INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Comparative Analysis of Kannada Formant Synthesized Utterances and their Quality

Alfred Vivek D'Souza[1]*, **D J Ravi**[2]

1 Research Scholar, Department of ECE, Vidyavardhaka College of Engineering, Mysuru, Karnataka, India
2 Research Supervisor, Department of ECE, Vidyavardhaka College of Engineering, Mysuru, Karnataka, India

OPEN ACCESS

Received: 27-10-2022
Accepted: 16-12-2022
Published: 04-02-2023

Citation: D'Souza AV, Ravi DJ (2023) Comparative Analysis of Kannada Formant Synthesized Utterances and their Quality. Indian Journal of Science and Technology 16(5): 309-317. https://doi.org/10.17485/IJST/v16i5.2091

* **Corresponding author**.

alfreddsouza@vvce.ac.in

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: The goal of this work is to synthesize Kannada utterances using a modified Klatt type formant synthesizer to evaluate its performance by comparing against eSpeak synthesizer in terms of intelligibility and quality of the utterances generated. **Methods:** Kannada utterances viz., vowels, diphthongs, Consonant-Vowel (CV) coarticulations and simple words are generated using a modified Klatt type formant synthesizer and eSpeak. The vowels and diphthongs generated by both the synthesizers are compared with natural recorded utterances using F1-F2 formants and the CV co-articulations are compared using spectrograms. The synthesized word utterances are compared with natural recorded utterances using Log Spectral Distance to find out which synthesizer outputs the frequency spectrum that is closest to the frequency spectrum of the natural utterances. Also, the synthesized word utterances are evaluated for their intelligibility and quality using Mean Opinion Score (MOS) obtained from 10 native Kannada language speakers. **Findings:** The word utterances synthesized by the modified Klatt type formant synthesizer scored a MOS of 86% and 4.46 out of 5 for the parameters of intelligibility and quality whereas for the same two parameters eSpeak scored 70% and 4.14 out of 5 respectively. **Novelty:** Klatt type formant synthesizer that uses pitch synchronous parameter update method synthesizes good quality Kannada sound utterances and storing the control parameters of the synthesizer using polynomials reduces the database footprint.

Keywords: Kannada Formant Synthesizer; Klatt type Synthesizer; eSpeak; Kannada TTS; Formant synthesis quality

1 Introduction

Speech Synthesizers form the core of Text-To-Speech (TTS) conversion systems and have a plethora of applications like screen readers for visually impaired, talking help for those who cannot speak, language teaching and learning aid [(1)] and in Natural Language Processing (NLP) for Human Computer Interface (HCI). There are many speech synthesis techniques available and the important ones are listed below[2,3].

1. Concatenative synthesis : A set of pre-recorded messages/utterances are stored in the database which are retrieved, concatenated and played back according to the input text. Pre-recorded utterances can be phonemes, diphones, words or even commonly used phrases or sentences.

2. Articulatory synthesis: In this method, a detailed model is developed for each articulator and is used for speech synthesis. Epiglottis, tongue, teeth, lips etc., are involved in human speech production mechanism and are called as articulators. Advanced neural networks are also employed sometimes to control the model inputs.

3. Formant synthesis: It is a rule-based method that uses the model of human vocal tract to synthesize speech utterances. These synthesizers require a number of control parameters[4] to generate different utterances.

Following are some of the advantages of using formant synthesis technique for synthesizing speech utterances over the other techniques.

1. Formant synthesizers require only the control parameters extracted from the speech utterances to be stored in the database[4]. Whereas, in concatenative speech synthesizers the recorded utterances are stored leading to huge database size[5,6].

2. Formant synthesis is a model-based approach unlike concatenative synthesis. The use of model helps in understanding speech production mechanism[7] and language structure.

3. Formant synthesizers can produce intelligible speech at high rates of utterance[7,8] and avoids the acoustic glitches that commonly plaques concatenative speech synthesizers.

4. Formant synthesizers use vocal tract models that are simpler compared to that of articulatory synthesizers.

5. Complete control over all aspects of output speech including intonation and prosody is possible in formant synthesizers[4,7].

