

# Performance analysis of different intonation models in Kannada speech synthesis

Sadashiva Veerappa Chakrasali<sup>1,2</sup>, Krishnappa Indira<sup>1</sup>, Sunitha Yariyur Narasimhaiah<sup>3</sup>,  
Shadaksharaiah Chandraiah<sup>4</sup>

<sup>1</sup>Department of Electronics & Communication Engineering, Ramaiah Institute of Technology, Bengaluru, India

<sup>2</sup>Department of Electronics & Communication Engineering, VTU Research Center, Ramaiah Institute of Technology, Bengaluru, India

<sup>3</sup>Department of Electronics & Communication Engineering, SJB Institute of Technology, Bengaluru, India

<sup>4</sup>Department of Electrical & Electronics Engineering, Bapuji Institute of Engineering and Technology, Davangere, India

## Article Info

### Article history:

Received Mar 23, 2021

Revised Jan 10, 2022

Accepted Feb 21, 2022

### Keywords:

CART model  
Fujisaki parameters  
Intonation models  
Kannada TTS  
Neural network  
Pitch frequency  
Tilt model

## ABSTRACT

Text to speech (TTS) is a system that generates artificial speech from text input. The prosodic models used improve the quality of the synthesized speech especially naturalness and intelligibility. The prosody involves intonation, intonation refers to the variations in the pitch frequency (F0) with respect to time in an utterance. This work mainly concentrates on building feedback neural network model to predict F0 contour in the utterances using Fujisaki intonation model parameters as the input features to the network since the Fujisaki intonation model is data driven and not a rule based one. In this work we have built 4-layer feedback neural network in the festival framework. Finally, the synthetically generated Kannada speech using the neural network model, is compared for its performance with the classification and regression tree (CART) model and Tilt model. Database of simple declarative Kannada sentences created by Carnegie Mellon University have been deployed in this work. From the study it is very clear that F0 contours can be accurately predicted using CART and neural network models, whereas naturalness and intelligibility is high in CART model rather than neural network model.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



## Corresponding Author:

Sadashiva Veerappa Chakrasali  
Department of Electronics & Communication Engineering, VTU Research Center  
Ramaiah Institute of Technology  
Bengaluru, India  
Email: sadashivavc@gmail.com

## 1. INTRODUCTION

Recent developments in voice synthesis are huge in foreign languages, but they are almost non-existent in Indian languages like Kannada. In a previous paper [1], we created a Kannada speech synthesiser based on the hidden markov model (HMM) and the festival framework for simple declarative sentences, with prosody generated using the Tilt model. The developed Kannada speech synthesizer was able to produce a speech with m-nary-coded-decimal (MCD) score ranging between 3.5 dB to 5 dB, which is considered to be a good synthesizer but the generated speech is unsteady and unnatural sounding. Mixdorff and others [2], [3] have demonstrated that integrating prosody models improves the quality (naturalness and intelligibility) of synthetic speech. Prosody refers to the duration, gain, and intonation (pitch pattern) of speech portions. In the case of spoken speech sounds, intonation provides information on the glottal pulse source's periodicity. Depending on the nature or style of speech, different utterances for the same phoneme segment may have different intonation patterns. To forecast intermediate peaks and valleys, Madhukumar *et al.* [4] have made

available a model for linear intonation for the Hindi language for common sentences which are declarative in nature and its content limited to most commonly used functional words. This model fails to capture abrupt changes in pitch. In the paper, Patel *et al.* [5] have reported that the naturalness of the synthesized language can be increased by using proper models for the variations in the pitch, its rise and fall in the phrases and accent used in the language synthesized. Fujisaki and others [6], [7] have demonstrated the improvement in the synthesized speech in Japanese language by incorporating prosodic model. Details on Fujisaki model and extraction of model parameters are available in [7]–[11]. In the paper, Mnasri *et al.* [12] have built a neural network model to predict pitch frequency using Fujisaki parameters to synthesize Arabic language. In the paper, Rao and Yegnanarayana [13] has developed a neural network model to predict F0 of a syllable for Telugu and Bengali languages by feeding the attributes obtained directly from the utterances. In this work, Fujisaki Intonation Model features are used as input attributes of the neural network (NN) Model to predict the pitch of the phoneme in an utterance. Section 2, explains Fujisaki model and extraction of its parameters, whereas neural network model built in this work along with input and output features is explained in section 3. Section 4 explains the performance of neural network model and its performance analysis in synthesis.

