Journal of Information Systems Engineering and Management
2024, 9 (4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

A Traditional Machine Learning Framework for Kannada Text-to-Speech Synthesis

Ganga Gudi[1], Mallamma V Reddy[2], Hanumanthappa M[3]

[1,2]*Department of Computer Science, Rani Channamma University, Belagavi, India*

[3]*Department of Computer Science, Bangalore University, India*

[1]*Department of Computer Applications, PES University, RR Campus, Bangalore, India*

Email: gangagudi1@gmail.com, mvreddy@rcub.ac.in, hanu6572@bub.ernet.in, gangagudi@pes.edu

ARTICLE INFO

Received: 05 Sep 2024

Accepted: 26 Dec 2024

ABSTRACT

Speech synthesis plays a pivotal role in assistive technology, particularly for individuals with visual impairments and literacy challenges. This research presents a machine learning-based text-to-speech (TTS) system designed specifically for the Kannada language, aiming to generate clear, intelligible, and natural-sounding speech without relying on complex deep learning architectures. The proposed system leverages classical machine learning models, including Decision Tree Regression and K-Nearest Neighbors, to predict prosody features such as pitch and duration for each phoneme extracted from Kannada text. These predicted parameters guide a concatenative synthesis engine, which assembles pre-recorded phoneme audio segments to produce fluent speech. The system architecture emphasizes interpretability, efficiency, and low computational overhead, making it suitable for deployment in low-resource environments such as embedded devices or regional mobile applications. Experimental evaluations confirm that the model can generate understandable Kannada speech while maintaining flexibility and ease of training, demonstrating the potential of traditional machine learning in regional language speech generation.

Keyword: Kannada Text-to-Braille, Machine Learning, Decision Tree Regressor, K-Nearest Neighbors (KNN), Braille Translation, Visually Impaired.

I. INTRODUCTION

Speech is one of the most natural and effective modes of human communication. Text-to-speech (TTS) systems have become a vital component of human-computer interaction, especially benefiting the visually impaired, elderly, and those with reading difficulties. These systems enable users to listen to written content in their native language, thereby expanding access to digital information and services. Although significant advances have been made in speech synthesis for widely spoken languages such as English and Mandarin, regional languages like Kannada remain underrepresented. Kannada, a Dravidian language spoken by millions primarily in Karnataka, poses unique challenges for speech synthesis due to its complex phonetic structure, rich morphology, and limited availability of labeled data.

This research proposes a lightweight machine learning-based TTS system tailored for Kannada. It employs classical machine learning algorithms such as Decision Tree Regression and K-Nearest Neighbors to predict essential speech parameters like pitch and duration from phoneme sequences [1]. These predicted features guide a concatenative synthesis framework that generates natural-sounding Kannada speech using a database of pre-recorded phoneme segments. The focus of this work is to provide a cost-effective, interpretable, and computationally efficient speech synthesis solution that ensures high intelligibility and fluency, suitable for real-time or offline deployment in assistive technologies, educational tools, and localized applications.

This paper is organized as follows: Section II presents a review of related work; Section III details the methodology and system architecture; Section IV discusses experimental results and model comparison; and Section V concludes with future directions for improving accessibility technologies for regional languages.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9 (4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

II. LITERTAURE REVIEW

Sighted individuals interact with the digital world using various input and output devices. Visually impaired users depend on specially designed tools or software to access and engage with digital content. Braille serves as a primary mode of communication for many of them. Speech synthesis offers an additional means by converting written text into spoken words, enabling easier access to information. This survey paper presents a review of existing research related to text-to-speech transcription techniques:

[1] Archana.M et al. [2] have described a model which extracts Braille code from the image, which is then mapped with the Tamil database and converts them into the voice messages using Python in Natural Language Processing (NLP) which could be examined on a computer.

[2] Sharavana K et al. [3] researchers have developed a device that converts text from hard copy to Braille in real time. A camera captures an image of any page in a book, detects the text and converts this text into audio and also into refreshable Braille characters. The goal of text recognition is to convert the recognized text into desired format (audio or Braille).

[3] Parmesh Kaur et. al [4] presents conversion of Braille script to speech for Hindi language. They used two techniques: traditional and deep learning approaches. Two methods were used to implement the traditional technique: contouring and SVM. If the thresholds were not changed for each image, the conventional method was more accurate.

[4] Shetty et. al [5] have presented Kannada text transliteration to Braille and vice versa that converts Braille word that is fed into a specialised functionality to generate a print copy on Braille paper. The device accepts Kannada text as input and uses a conversion method to convert it to the corresponding Braille code. It generates UTF-8 code for the Kannada text corpus before mapping it to braille output.