Formant synthesizers use acoustic models based on source-filter theory[9,10] and the main components used to model human speech production system is shown in Figure 1 .

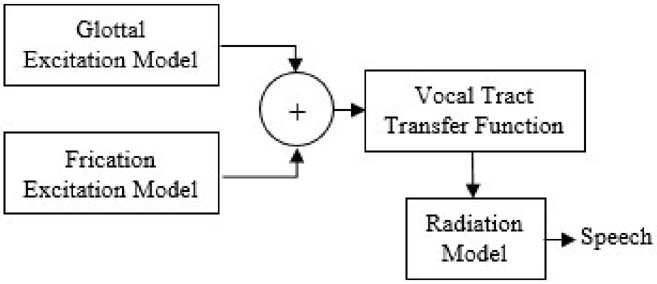


Fig 1. Source-Filter model of speech production

The idea behind source-filter model of human speech production mechanism can be expressed as Equation 1 in frequency domain.

$$P(f) = S(f) \cdot T(f) \cdot R(f) \tag{1}$$

P(f) is the spectrum of the output speech, S(f) is the spectrum of the source, T(f) is the linear transfer function of the vocal tract and R(f) is the radiation characteristics. The glottal excitation model generates harmonic rich quasi periodic signal called as glottal wave for producing voiced sounds such as vowels, diphthongs etc., The frequency of the glottal wave is called the Pitch. The frication excitation model is used for generating noise like signal for producing unvoiced sounds such as fricatives. However, at times both sources may be combined to get natural sounding utterances. The combined spectrum of glottal excitation model and frication excitation model appears as S(f) in Equation 1. The human vocal tract acts as frequency selective network and is represented as vocal tract transfer function T(f) which can be implemented as combination of digital filters each having its own resonant frequency and bandwidth. The radiation model whose spectrum is represented as R(f) in Equation 1 is used for modelling the sound radiating out of lips and nostrils and is normally implemented as a High Pass Filter.

Obtaining natural sounding speech utterances from a formant synthesizer by optimizing different parameters is still a challenging and an active area of research[8]. Much attention is also not given for developing or evaluating formant synthesizers for Indian languages[7,8] especially Kannada, which is predominantly spoken in the state of Karnataka, India with atleast 35 million native speakers. This work tries to overcome these lacunae by attempting to synthesize Kannada utterances such as vowels, diphthongs, Consonant-Vowel (CV) coarticulations and simple words using a modified Klatt type formant synthesizer[11] and eSpeak[12], a popular formant synthesizer. This work also compares the intelligibility and quality of Kannada utterances generated by both the synthesizers.

1.1 Modified Klatt Type Formant Synthesizer

The architecture of modified Klatt type formant synthesizer[11] is shown in Figure 2 . This synthesizer has a cascade and a parallel vocal tract model with 5 resonators each. In the cascade arm, an anti-resonator is used for nasal coupling. In parallel arm, each resonator has an independent amplitude control. Resonators and anti-resonator have their own center frequency and bandwidth parameters.