## 2. FUJISAKI MODEL

From the literature it is evident that the Fujisaki model is widely used in intonation modeling, the main reason to use this is that it is a data driven method which is independent of language. The Fujisaki model provides a high-accuracy method for generating fundamental frequency (F0) changes in natural speech. The supplied pitch contour is decomposed into various components for parameters like Fb, the base frequency, the language phrases commonly used and the accent in which the words and phrases are pronounced. These three parameters of F0 are compared by superimposing their contours generated on a log scale. The Figure 1 displays the contour generated on the log scale for the pitch F0. The components or the phrases of the language are obtained as a result of the impulse reaction caused due to exposure of the linear system to critical damping to the second order. The accent components were seen to develop when the above dampening system was actuated by accent commands which were in the form of rectangular pulses, varying in length and amplitude. When the value of the speech which is constant is superimposed on two different components of the same utterance, it is seen to result in a basic model of the pitch contour relating to the utterance.

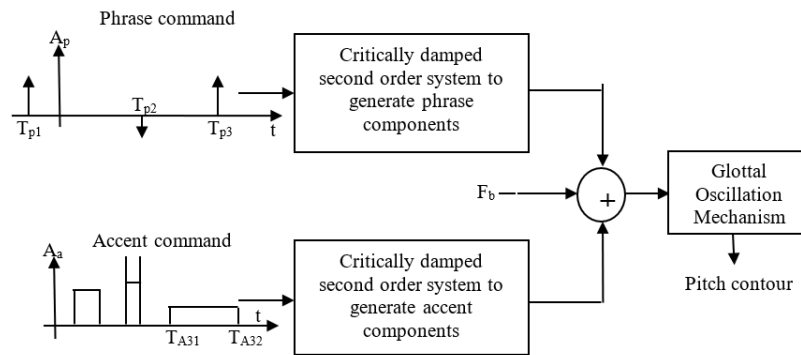


Figure 1. Block diagram for pitch contour generation

$$\ln(F_0(t)) = \ln(F_b) + \sum_{i=1}^I A_{pi} G_{pi}(t - T_{pi}) + \sum_{j=1}^J A_{aj} \{G_{aj}(t - T_{aj1}) - G_{aj}(t - T_{aj2})\}$$

where,

$$G_{pi}(t) = \begin{cases} \alpha_i^2 t \exp(-\alpha_i t), & t \geq 0 \\ 0, & t < 0 \end{cases}$$

$$G_{aj}(t) = \begin{cases} \min [1 - (1 + \beta_j t) \exp(-\beta_j t), \theta_j], & t \geq 0 \\ 0 & t < 0 \end{cases}$$

F<sub>b</sub>: represents the utterance base frequency

I: represents the quantity of phrase commands available in any utterance

J: represents the quantity of accent commands available in any utterance  
 $A_{pi}$  and  $A_{aj}$ : respectively represent the amplitude of the  $i^{th}$  phrase command and accent command in the utterance  
 $T_{pi}$ : represents the instant of occurrence of the  $i^{th}$  phrase command available in the utterance  
 $T_{aj1}$  and  $T_{aj2}$ : respectively represent the Starting Time and End Time of the  $j^{th}$  accent command available in the utterance  
 $\alpha_i$ : represents the natural angular frequency of phrase control mechanism of  $i^{th}$  phrase command  
 $\beta_j$ : represents the natural angular frequency of accent control mechanism of  $j^{th}$  accent command

