Utilizing POS-Driven Pitch Contour Analysis for Enhanced Tamil Text-to-Speech Synthesis

Preethi Thinakaran[1], Anushiya Rachel Gladston[1], P. Vijayalakshmi[2], T. Nagarajan[1], Malarvizhi Muthuramalingam[1], Sooriya S[1]
[1]Department of CSE, Shiv Nadar University Chennai, India
[2]Department of ECE, Sri Sivasubramaniya Nadar College of Engineering, Chennai, India
preethit@snuchennai.edu.in, anushiyarachelg@snuchennai.edu.in,vijayalakshmip@ssn.edu.in
nagarajant@snuchennai.edu.in, malarvizhim@snuchennai.edu.in, sooriyas@snuchennai.edu.in

Abstract

A novel approach to text-to-speech synthesis that integrates pitch contour labels derived from the highest occurrence analysis for each Part-of-Speech (POS) tag. Using the Stanford POS Tagger, grammatical tags are assigned to words, and the most frequently occurring pitch contour labels associated with these tags are analyzed, focusing on both unigram and bigram statistics. The primary goal is to identify the pitch contour for each POS tag based on its frequency of occurrence. These pitch contour labels are incorporated into the output of the synthesized waveform using the TD-PSOLA (Time Domain Pitch Synchronous Overlap and Add) signal processing algorithm. The resulting waveform is evaluated using Mean Opinion Scores (MOS), demonstrating significant enhancements in quality and producing a prosodically rich synthetic speech.

1 Introduction

Text-to-speech (TTS) systems convert written text into natural-sounding speech, with prosody playing a vital role in achieving high-quality synthesis. Prosody encompasses pitch, rhythm, and intensity, with pitch contour being particularly important for conveying meaning and emotion in spoken language. Natural human communication sees pitch contour influenced by various factors, including pitch accents, phrase boundaries, and intonation patterns (Pierrehumbert , 1980).

While significant advancements have been made in integrating prosodic features into TTS systems across various languages, Tamil—a Dravidian language known for its complex tonal and prosodic structures—remains underexplored. Existing research has investigated methods for enhancing prosody in TTS, such as Statistical Parametric Speech Synthesis (SPSS), which offers a range of techniques to improve spoken output (Zen et al. , 2009). Deep learning models like Tacotron and WaveNet have made substantial strides in generating natural-sounding speech by learning direct mappings from text to audio waveforms (Wang et al. , 2017; van den Oord et al. , 2016). However, there is a need for approaches that effectively combine linguistic elements—such as Part-of-Speech (POS) tags—with prosodic features to enhance the naturalness and intelligibility of synthesized speech. In this regard, techniques like Linear Prediction (LP) and Time-Domain Pitch Synchronous Overlap and Add (TD-PSOLA) have been utilized to enhance naturalness and expressiveness (Gladston et al. , 2014, 2022; Rachel et al. , 2015).

The objective of the current work is to enhance a synthesized speech signal, by analysing pitch contour labels associated with various POS tags. Utilizing the Stanford POS Tagger, grammatical tags are assigned to words in Tamil utterances, and the most frequently occurring pitch contours for each tag are analysed through unigram and bigram statistics. This analysis enables the integration of the most common pitch contours into synthesized speech. The identified pitch contours will be incorporated to the synthesized speech signal using the TD-PSOLA technique, resulting in more natural and expressive Tamil speech synthesis.

To evaluate the effectiveness of this approach, a Mean Opinion Score (MOS) is obtained. Preliminary results indicate that this method achieves a higher MOS score compared to the output from the HMM-based TTS system, demonstrating its potential to significantly improve the naturalness of synthesized Tamil speech.

The paper is organized as follows: Section 2 outlines the methodology employed for POS tagging and pitch contour analysis. Section 3 presents the proposed approach, highlighting the integration of pitch contour labels based on POS tags. Section 4 discusses the results and analysis. Finally, Section 5 concludes the paper and outlines directions for future work.

2 Methodology

The objective of this study is to identify the predominant pitch contour labels associated with different Part-of-Speech (POS) tags in Tamil language utterances using the MILE corpus and validate these findings through a testing phase.

2.1 Speech Corpus

The IISc-MILE Tamil ASR Corpus (Madhavaraj et al. , 2022), originally developed for Automatic Speech Recognition (ASR) in the Tamil language is used for this study. The dataset comprises approximately 150 hours of read speech data (and the corresponding transcriptions), recorded from 531 native Tamil speakers in a controlled, noise-free environment using high-quality USB microphones. The data is split into training and testing sets, consisting of 6000 and 2000 utterances respectively, to ensure the model's generalization. The training set was used to learn the patterns of pitch labels associated with each POS, while the testing set was reserved for validating the system's performance.