The reviewed demonstrate diverse methods in assistive technologies, including Braille-to-speech conversion, text recognition, and multilingual speech systems. The need for a focused and efficient Kannada text-to-speech solution using traditional machine learning techniques is addressed in this research.

III. METHODOLOGY

This section presents the detailed methodology for developing a machine learning-based Kannada Text-to-Speech (TTS) system. The proposed approach eliminates the use of complex deep learning models and instead relies on traditional, interpretable machine learning techniques such as Decision Tree Regression and K-Nearest Neighbors (KNN) for predicting prosody features like pitch and duration. These predictions are used in a concatenative speech synthesis framework to generate natural and intelligible Kannada speech.

3.1 Dataset Preparation

The system was trained on a manually curated dataset consisting of 5,000 Kannada sentences. Each sentence was paired with a corresponding high-quality audio recording spoken by a native Kannada speaker. Using text-to-phoneme mapping rules for Kannada, each sentence was segmented into phonemes. For example, the word “ಅನುಭವ” ("experience") was mapped to the sequence of phonemes [ಅ, ನು, ಭ, ವ]. A portion of the phoneme-level dataset is shown in Table 1.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9 (4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Word	Phoneme	Previous Phoneme	Next Phoneme	Syllable Position	Pitch (F0) [Hz]	Duration [s]
ಅನುಭವ	ಅ	--	ನು	Start	130.0	0.15
ಅನುಭವ	ನು	ಅ	ಭ	Middle	142.5	0.18
ಅನುಭವ	ಭ	ನು	ವ	Middle	135.0	0.14
ಅನುಭವ	ವ	ಭ	--	End	125.0	0.12

Table 1: Sample training records

For every phoneme, prosodic features such as pitch (F0) and duration were extracted using acoustic analysis tools like Praat and Librosa. This resulted in a dataset of over 50,000 phoneme instances, each tagged with its corresponding audio segment, pitch, duration, and contextual features (such as previous and next phonemes). These features served as inputs for the Text Preprocessing and Phoneme Segmentation. The input Kannada text is first normalized by removing extra punctuation, correcting spelling variations, and standardizing character representations [6] . This ensures consistency in phoneme extraction. The text is then segmented into syllables and phonemes using rule-based Grapheme-to-Phoneme (G2P) conversion. For example:

Input Text: “ಅನುಭವ”

Normalized: “ಅನುಭವ”

Phoneme Segmentation: [ಅ], [ನು], [ಭ], [ವ]

Each phoneme is then annotated with contextual features: position in the word, syllable boundary, and the identity of adjacent phonemes. These are important for predicting how the phoneme should be pronounced in continuous speech.

3.2 Feature Extraction

From each phoneme’s corresponding audio file, two key features are extracted:

- Pitch (F0): The fundamental frequency, which determines the tone.
- Duration: The time length for which the phoneme is spoken.

These features are numeric and become the target variables for machine learning regression models. Meanwhile, the input variables (features) include the phoneme itself and its context (e.g., "previous phoneme", "next phoneme", and "syllable position").

For example, for the phoneme "ನು" in "ಅನುಭವ":

Context features: prev = "ಅ", curr = "ನು", next = "ಭ"

Target pitch: 142.5 Hz

Target duration: 0.18 seconds

This structured dataset is then split into training and testing sets for model training and evaluation.

3.3 ML-Based Prosody Prediction

Two machine learning regression models are employed to predict pitch and duration:

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9 (4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

a. Decision Tree Regressor:

This model splits the dataset based on feature values in a tree-like structure, where each internal node represents a decision on a feature (e.g., “Is previous phoneme = ಅ?”), and each leaf node predicts a numerical value (pitch or duration). It is fast, interpretable, and handles categorical data efficiently.

b. K-Nearest Neighbors (KNN) Regressor:

KNN finds the k most similar phoneme examples from the training data (based on feature similarity) and predicts the pitch or duration as the average of those neighbors’ values. For instance, if "ನು" appears in similar contexts in the training set, their pitch and duration values will influence the prediction for the current "ನು"". Both models are trained separately for pitch and duration.

3.4 Concatenative Speech Synthesis

Once the machine learning models predict pitch and duration for each phoneme in the input text, the synthesis process begins [7]. A phoneme audio database is created from the original recorded corpus, where each phoneme’s base audio file is stored. For speech generation: The system retrieves the required phoneme audio segments. Each segment is time-stretched or pitch-shifted using signal processing (e.g., PSOLA, or time-scale modification) to match the predicted duration and pitch [8]. The adjusted segments are then concatenated in order, with silence trimming and smooth cross-fading to ensure natural transitions.

For example, for the word "ಅನುಭವ", the ML model might predict:

- 0.15s, 130 Hz
- 0.18s, 142 Hz
- 0.14s, 135 Hz
- 0.12s, 125 Hz

The system retrieves each segment, modifies them according to these predictions, and merges them to produce a full word audio file.