**2.1. Parameter extraction**

When a language is spoken by different individuals there is a quite naturally a marginal difference in the contours with regard to the pitch at a micro level. The individual prosodic changes caused by different individuals at the micro level need to be ironed out or eliminated to arrive at a pitch contour specific to a particular utterance and phrase. Only then it would be possible to extract the parameters of the Fujisaki model. This is possible by using an interpolation technique known as Cubic Spline Interpolation [6], [8], [9]. This interpolation creates a continuous pitch contour by removing silence portions from the speech, which is very essential to take derivatives of interpolated pitch contour. By taking derivatives we get high frequency contours (HFC) and low frequency contour (LFC). LFC is obtained by deducting the HFC from the obtained contour for interpolated pitch (IC). The IC is processed through a high pass filter with a cut off frequency of 0.5 Hz. to separate the accent and phrase components. As shown in Figure 2(a)-(d), the LFC is produced by subtracting HFC from the interpolated contour. The approximation derivative is used to HFC to extract the accent commands, i.e. the accent commands are analysed between successive minima in HFC. The largest F0 in the Log scale is obtained based on the value of Aa, the accent command maxima. Fb the base frequency is revealed by the LFC Global minimum. In view of the fact that the commencement of any new phase is distinctly marked by a local minima in that phrase component, the phrase components are spaced 1 second apart and the LFC searched for local minimas. The segment of the LFC after the potential onset time  $T_p$  is searched for the next local maximum to initialize the magnitude value of  $A_p$  assigned to each phrase command. At this stage,  $A_p$  is determined in proportion to F0, taking into account the contributions of previous orders. 1/S, 20/S, and 0.9 are the values for the remaining model parameters [11].

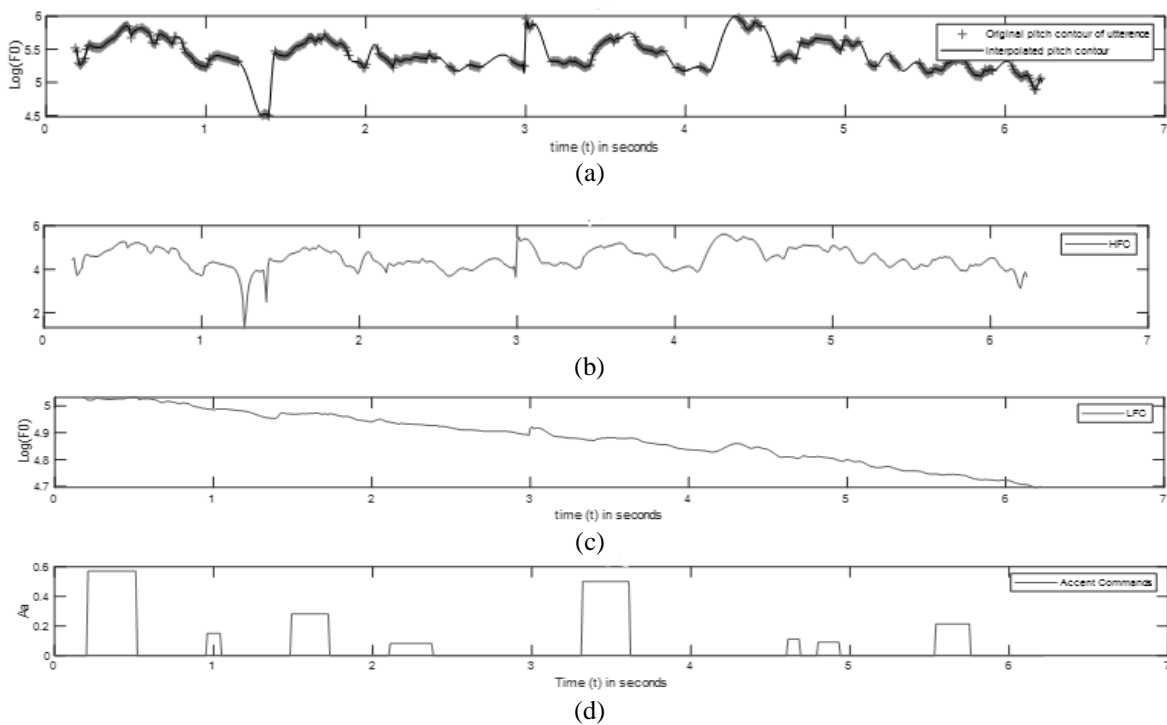


Figure 2. Extraction of the Fujisaki Parameters in an utterance, (a) cubic interpolation of pitch contour, (b) HFC, (c) LFC of interpolated pitch contour, and (d) extracted phrase and accent commands of an utterance