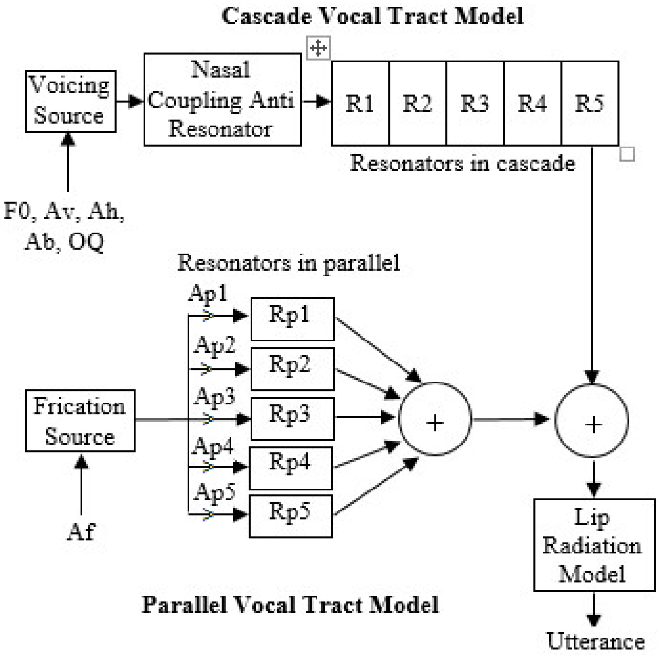


Fig 2. Modified Klatt type formant synthesizer

This synthesizer uses Rosenberg's glottal model as excitation source for generating voiced sounds. The control parameters for voicing source are pitch (F0), voicing amplitude (Av), aspiration amplitude (Ah), breathiness amplitude (Ab) and open quotient (OQ). For generating unvoiced sounds, a random number generator is used as frication source with frication amplitude (Af) as its control parameter.

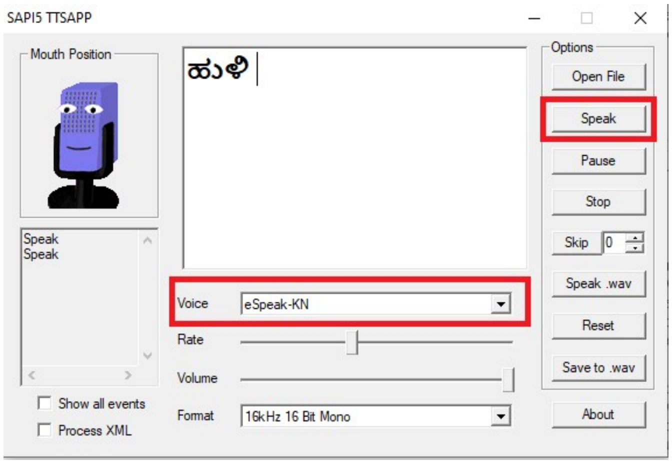
The utterance is synthesized by applying appropriate control parameter values which are stored in the database as time normalized polynomial curves. The control parameters are updated once per pitch period. Synthesizing utterances requires the control parameters to be generated first by evaluating time normalized polynomial curves of pitch, formants, bandwidths, open quotient, voicing/frication amplitude, breathiness amplitude and aspiration amplitude etc., The synthesizer then synthesizes the sound utterance using the generated parameters.

Generating control parameters for isolated vowels, diphthongs and consonants are straight forward and involves evaluation of time normalized polynomial curves of various parameters. However, generating control parameters for CV coarticulations involves evaluation of locus and transition equations[11].

1.2 eSpeak Formant Synthesizer

eSpeak was originally known as speak, created by Jonathan Duddington and was written for Acorn/FISC_OS computers. The development started in 1995 and today it is available as an open source speech synthesizer for Linux, Windows and Android platforms that supports over 100 different languages including Kannada. The eSpeak engine is written in C/C++ language and is available as a command line program, a shared library and also a SAPI5 version for windows. The eSpeak engine supports Klatt formant synthesis along with Multi-Band Resynthesis OverLap Add (MBROLA) diphone synthesis. However, MBROLA voices are not used in this work.

The implementation of eSpeak synthesizer is based on Klatt formant synthesizer. The cascade vocal tract model has 6 resonators and two anti-resonators for nasal coupling and the parallel vocal tract model has 6 resonators. The utterance is generated by applying values of different parameters to the synthesizer such as pitch, formant frequencies, bandwidths etc., which are stored for each sound in the database. The parameters are updated once per frame. Synthesizing the utterances using eSpeak is a straight forward task. The software package is available with a Graphical User Interface (GUI) which is shown in Figure 3 . The required text is entered in the textbox provided and eSpeak-KN voice is selected for Kannada language. Upon pressing ‘Speak’ button the parameters are fetched from the database and the sound is synthesized by the synthesizer. The synthesized utterance can be saved in .wav format file for further analysis.



