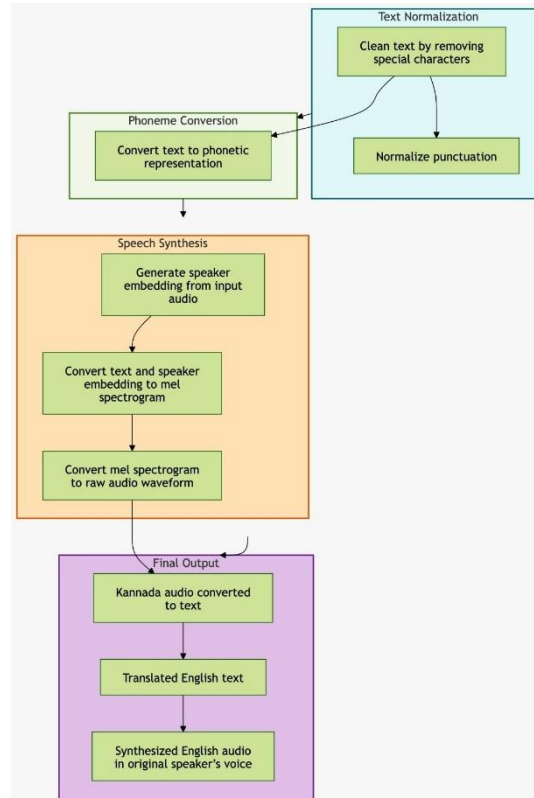
The dictionary was designed x carefully by gathering inputs from literature, media, and from the community in order to maintain the regional and cultural sensitivity. This dictionary is then applied on top of the Google Translator API outputs to make machine translations more relevant by district. System architecture is based on Python and Streamlit that provides convenient input form for text, region selection and visualization of translations. Tokenization is integrated into the backend then it will compare it with the custom dictionary, and other features such as APIs to ensure the translation process is smooth.

Moreover, the application is deployed on the cloud environment, and it uses the microservice patterns to split the application into single functioning units such as the front end, the back end and the dictionary service. CI/CD process provides efficient ways of updates, and steady system which does not fail all the time. Both training and verification are carried

out with real-life data, and the results indicate that the system achieves up to 90% of specific dialects' precision. In addition to facilitating precise translation, there are several multifaceted benefits: the preservation of the language's cultural heritage, better understanding of the culture by other nations, and design of more effective technologies for use by the Kannada-speaking population.

### **Real time Voice Cloning System for Kannada-English Translation**

This work concentrates on designing a voice cloning system for real time in Kannada-English translation module. It uses the current state of the art of neural networks to maintain the speakers voice identity while for English output it is fluent and easily understandable. Unlike prior work where voice synthesis and translation are performed independently, this model unifies both tasks, so that the tone and emotion of the speaker is preserved in the resulting translated text.



**Fig. 4:** Proposed Architecture for Real time Voice Cloning System.

The system architecture comprises three key components: They include a Speaker Encoder [24], a Synthesizer [25], and a Vocoder [26]. The Speaker Encoder, which has the ResNet architecture, extracts speaker-specific features, which will create an embedding of 256 dimensions. The Synthesizer, designed with Tacotron 2 as its backbone, produces mel spectrograms from the input text, guaranteeing that it matches the speaker's voice. Last but not least, by means of WaveNet, Vocoder produces spectrograms transformed into high quality raw audio waveforms.

Training is dedicated and requires preparation of data, rich as Kannada speech selected along with its English translation. Speech data is resampled to a standard rate of 16kHz, and filter banks for generating mel spectrograms are derived for model inputs. The text data is preprocessed to normalize, translate and convert into phoneme form in order to synthesize right. The training pipeline

incorporates the stages of speaker encoding, text-to-speech synthesis, and vocoder and its improvement to achieve durable and realistic sound.

The quantitative analysis shows that it has a very high value of speaker similarity (average cosine similarity score of 0.92), high level of translation accuracy (85% as identified by the BLEU score), as well as highly positive evaluation from users on the quality of audio provided (Mean Opinion Score of 4.3/5). The current qualitative analysis reproduces the system's capacity of preserving the emotional tone and applying prosody in a smooth and automatic manner to English. However, some of them are; dealing with noise in input audio, computational cost for real-time application and phoneme mismatch due to different languages.

This research will be highly relevant to universal and personalized AI and communication, which contributes to cross

lingual voice cloning and translation. Further work shall seek to increase the database with varying accent, further augment the noise suppression, and finally adapt the system for smart mobile devices and edge computing platforms.

### ACKNOWLEDGMENT

We would like to express our sincere gratitude to the following individuals and resources for their invaluable contributions to the authors of the referenced research papers, The research community. We extend our heartfelt thanks to our guide Ms. Shilpa M I, Assistant Professor and Dr. Likewin Thomas, Head of Department and Associate Professor, Department of AIML, PESITM, for his guidance and support throughout the project. Special thanks to Mrs. Poornima B.P., Assistant Professor and Project Coordinator, for her valuable coordination and insights during the research. We are also grateful to all the other faculty members of the Department of AIML, PESITM, for their encouragement and assistance. We are also grateful to every individuals who have provided the support or encouragement during the research and writing process we acknowledge the limitations of this report. It is not an exhaustive review of the field and does not claim to present all aspects of the topic. It serves as a starting point for further exploration and encourages the AI based machine Translation.