### 3. NEURAL NETWORK MODEL

Neural network (NN) model is nothing but the mutual functional relation between the input and output. This study uses a four layered feedback neural network (FBNN) shown in Figure 3, which is used to predict the average F0 [14], [15] value for each phoneme in an utterance. The positional, contextual and phonological features are used to train the neural network. There are 23 input feature vectors and 1 output feature vector. The list of features affecting the F0 of a phoneme are given in Table 1 and Table 2 respectively. The first layer in the four-layer neural network is the input layer with 23 input feature vectors and the first layer is having linear activation function. The middle two layers are the hidden layers with sigmoid as an activation function. The third layer is having a single node whereas the performance of the network is analyzed by varying number of nodes in the second layer. The last or the fourth layer is the output layer with 1 output feature vector that is a non-linear unit. The back-propagation method is used in order to achieve minimum mean square error (MSE).

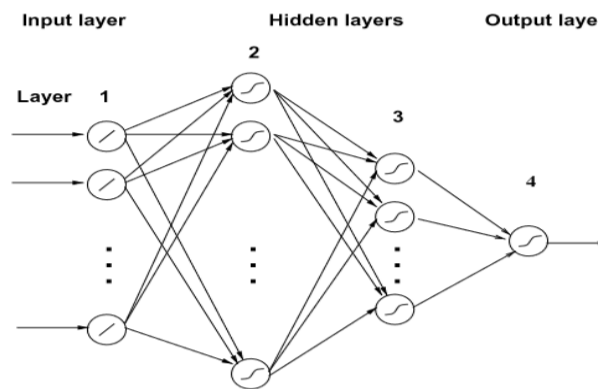


Figure 3. Four-layer FBNN

#### 3.1. Training of neural network

In this work, we used a four layered FBNN shown in Figure 3, which consists of one input layer, two hidden layers and one output layer. There are 23 input feature vectors and 1 output feature vector considered that are fed to train the network. The input features are based on the positional, contextual and phonological constraints of a phoneme and the output feature vector is the pitch of the respective phoneme as shown in Table 1.

Factors	Input attributes	No. of Nodes
Phoneme position in the word	Position of phoneme from the beginning of the word	3
	Position of phoneme from the end of the word	
	Number of phonemes in the word	
Phoneme position in the phrase	Position of phoneme from the beginning of the phrase	3
	Position of phoneme from the end of the phrase	
	Number of phonemes in the phrase	
Word position in the phrase	Position of the word (corresponding to required phoneme) from the beginning of the phrase	3
	Position of the word from the end of the phrase	
	Number of words in the phrase	
Context of phoneme	Present phoneme	3
	Previous phoneme	
	Successive phoneme	
	Phrase command amplitude	
	Accent command amplitude ( $A_a$ ) at Present phoneme	
Fujisaki model parameters	Duration of the accent command at present phoneme	9
	Accent command amplitude at Previous phoneme	
	Duration of the accent command at previous phoneme	
	Accent command amplitude at successive phoneme	
	Duration of the accent command at successive phoneme	
Pitch	Base frequency of the utterance	2
	Number of accent commands in the utterance	
	Pitch of the previous phoneme	
	Pitch of the successive phoneme	