ಹುಳಿ

Speak
Speak

Voice eSpeak-KN

Format 16kHz 16 Bit Mono

Skip 0

Fig 3. Screenshot of eSpeak – SAPI5 TTSAPP GUI

2 Methodology

The following types of Kannada utterances were generated using the modified Klatt type formant synthesizer and eSpeak for performing analysis

1. Isolated vowels and diphthongs
2. Isolated Consonant-Vowel(CV) coarticulations
3. Simple words

2.1 Isolated Vowels and Diphthongs

Ten isolated vowels and two diphthongs of Kannada were synthesized. The waveform and spectrogram of vowel /o/(☒) and diphthong /ai/(☒) synthesized using both formant synthesizers are shown in Figure 4 as example. Vowels being continuant sounds have almost constant formant contours but, on the other hand, the formant contours of the diphthongs start at values corresponding to one vowel and moves towards the values of a different vowel. The first three formants F1, F2 and F3 are marked as red lines on the spectrogram in Figure 4 .

The first two formants ie., F1 and F2 help listeners to perceive the vowel and diphthong sounds correctly [9]. Thus, F1 and F2 formants of the synthesized vowels and diphthongs were extracted and compared with the average values of F1 and F2 formants obtained by recording naturally uttered vowels and diphthongs by five native Kannada language speakers (2 Male and 2 Female) belonging to age group of 22 years to 30 years. The first two formants were treated as vectors (F1, F2) and Euclidean distance between synthesized utterance formant vectors and natural recorded utterance formant vectors were found using Equation 2.

$$\|F_R - F_S\| = \sqrt{(F1_R - F1_S)^2 + (F2_R - F2_S)^2} \qquad (2)$$

Where F_R is the (F1, F2) formant vector of recorded natural utterances and F_S is the (F1, F2) formant vector of synthesized utterances. The extracted F1 and F2 formants along with Euclidean distances are listed in Table 1 . It is evident from Table 1 that the Euclidean distance between (F1, F2) of utterances synthesized using the modified Klatt type formant synthesizer and (F1, F2) of recorded utterance is less than that of eSpeak suggesting that the former produces vowels and diphthongs that is closer to natural sound.

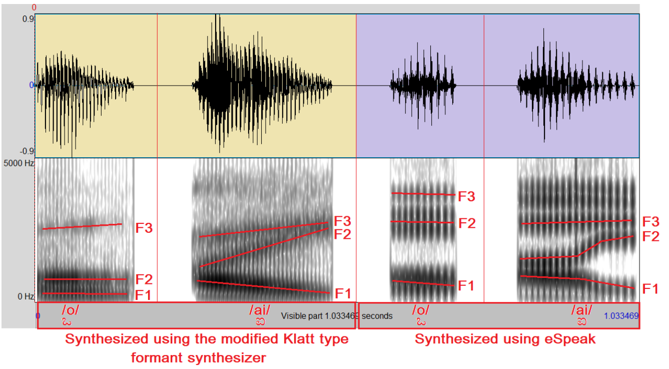


Fig 4. Waveform and Spectrogram of synthesized vowel /o/(⊠) and diphthong /ai/(⊠)

Table 1. F1 – F2 formant values of vowels and diphthongs

Sound utterance		**Average of 5 Kannada Speakers**	**Modified Klatt type formant synthesizer**		**eSpeak**	
		F_R = (F1, F2) in Hz	**F_{S1} = (F1, F2) in Hz**	**$\|\|F - F_{S1}\|\|$**	**F_{S2} =(F1, F2)in Hz**	**$\|\|F - F_{S2}\|\|$**
/a/(⊠)		(666,1256)	(640,1226)	39.69	(706,1297)	57.42
/aa/(⊠)		(765,1277)	(821,1197)	98.09	(772,1178)	99.66
/i/(⊠)		(258,2592)	(297,2337)	257.51	(310,2208)	387.05
/ii/(⊠)		(241,2588)	(321,2380)	222.63	(287,2280)	311.08
/u/(⊠)		(358,931)	(318,903)	48.38	(375,1066)	136.32
/uu/(⊠)		(329,729)	(320,691)	38.57	(411,902)	191.98
/e/(⊠)		(464,2328)	(450,2137)	191.51	(444,1989)	339.58
/ee/(⊠)		(423,2363)	(402,2026)	337.88	(444,1987)	376.76
/o/(⊠)		(494,910)	(347,780)	196.25	(564,1419)	513.96
/oo/(⊠)		(467,992)	(464,737)	254.62	(541,1295)	312.2
/ai/	Start	(791,1623)	(697,1433)	212.25	(882,1283)	352.41
(⊠)	End	(335,2516)	(380,2248)	271.72	(392,2181)	339.78
/au/	Start	(747,1128)	(742,1136)	9.67	(723,1127)	23.81
(⊠)	End	(410,824)	(456,926)	111.61	(476,690)	149.82