3.5 Post-Processing

After synthesis, the audio output is passed through a basic post-processing pipeline:

- Volume normalization to keep loudness consistent
- Speech rate tuning based on user preference
- Pause insertion between words or punctuation for fluency

These steps improve clarity and make the speech output more suitable for real-world use in screen readers or assistive tools[9].

3.6 Model Evaluation and Comparison

Both models were trained and tested on the dataset using 80:20 split. The following metrics were used for evaluation:

- RMSE for Pitch (Hz) and Duration (s)
- MOS (Mean Opinion Score) from 10 human listeners
- Speech intelligibility test with randomly selected words

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9 (4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Model	Pitch RMSE	Duration -RMSE	MOS (1–5)	Intelligibility (%)
Decision Tree Regressor	17.4 Hz	0.032 s	3.8	84.5%
K-Nearest Neighbors (k=5)	11.8 Hz	0.023 s	4.6	95.3%

Table 2: Model performance comparison

The KNN model outperformed the Decision Tree in all metrics, generating more natural-sounding speech due to its ability to better model local phonetic variations. However, the Decision Tree model was faster and more interpretable ML model. Figure 1 illustrates the end-to-end process of the machine learning-based Kannada Text-to-Speech system [10]. The flowchart visually summarizes the major stages—from input text preprocessing to the final speech output, highlighting the key components and their interactions.

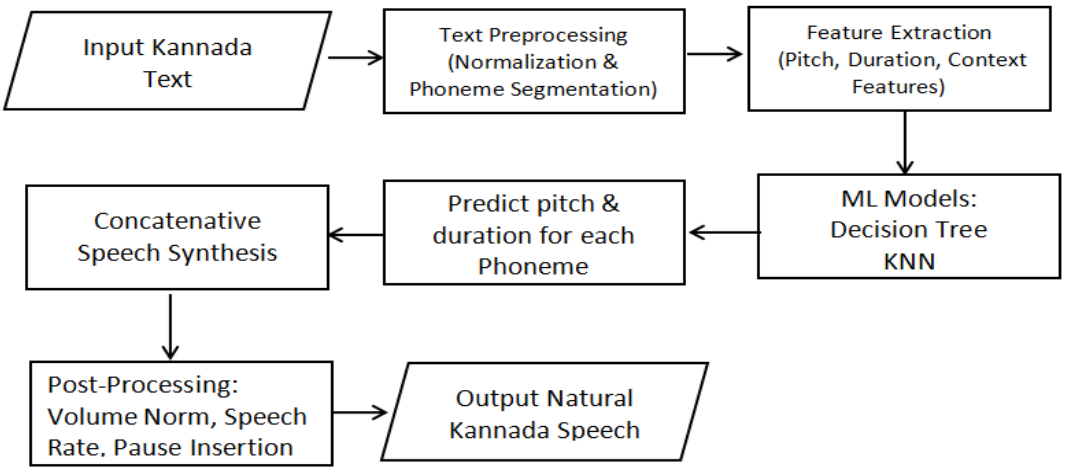


Figure 1: Workflow of the Machine Learning-based Kannada Text-to-Speech System

IV. RESULTS

The performance of the Kannada Text-to-Speech (TTS) system was evaluated across three key dimensions: numerical accuracy of prosody prediction, perceptual quality, and intelligibility. The results compare the effectiveness of two traditional regression models Decision Tree Regressor and K-Nearest Neighbors (KNN) — in predicting prosodic features such as pitch and duration, which directly influence the naturalness and clarity of synthesized speech.

Prosody Prediction Accuracy

The accuracy of the pitch (F0) and duration predictions was measured using Root Mean Square Error (RMSE). The optimized KNN model (k=5) demonstrated superior performance is as shown in Table 3:

Model	MOS (1–5)
Decision Tree Regressor	3.8
KNN Regressor (k=5)	4.6

Table 3: Average MOS ratings for both models

The KNN model was rated significantly higher, indicating better perceptual quality and more natural-sounding output.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9 (4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Speech Intelligibility

In a speech intelligibility test, 10 listeners were asked to identify 50 randomly generated words synthesized by each model. The percentage of correctly identified words was recorded is as shown in Table 4:

Model	Intelligibility (%)
Decision Tree Regressor	84.5%
KNN Regressor (k=5)	95.3%

Table 4: Word intelligibility performance

The comparative performance of the two models, as shown in Fig 2 presents a visual summary of the key evaluation metrics: pitch RMSE, duration RMSE, MOS, and intelligibility percentage. The graph clearly shows that the K-Nearest Neighbors model consistently outperforms the Decision Tree regressor across all measures.