2.2 Pitch Contour Label Extraction

A ToBI annotation tool is developed for extracting the prosodic features from speech signal. The ToBI annotation tool is used to extract the pitch contour from the speech signal. The tool first converts the audio into a single channel and resamples it to 16 kHz. Utilizing Automatic Speech Recognition (ASR), the tool generates an orthographic transcription, which is combined with the speech signal to identify phoneme, syllable, and word-level boundaries. Furthermore, the tool analyses the speech signal to estimate the relative intensity index, pitch contour labels, and break indices.

Pitch contours are calculated by estimating the fundamental frequency (F0) using the autocorrelation method on a frame-by-frame basis, with each frame set at 20 ms and a hop length of 10 ms. To smooth out the pitch variations at the word level, a third-order polynomial is applied, preserving the overall contour shape while reducing noise. The pitch contours are classified into eleven predefined shapes: L (Low), H (High), HLL, HHL, LLH, LHH, HLH, LHL, "hat," "bucket," and "flat." The L shape represents a falling pitch, while H indicates a rising pitch. If the dynamic range is less than 10 Hz, the contour is considered flat, with no significant pitch variation. The basic pitch contour shapes (except flat) are portrayed in Fig. 1.

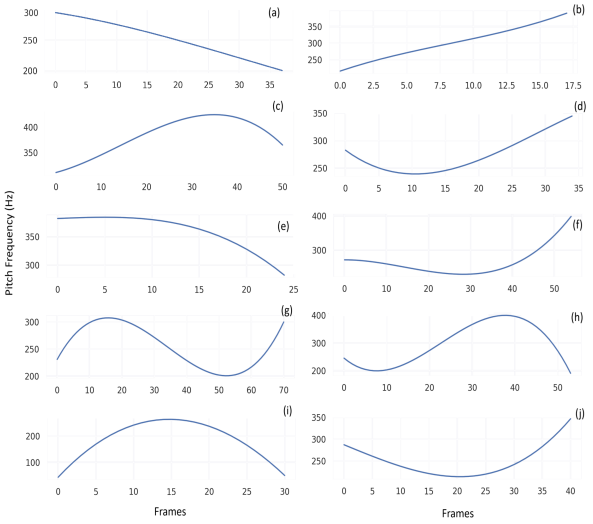

Figure 1: Basic Pitch Contour Shapes Considered: (a) H, (b) L, (c) HHL, (d) LHH, (e) HLL, (f) LLH, (g) HLH, (h) LHL, (i) hat, (j) bucket

2.3 Unigram and Bigram Statistics

As mentioned earlier, the MILE corpus is used for this analysis, with 6000 utterances employed for training and 2000 utterances for testing the analysis. The utterances are tagged using the Stanford POS tagger, along with pitch contour extraction using the ToBI annotation tool. The unigram and bigram statistics for 31 POS tags and 11 pitch contour labels are analysed, focusing on the highest occurrence counts of the pitch contour labels. The shortest path through the POS tag sequences was calculated by taking the logarithm of the highest counts of the occurrences of the first three labels for each tag. With the shortest path analysis, the path with minimum cost is considered, and the identified label is designated as the respective label for each POS tag. Pitch contour labels from the utterances were matched against those obtained from the annotation tool. For better analysis and to reduce complexity, some pitch contour labels are grouped together. The grouping of labels [HHL, HLL, HLH] and [LLH, LHH, LHL] can be justified based on their overall pitch contour patterns and dominant pitch behaviour. The first group, $[HHL, HLL, HLH]$, is characterized by high-pitch dominance, as these contours either start with or predominantly feature high pitch (H), often followed by a drop to low pitch (L) or alternating patterns with an emphasis on high. For instance, HHL and HLL exhibit a falling contour, while HLH shows a rising-falling pattern with a strong presence of high pitch at the start or end.

In contrast, the second group, $[LLH, LHH, LHL]$, demonstrates low-pitch dominance, as these contours predominantly feature low pitch (L), with a tendency to rise to high pitch (H) or alternate while maintaining a strong presence of low. For example, LLH and LHH represent rising contours, while LHL shows a falling-rising pattern with a low-pitch emphasis at the start or end.

This classification reflects the contours tonal movement, focusing on whether high or low pitch is more prominent in the overall shape. An exact match was assigned a score of 1, while matches within the same group received a score of 0.5. The analysis yielded a match percentage of 35% for unigram statistics and 40% for bigram statistics, providing insights into the alignment between POS tagging and pitch contour labeling processes.