2.2 Isolated Consonant Vowel (CV Co-articulations)

The coarticulation is a process of transition from one utterance to the another that causes the vocal tract to vary its shape and its parameters in a complex way as opposed to same two sound units uttered in rapid succession. In CV coarticulations, a vowel sound succeeds a consonant sound and the formants of the succeeding vowels get modified in the beginning portion of the utterance in accordance with the locus theory. Figure 5 shows waveform and spectrogram of utterance /ka/(⊠) synthesized using both the synthesizers as an example.

In Figure 5 , it is worth noting that the onset of the formants is different from that of the middle or ending portion of vowel part in the spectrogram of utterance synthesized by the modified Klatt type formant synthesizer as per locus theory. However, the same is not so evident in case of spectrogram synthesized by eSpeak. Most of the CV coarticulations were synthesized correctly by both the synthesizers with some exceptions. eSpeak generated a faint voice bar for voiced plosives such as /ga/(⊠), /ja/(⊠), /da/(⊠) and /ba/(⊠)/ making them difficult to recognize if uttered in isolation. On the other hand, the modified Klatt type formant synthesizer synthesized /L/(⊠)that sounded like /l/(⊠)and the nasal sound /N/(⊠)that sounded like /n/(⊠)when uttered in isolation.

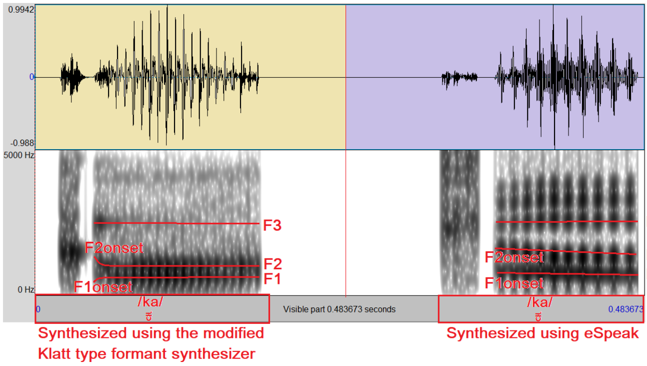


Fig 5. Waveform and Spectrogram of synthesized CV coarticulation /ka/(⊠)

2.3 Simple Words

Five sample word utterances viz., /huLi/ (⊠⊠⊠-sour), /maavu/ (⊠⊠⊠⊠ - mango), /kaage/ (⊠⊠⊠⊠- crow), /ravi/ (⊠⊠⊠- sun, proper noun) and /ruupaayi/ (⊠⊠⊠⊠⊠⊠- rupee) were synthesized using the modified Klatt type formant synthesizer and eSpeak. The same word utterances were also recorded from a native Kannada language male speaker as reference for comparison. Figure 6 shows an example of waveform and spectrogram of synthesized utterance /Ruupaayi/(⊠⊠⊠⊠⊠⊠)

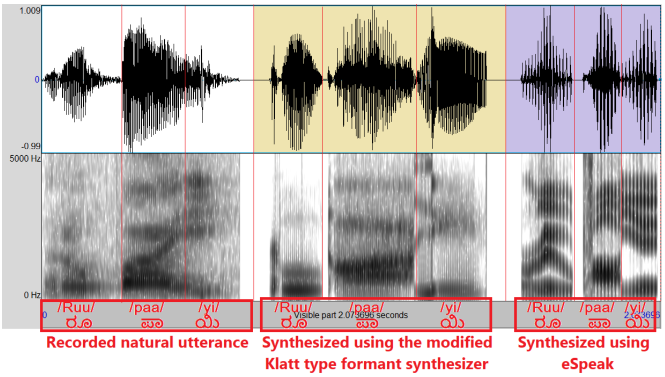


Fig 6. Waveform and Spectrogram of utterance /Ruupaayi/(⊠⊠⊠⊠⊠⊠)