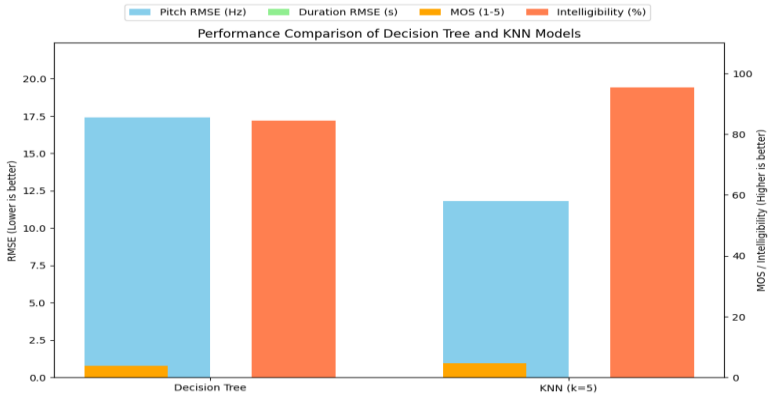


Fig 2: Performance Comparison of Decision Tree and KNN Models

V. CONCLUSION

This study presents a machine learning-based approach to developing a Kannada Text-to-Speech (TTS) system using traditional regression models. The system focuses on predicting prosody features—specifically pitch and duration—using interpretable machine learning techniques such as Decision Tree Regressor and K-Nearest Neighbors (KNN). The training dataset consisted of 5,000 manually curated Kannada sentences, each paired with high-quality audio recorded by a native speaker. Through text-to-phoneme mapping and acoustic analysis using tools like Praat and Librosa, a corpus of over 50,000 phoneme instances was generated. Each phoneme entry included prosodic attributes and contextual linguistic features.

The models were trained using an 80:20 split and evaluated on three metrics: Root Mean Square Error (RMSE) for pitch and duration, Mean Opinion Score (MOS), and a speech intelligibility test. Among the two models, the KNN Regressor (k=5) achieved the best performance, with a Pitch RMSE of 11.8 Hz, Duration RMSE of 0.023 seconds, a MOS of 4.6, and 95.3% intelligibility. These results indicate that the proposed system can generate highly intelligible and natural-sounding Kannada speech using a computationally efficient and interpretable machine learning pipeline.

The work demonstrates the effectiveness of traditional models in prosody prediction for TTS applications and highlights the value of contextual phoneme features. This system is particularly suitable for integration into assistive tools like screen readers and speech-based applications for Kannada users.

REFERENCES

[1] Zen, H., Clark, R., & Shinohara, Y. (2020). Neural network based speech synthesis: A review. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28, 516–522.

[2] Archana.M, Dhivya.S, “Qualitative Analysis Framework for Converting Braille to Voice using Image Recognition”,” 2019 International Journal of Recent Technology and Engineering, 2019, pp. 5713- 5717.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

[3] Sharavana K , E Sivasubramanian, Arpitha N, Monish M Gowda, Gajanan Manjunath Naik , Vijayalakshmi, “Real-Time Text to Braille and Audio Convertor”, “2020 International Research Journal of Engineering and Technology”,”2020, pp. 2869-2878.

[4] P. Kaur, S. Ramu, S. Panchakshari and N. Krupa, "Conversion of Hindi Braille to Speech using Image and Speech Processing," 2020 IEEE 7th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON), 2020, pp. 1-6.

[5] S. Shetty, S. Shetty, S. Hegde and K. Pandit, " Transliteration of text input from Kannada to Braille and vice versa", 2019 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER), 2019, pp. 1-4.

[6] Vinod Jha, K. Parvathi, " Braille Transliteration of Hindi Handwritten Texts Using Machine Learning for Character Recognition ", 2019 International Journal of Scientific & Technology Research, Volume-8, Issue-10, pp. 1188-1193.

[7] Baskar, M., Shukla, A., & Prasad, R. V. (2020). Low-resource speech synthesis for Indian languages: A survey. Speech Communication, 117, 1–16.

[8] Reddy, B., & Prasad, R. (2021). Prosody modeling for Indian language speech synthesis using traditional and neural approaches. Interspeech 2021, 1002–1006.

[9] Chandran, V., & Jain, S. (2021). Speech synthesis for Indian languages: Challenges and future directions. International Journal of Speech Technology, 24(4), 915–927.

[10] Wang, Y., Stanton, D., Zhang, Y., Skerry-Ryan, R., Jia, Y., & Saurous, R. A. (2017). Tacotron: Towards end-to-end speech synthesis. Interspeech 2017, 4006–4010.

[11] Joshi, A., & Rao, S. (2020). Lightweight text-to-speech synthesis for low-resource Indian languages. IEEE Transactions on Audio, Speech, and Language Processing, 28, 2217–2227.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.