Table 1: Analysis Results of POS Tagging and Pitch Contour Labeling

Statistic Type	Match Percentage
Unigram	35%
Bigram	40%

3 Prosody Modification in Synthesized Speech

Given a text, corresponding to a sentence, using unigram and bigram statistics, for each of the words, a suitable pitch contour is assigned. The identified pitch contour labels are incorporated into the synthesized waveform generated by the Hidden Markov Model (HMM)-based TTS system (Rachel et al. , 2015), known for its small footprint and reduced redundancy. The output waveform from the HTS system typically has a neutral tone. To enhance expressiveness, pitch contour modifications are applied using the Time-Domain Pitch-Synchronous Overlap-Add (TD-PSOLA) technique. This method allows for accurate pitch adjustments while preserving the natural quality of the original audio signal, ensuring the synthesized speech retains both clarity and emotional nuance.

To directly modify the prosody of speech without relying on modeling, a signal processing algorithm known as Time-Domain Pitch-Synchronous Overlap and Add (TD-PSOLA) is employed (Gladston et al. , 2014, 2022). Since it operates pitch-synchronously, the instants of excitation, or Glottal Closure Instants (GCIs), are derived using a phase-based algorithm (PD) (Rachel et al. , 2017). The speech signal is segmented using a Hanning window, with segments centered around the GCIs. To modify both the pitch contour, the algorithm selects, overlaps, and adds the segments aligned with the new excitations, preserving naturalness in the synthesized speech.

This approach allows for precise pitch modifications without altering the duration of the original signal, resulting in a synthesized output that retains the natural characteristics of the speech while achieving the desired pitch contour adjustments.

Changes in pitch contours significantly influence the expression of emotions in the synthesized waveform. For instance, a rising pitch contour may convey excitement or enthusiasm, while a falling contour can indicate sadness or calmness. By accurately adjusting the pitch according to these emotional cues, the synthesized speech can better reflect human-like emotional expressions.

Finally, the generation of the modified waveform aims to produce output with natural prosody. By integrating pitch adjustments that align with the prosodic features of human speech, the synthesized audio achieves a more fluid and realistic sound, enhancing listener engagement and comprehension.

The figure 2 illustrates the comparison between the pitch contour of the synthesized speech signal and the pitch contour processed using the TD-PSOLA-based signal processing algorithm. The effectiveness of TD-PSOLA lies in its ability to operate at the signal level, minimizing distortion and ensuring a high-quality synthesized output.

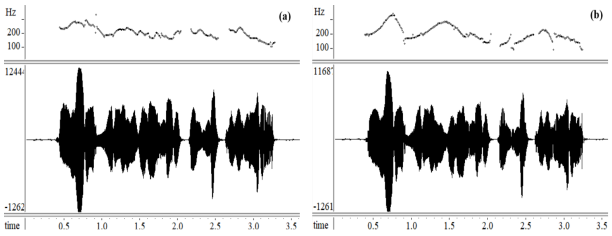

Figure 2: (a) Synthesized wave from HTS system; (b) The pitch contour modified for the entire speech signal using TD-PSOLA algorithm.

4 Results and Analysis

The performance of the proposed method was evaluated using the Mean Opinion Score (MOS) as a subjective measure to assess the effectiveness of pitch contour fitting for each word segment, incor-

porating the contours into the full utterance. A total of 10 audio samples were synthesized, and MOS scores were obtained from 10 listeners, resulting in an average score of 5.

The MOS score obtained from HTS is 3.58. When comparing the synthesized output with pitch contour modifications applied through the TD-PSOLA technique to the baseline synthesized output from the HTS system, the modified output achieved an MOS score of 4.1 out of 5. This score indicates a significant improvement in prosody, showcasing the effectiveness of pitch contour fitting based on Part-of-Speech (POS) tag analysis.

To extend the scope of this analysis, the pitch contour modifications were also incorporated into synthesized waveforms generated by state-of-the-art neural TTS models, namely Tacotron and FastSpeech2. Tacotron, known for its sequence-to-sequence modeling capability, achieved an MOS score of 4.35 when augmented with the pitch contour modifications. FastSpeech2, a parallel TTS model designed for faster synthesis, achieved an MOS score of 4.3 with the same enhancements. These results indicate that the integration of pitch contour fitting is effective across different TTS frameworks, improving prosody and naturalness consistently.

TTS System / Configuration	**MOS Score (Out of 5)**
Baseline HMM-Based TTS (HTS)	3.58
HMM-Based TTS + Pitch Contour Modifications	4.10
Tacotron (Baseline)	4.00
Tacotron + Pitch Contour Modifications	4.35
FastSpeech2 (Baseline)	4.05
FastSpeech2 + Pitch Contour Modifications	4.30

Table 2: Comparison of MOS Scores for Different TTS Systems and Configurations

The enhanced scores across all models highlight the ability of the modified synthesis process to convey emotional expressiveness and naturalness, bridging the gap between machine-generated and human-like speech. By integrating statistical analyses of pitch contours into the synthesis process, the results show a clear improvement in the prosody of synthesized speech, providing a more engaging and natural listening experience. These findings underscore the potential of incorporating advanced prosodic features into TTS systems for applications requiring high-quality, expressive speech synthesis.