The sample word utterances contain various sound types such as plosives (/k/-⊠⊠, /g/-⊠⊠and /p/-⊠⊠), nasal (/m/-⊠⊠), fricative (/h/-⊠⊠), Trill (/r/-⊠⊠), approximant (/L/-⊠⊠), semi-vowels (/y/-⊠⊠and /v/-⊠⊠)and trailing vowels (/a/-⊠, /aa/-⊠, /i/-⊠, /u/-⊠, /uu/-⊠ and /e/-⊠). Following are some of the observations that can made by visual comparison of the spectrograms of synthesized word utterances by the modified Klatt type formant synthesizer and eSpeak.

1. The spectrogram generated by eSpeak has striped appearance compared to that of the modified Klatt type formant synthesizer.

2. The spectrogram generated by eSpeak has very less or no frequency components outside bandwidths of formant frequencies.

3. The formant contours generated by modified Klatt type formant synthesizer is closer to the formant contours of recorded utterance than that of eSpeak.

The striped appearance of spectrogram in eSpeak is caused because of fixed parameter update rate and lack of frequency contents outside formant bandwidths is indicative of a robotic voice with less naturalness. On the other hand, the spectrogram produced by modified Klatt type formant synthesizer which employs pitch synchronous parameter update method shows less striped appearance and is closer to the spectrogram of recorded utterances.

3 Results and Discussion

The synthesized word utterances were compared with recorded word utterances using the following two techniques.

1. Log Spectral Distance
2. Mean Opinion Score

3.1 Log Spectral Distance

Log Spectral Distance (D_{LS}) is defined in Equation 3.

$$D_{LS} = \sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi}\left[10\,log_{10}\left(\frac{P_1(\omega)}{P_2(\omega)}\right)\right]^2 d\omega} \tag{3}$$

Where $P_1(\omega)$ and $P_2(\omega)$ are the two spectra between which the log spectral distance is to be calculated. D_{LS} is 0 if and only if the spectra under comparison are a complete match. A non-zero D_{LS} means spectral mismatch but is not indicative of the degree of mismatch. Thus, D_{LS} is only a figure of merit and can be used for relative comparisons. Using Equation 3, D_{LS} was calculated between recorded and synthesized word utterances to assess how close the spectra of synthesized speech utterances are to the spectra of recorded ones. The Log Spectral Distances D_{LS} are tabulated in Table 2 .

Table 2. Log Spectral Distancebetween recorded and synthesized utterances

Word	**Log Spectral Distance DLS**	
	Modified Klatt type synthesizer	**eSpeak**
/huLi/ ⊠⊠⊠⊠	2.4763	2.8275
/maavu/ ⊠⊠⊠⊠	1.5380	1.9529
/kaage/ ⊠⊠⊠⊠	2.0322	2.5132
/ravi/ ⊠⊠⊠	1.8903	2.1721
/ruupaayi/ ⊠⊠⊠⊠⊠⊠	1.6153	1.9338

Comparing recorded and synthesized utterances using Log Spectral Distance always yielded a D_{LS} value that is greater than 0. This is due to the fact that, both the synthesizers use parameters that are obtained from different set of speakers and sound differently compared to recorded reference utterances. But it can be noted that the D_{LS} values of the modified Klatt type synthesized utterances are less than that of eSpeak indicating that the spectra of utterances synthesized by modified Klatt type synthesizer match more closely to recorded sound utterances than that of eSpeak.

The eSpeak tool uses control parameters that were originally extracted for synthesizing non-Indic languages like English, French, German etc., for synthesizing Kannada utterances leading to spectral mismatch. However, the modified Klatt type formant synthesizer use control parameters extracted from utterances of native Kannada speakers leading to generation of spectra closer to natural utterances.

3.2 Mean Opinion Score (MOS)

Judging the degree of intelligibility and quality of the synthesized speech utterance is a subjective task and for this purpose Mean Opinion Score (MOS) was calculated for each of the synthesized utterance. Intelligibility is defined as certainty with which the listener can correctly identify the utterance and quality is defined as degree of naturalness of utterance. 10 native Kannada language speakers participated in the MOS survey and they were made aware of the definition of intelligibly and quality apriori.