### REFERENCES

1. Swamy, M. (2022). Robust automatic speech recognition system for Kannada speech sentences in the presence of noise.
2. Zhang, Z., et al. (2023). Speak foreign languages with your own voice: Cross-lingual neural codec language modeling. *arXiv preprint arXiv:2303.03926*.
3. Song, Y., Chen, Z., Wang, X., Ma, Z., & Chen, X. (2024). ELLA-V: Stable neural codec language modeling with alignment-guided sequence reordering. *arXiv preprint arXiv:2401.07333*.
4. Kain, A., & Macon, M. (1998). Personalizing a speech synthesizer by voice adaptation. In *The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*.
5. Xin, D., Saito, Y., Takamichi, S., Koriyama, T., & Saruwatari, H. (2021). Cross-lingual speaker adaptation using domain adaptation and speaker consistency loss for text-to-speech synthesis. In *Interspeech* (pp. 1614-1618).
6. Le, M., Vyas, A., Shi, B., Karrer, B., Sari, L., Moritz, R., ... & Hsu, W. N. (2024). Voicebox: Text-guided multilingual universal speech generation at scale. *Advances in Neural Information Processing Systems*, 36.
7. Sun, J., Chen, H., Tian, J., & Xie, L. (2022). Speaker embedding for cross-lingual speech synthesis. *arXiv preprint arXiv:2204.09042*.
8. Chen, Z., Rosenberg, A., Zhang, Y., Wang, G., Ramabhadran, B., & Moreno, P. J. (2020, October). Improving speech recognition using GAN-based speech synthesis and contrastive unspoken text selection. In *Interspeech* (pp. 556-560).
9. Baevski, A., Srinivasan, A., Shankar, S., Bengio, Y., & Pineau, J. (2022). Meta-learning for zero-shot cross-lingual speech synthesis. *arXiv preprint arXiv:2206.05150*.
10. Nguyen, H., Li, K., & Unoki, M. (2022). Automatic mean opinion score estimation with temporal modulation features on gammatone filterbank for speech assessment. In *INTERSPEECH* (pp. 4526-4530).
11. Dehak, N., Dehak, R., Glass, J. R., Reynolds, D. A., & Kenny, P. (2010, June). Cosine similarity scoring without score normalization

- techniques. In *Odyssey* (p. 15).
12. Kamil, D., Ariadna, S., Julian, R., & Marius, C. (2022). Automatic evaluation of speaker similarity. *arXiv preprint arXiv:2207.00344*.
  13. Nosek, T. V., Suzic, S. B., Pekar, D. J., Obradovic, R. J., Secujski, M. S., & Delic, V. D. (2021). Cross-lingual neural network speech synthesis based on multiple embeddings.
  14. Langlois, Q., & Jodogne, S. (2023, July). Practical study of deep learning models for speech synthesis. In *Proceedings of the 16th International Conference on Pervasive Technologies Related to Assistive Environments* (pp. 700-706).
  15. Zhu, X., Lei, Y., Li, T., Zhang, Y., Zhou, H., Lu, H., & Xie, L. (2024). METTS: Multilingual emotional text-to-speech by cross-speaker and cross-lingual emotion transfer. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.
  16. Tavares, C., Tallone, L., Oliveira, L., & Ribeiro, S. (2023). The challenges of teaching and assessing technical translation in an era of neural machine translation. *Education Sciences*, 13(6), 541.
  17. Shen, J., et al. (2018). Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
  18. Ostling, R., & Tiedemann, J. (2017). Neural machine translation for low-resource languages. *arXiv preprint arXiv:1708.05729*.
  19. Van Esch, T. E. M., & Dreschler, W. A. (2015). Relations between the intelligibility of speech in noise and psychophysical measures of hearing measured in four languages using the auditory profile test battery. *Trends in Hearing*, 19, 2331216515618902.
  20. Rane, A. L., Kangune, S. P., Sahastrabuddhe, K. N., & Patil, P. P. (2019). A novel approach to machine translation: Example-based systems.
  21. Sheng, V. S., Provost, F., & Ipeirotis, P. G. (2008, August). Get another label? Improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 614-622).
  22. Zhang, Z., Han, J., Deng, J., Xu, X., Ringeval, F., & Schuller, B. (2018). Leveraging unlabeled data for emotion recognition with enhanced collaborative semi-supervised learning. *IEEE Access*, 6, 22196-22209.
  23. Costa-jussà, M. R., Cross, J., Celebi, O., Elbayad, M., Heafield, K., Heffernan, K., & NLLB Team. (2022). No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.
  24. Wan, L., et al. (2018). Generalized end-to-end loss for speaker verification. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
  25. Wang, Y., et al. (2017). Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*.
  26. Kalchbrenner, N., et al. (2018). Efficient neural audio synthesis. *International Conference on Machine Learning*. PMLR.