5 Conclusion

In this paper, a novel approach to Tamil text-to-speech synthesis that integrates pitch contour based on analysis of Part-of-Speech (POS) tagging is proposed. Our results demonstrate significant improvements in the naturalness and emotional expressiveness of synthesized speech, as indicated by the Mean Opinion Scores (MOS) obtained from listener evaluations. By leveraging the TD-PSOLA technique, we successfully incorporated pitch contour modifications, leading to a more engaging auditory experience that reflects the nuances of human communication. While the model effectively captures pitch variations, further refinements are necessary to encompass a broader range of emotional expressions and linguistic features.

Acknowledgement

The current work is carried out as a part of the project titled, “Prosody Modeling”, under the subproject of the NLTM BHASHINI project, titled, “Speech technologies in Indian languages”, funded by the Ministry of Electronics and Information Technology, Government of India, with reference number, 11(1)/2022-HCC(TDIL).

6 Limitations

This study has limitations that impact the generalizability and scope of the results. First, the IISc-MILE Tamil ASR Corpus, while consisting of approximately 150 hours of speech data, may not fully encompass the wide range of dialects, accents, and prosodic variations inherent in Tamil spoken across different regions. This limits the application of the pitch contour modification method to speakers outside the dataset’s demographic. Second, the focus on pitch contour modification improves prosody but neglects other prosodic elements such as rhythm and intensity, which are critical to achieving fully natural speech synthesis. The use of the TD-PSOLA algorithm for pitch modification may also introduce artifacts, particularly in cases of extreme pitch adjustments, potentially degrading the quality of the synthesized speech. Moreover,

the subjective evaluation using the Mean Opinion Score (MOS) can introduce bias, as perceptions of speech naturalness can vary widely among listeners, suggesting the need for more objective evaluation metrics to supplement the subjective analysis.

7 Ethical Considerations

The current work complies with the ACL ethics policy. All data used in this study was sourced from publicly available datasets with appropriate permissions for research use.

References

Aarabi, P., Shi, G., Shanechi, M., and Rabi, S., *Phase-Based Speech Processing*, World Scientific Publishing Co., 2005.

Black, A. W., and Hunt, A. J., "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proceedings of ICASSP*, 1996.

Gladston, A. R., Sreenidhi, S., Vijayalakshmi, P., and Nagarajan, T., "Incorporation of happiness into neutral speech by modifying emotive keywords," in *Proceedings of IEEE TENCON 2014*, Bangkok, Thailand, Oct. 2014, pp. 1–6.

Gladston, A. R., Sreenidhi, S., Vijayalakshmi, P., and Nagarajan, T., "Incorporation of Happiness in Neutral Speech by Modifying Time-Domain Parameters of Emotive-Keywords," *Circuit Systems and Signal Processing*, vol. 41, no. 4, pp. 2061–2087, Apr. 2022.

Madhavaraj, A., Bharathi, P., and Ramakrishnan, G. A., "Subword dictionary learning and segmentation techniques for automatic speech recognition in Tamil and Kannada," arXiv, 2022. [Online].

Pierrehumbert, J. B., "The phonology and phonetics of English intonation," Ph.D. dissertation, Massachusetts Institute of Technology, 1980.

Rachel, G. A., Vijayalakshmi, P., and Nagarajan, T., "Estimation of glottal closure instants from degraded speech using a phase-difference-based algorithm," *Computer Speech & Language*, vol. 46, pp. 136–153, 2017.

Rachel, R. A., Solomi, V. S., Vijayalakshmi, P., and Nagarajan, T., "A small-footprint context-independent HMM-based synthesizer for Tamil," *International Journal of Speech Technology*, vol. 18, no. 2, pp. 281–289, Apr. 2015.

Shen, J., Pang, R., Weiss, R. J., et al., "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proceedings of ICASSP*, 2018.

Tokuda, K., Zen, H., and Black, A. W., "An HMM-based speech synthesis system applied to English," in *Proceedings of the IEEE Workshop on Speech Synthesis*, 2002.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y., "Feature-rich part-of-speech tagging with a cyclic dependency network," in *Proceedings of NAACL*, 2003.

van den Oord, A., Dieleman, S., Zen, H., et al., "WaveNet: A generative model for raw audio," in *Proceedings of SSW*, 2016.

Wang, Y., Skerry-Ryan, R., et al., "Tacotron: Towards end-to-end speech synthesis," in *Proceedings of Interspeech*, 2017.

Zen, H., Tokuda, K., and Black, A. W., "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.