Intelligibility of the synthesized word utterances was measured using Rhyme Test. In this test, the listeners were made to listen to the utterances and choose its equivalent from two visually presented words. The two visually presented words differ only in a single distinctive acousticphonetic feature. Table 3 shows homophones words used for each utterance for performing Rhyme Test. The MOS for intelligibility for each word utterance is calculated as percentage of listeners correctly identifying the word utterance.

Quality of the synthesized word utterances was measured using 5-point rating scale (Excellent – 5, Good – 4, Fair – 3, Poor – 2 and Bad - 1). The listeners rated each word utterance after listening to them. The MOS for quality were then calculated using Equation 4.

$$MOS = \frac{\sum_{i=1}^{N} R_i}{N} \tag{4}$$

Where R_i is the rating given by i^{th} person and N is the total number of people participating in MOS survey.

Table 3. Homophones used for Rhyme test

Actual Words Used		**Homophones used for Rhyme Test**	
Word	**Meaning**	**Word**	**Meaning**
/huLi/ ⊠⊠⊠⊠	Sour	**/huli/ ⊠⊠⊠⊠**	Tiger
/maavu/ ⊠⊠⊠⊠	Mango	**/naavu/ ⊠⊠⊠⊠**	We
/kaage/ ⊠⊠⊠⊠	Crow	**/haage/ ⊠⊠⊠⊠**	Such, Like that
/ravi/ ⊠⊠⊠	Sun, also used as a proper noun	**/savi/ ⊠⊠⊠**	To taste
/ruupaayi/ ⊠⊠⊠⊠⊠⊠	Rupee	**/pipaayi/ ⊠⊠⊠⊠⊠**	Barrel

The calculated MOS is tabulated in Table 4 and it can be inferred that the intelligibility of the utterances produced by the modified Klatt type synthesizer is more than that of the utterances produced by eSpeak. However, a few listeners found it difficult to identify the utterance /huLi/ (⊠⊠⊠⊠)due to the presence of voiced retroflex lateral approximant /L/ (⊠⊠)and got /L/ (⊠⊠) confused with voiced alveolar lateral approximant /l/ (⊠⊠)and misidentified /huLi/ (⊠⊠⊠⊠)meaning sour as /huli/ (⊠⊠⊠⊠)meaning tiger. Those participants were able to identify the utterance correctly only after a second listening resulting in low score for both intelligibility and quality. Overall, the quality of utterances produced by the modified Klatt type formant synthesizer is more compared to the utterances produced by eSpeak.

Table 4. Mean Opinion Score (MOS)

Word	**Mean Opinion Score (MOS)**			
	Modified Klatt type synthesizer		**eSpeak**	
	Intelligibility	**Quality**	**Intelligibility**	**Quality**
/huLi/ ⊠⊠⊠⊠	7/10	3.4	5/10	3.0
/maavu/ ⊠⊠⊠⊠	9/10	4.7	8/10	4.5
/kaage/ ⊠⊠⊠⊠	8/10	4.5	7/10	4.3
/ravi/ ⊠⊠⊠	9/10	4.9	7/10	4.4
/ruupaayi/ ⊠⊠⊠⊠⊠⊠	10/10	4.8	8/10	4.5
Average	**86%**	**4.46**	**70%**	**4.14**

The reason for better performance of the modified Klatt type formant synthesizer in terms of intelligibility and quality over eSpeak is the way in which the control parameters are stored. The modified Klatt type formant synthesizer captures the variation of formants over time using polynomial curves unlike eSpeak which uses either constant formants or straight line contours leading to production of robotic speech.

4 Conclusion

The basic sound units viz., vowels, diphthongs and CV coarticulations were synthesized using eSpeak and modified Klatt type formant synthesizers. It was found that the modified Klatt type formant synthesizer produced vowels and diphthongs that are closer to natural utterances than eSpeak. However, in CV co-articulations the modified Klatt type formant synthesizer produced 2 ambiguous sounding approximants and eSpeak produced 4 ambiguous sounding plosives. Overall, the spectrogram produced by the modified Klatt type formant synthesizer had lesser striped appearance compared to eSpeak.

Five sample word utterances were also synthesized using eSpeak and the modified Klatt type formant synthesizer. The spectra of synthesized utterances were compared using Log Spectral Distance with the spectra of recorded reference utterances. The spectra of utterances produced by the modified Klatt type synthesizer were found to be closer to the spectra of recorded reference utterances than spectra of those utterances synthesized by eSpeak. Since, the quality of the synthesized speech utterance cannot be easily quantized, a MOS was also obtained from 10 Kannada native speakers for intelligibility and quality. The average intelligibility of 5 utterances was 86% for modified Klatt type synthesizer and 70% for eSpeak. Also the average quality of 5 utterances was 4.46 for modified Klatt type synthesizer and 4.14 for eSpeak on a 5-point rating scale. In both the cases, the modified Klatt type synthesizer outperformed eSpeak. Also, the successful synthesis of these sample word utterances also demonstrates the capability of the modified Klatt type formant synthesizer to synthesize various sound types of Kannada language.

Acknowledgement

This work was supported by Department of Electronics and Communication Engineering, Vidyavardhaka College of Engineering (VVCE), Mysuru and Visvesvaraya Technological University (VTU), Belagavi, Karnataka.

References

1) Koffi E, Petzold M. A Tutorial on Formant-based Speech Synthesis for the Documentation of Critically Endangered Languages. 2022. Available from: https://repository.stcloudstate.edu/stcloud_ling/vol11/iss1/3.
2) Trivedi A, Pant N, Shah P, Sonik S, Agrawal S. Speech to text and text to speech recognition systems-A review. *IOSR Journal of Computer Engineering*. 2018;20(2):36–43. Available from: https://www.iosrjournals.org/iosr-jce/papers/Vol20-issue2/Version-1/E2002013643.pdf.
3) Dutonde SK, Mapari GS, Wagh SJ, Kapse A. Review on Text to Speech Synthesizer. *International Journal of Advance Research and Innovative Ideas in Education*. 2022;8(3):592–596. Available from: https://ijariie.com/AdminUploadPdf/Review_on_Text_to_Speech_Synthesizer_ijariie16614.pdf.
4) Li X, Ma D, Yin B. Advance research in agricultural text-to-speech: the word segmentation of analytic language and the deep learning-based end-to-end system. *Computers and Electronics in Agriculture*. 2021;180:105908. Available from: https://doi.org/10.1016/j.compag.2020.105908.
5) Tan X, Qin T, Soong F, Liu TY. A survey on neural speech synthesis. 2021. Available from: https://arxiv.org/pdf/2106.15561.pdf.
6) Kuligowska K, Kisielewicz P, Włodarz A. Speech synthesis systems: disadvantages and limitations. *International Journal of Engineering & Technology*. 2018;7(2.28):234–234. Available from: https://doi.org/10.14419/ijet.v7i2.28.12933.
7) Sen A. Speech Synthesis in India. 2007. Available from: https://www.tandfonline.com/doi/abs/10.4103/02564602.10876616.
8) Panda SP, Nayak AK, Rai SC. A survey on speech synthesis techniques in Indian languages. *Multimedia Systems*. 2020;26(4):453–478. Available from: https://doi.org/10.1007/s00530-020-00659-4.
9) Hillenbrand JM. The acoustics and perception of North American English vowels. 2019. Available from: https://www.taylorfrancis.com/chapters/edit/10.4324/9780429056253-10/acoustics-perception-north-american-english-vowels-james-hillenbrand.
10) Lukose S, Upadhya SS. Text to speech synthesizer-formant synthesis. *2017 International Conference on Nascent Technologies in Engineering (ICNTE)*. 2017;p. 1–4. Available from: https://doi.org/10.1109/ICNTE.2017.7947945.
11) D'souza AV, Ravi DJ. An Approach for Formant Synthesis of Kannada. *Journal of Signal Processing*. 2022;8(2):31–38. Available from: https://doi.org/10.46610/JOSP.2022.v08i02.006.
12) Duddington J, Dunn R. eSpeak text to speech. . Available from: http://espeak.sourceforge.net.2012.http://espeak.sourceforge.net.