For an example the utterance corresponding to the text “ಶ್ರೀ ಅವರ ಜೊತೆಗೆ ದುಡಿದ ಗೆಳೆಯರು ಪ್ರೊಫೆಸರ್ ನರಸಿಂಹಾಚಾರ್ ಕಸ್ತೂರಿ ರಂಗಾಚಾರ್ ಮುಂತಾದವರುಗಳು“, tabulation is prepared to depict the input to the neural network. Table 2, indicates part of that Table 3, indicates the codepoints of Kannada vowels along with its corresponding English transliteration. The details of codepoints and the corresponding transliteration for the consonants in Kannada is given in appendix. Out of 696 simple Kannada declarative utterances taken from Carnegie Mellon University’s Indic data base, in this work we have used 130 utterances to train the neural network with 70% training and 15% testing and 15% validation. The detailed discussion of performance of this network in prediction of F0 is discussed in next section by varying number of neurons in layer 2. Then we measure the performance of the proposed network in speech synthesis in comparison to Tilt and Cart model.

Table 2. Input and output attributes for the words “\” ಶ್ರೀ ಯವರ “\”

Pho neme	Phoneme position in the word	Phoneme position in the phrase	Word position in the phrase	Context of phoneme	Fujisaki Model Parameters	Pitch of previous and succeeding phoneme	Pitch of current phoneme
C}	1 3 3	1 78 78	1 10 10	3254 3200 3248	0.4 0 0 0 0 0.57 0.035 110 8	0 270 0	0 327 270
9r	2 2 3	2 77 78	1 10 10	3248 3254 3208	0.4 0.57 0.035 0 0 0.57 0.185 110 8	0 327 270	315 327
l:	3 1 3	3 76 78	1 10 10	3208 3248 3247	0.4 0.57 0.185 0.57 0.035 0.57 0.015 110 8	327 300 315	300 315
j	1 6 6	4 75 78	2 9 10	3247 3208 3205	0.4 0.57 0.015 0.57 0.185 0.57 0.085 110 8	315 300 322	260 300
a	2 5 6	5 74 78	2 9 10	3205 3247 3253	0.4 0.57 0.085 0.57 0.015 0 0 110 8	322 260 300	260 280
v	3 4 6	6 73 78	2 9 10	3253 3205 3205	0.4 0 0 0.57 0.085 0.2 0.1 110 8	300 260 280	270 260
a	4 3 6	7 72 78	2 9 10	3205 3253 3248	0.4 0.2 0.1 0 0 0 0 110 8	280 270 260	265 270
9r	5 2 6	8 71 78	2 9 10	3248 3205 3205	0.4 0 0 0.2 0.02 0.26 0.04 110 8	260 265 270	
a	6 1 6	9 70 78	2 9 10	3205 3248 3228	0.4 0.26 0.08 0.2 0.02 0 0 110 8		

Table 3. Vowels and their code points with English transliteration

Kannada Grapheme	ಅ	ಆ	ಇ	ಈ	ಉ	ಊ	ಋ
Codepoint	3205	3206	3207	3208	3209	3210	3211
English utterance	A	A:	i	i:	u	u:	9r
Kannada Grapheme	ಎ	ಏ	ಐ	ಒ	ಓ	ಔ	ಌ
Codepoint	3214	3215	3216	3218	3219	3220	3202
English utterance	e	e:	aI	o	o:	aU	n

#### 4. RESULTS AND DISCUSSION

##### 4.1. Prediction of F0 by neural network model

The training of the neural network is carried to predict the pitch frequency of the phoneme. The performance of the network for the intended work is measured with different number of neurons in the hidden layer 1 (layer 2 in entire network). From the following regression graphs shown in Figure 4, representing correlation coefficient (R) for training, testing and validation. It is very clear that network is well performing when the network is having 5 neurons in the hidden layer 1. Comparative analysis with different number of neurons is indicated in Table 4.

Table 4. Performance analysis of proposed neural network with different neurons in hidden layer 1

Number of Neurons in hidden layer 1	Correlation Coefficient		
	Training	Validation	Testing
5	0.972	0.857	0.808
10	0.925	0.584	0.353
15	0.584	0.185	0.192
20	0.938	0.696	0.519
25	0.741	0.534	0.610
30	1	0.638	0.405

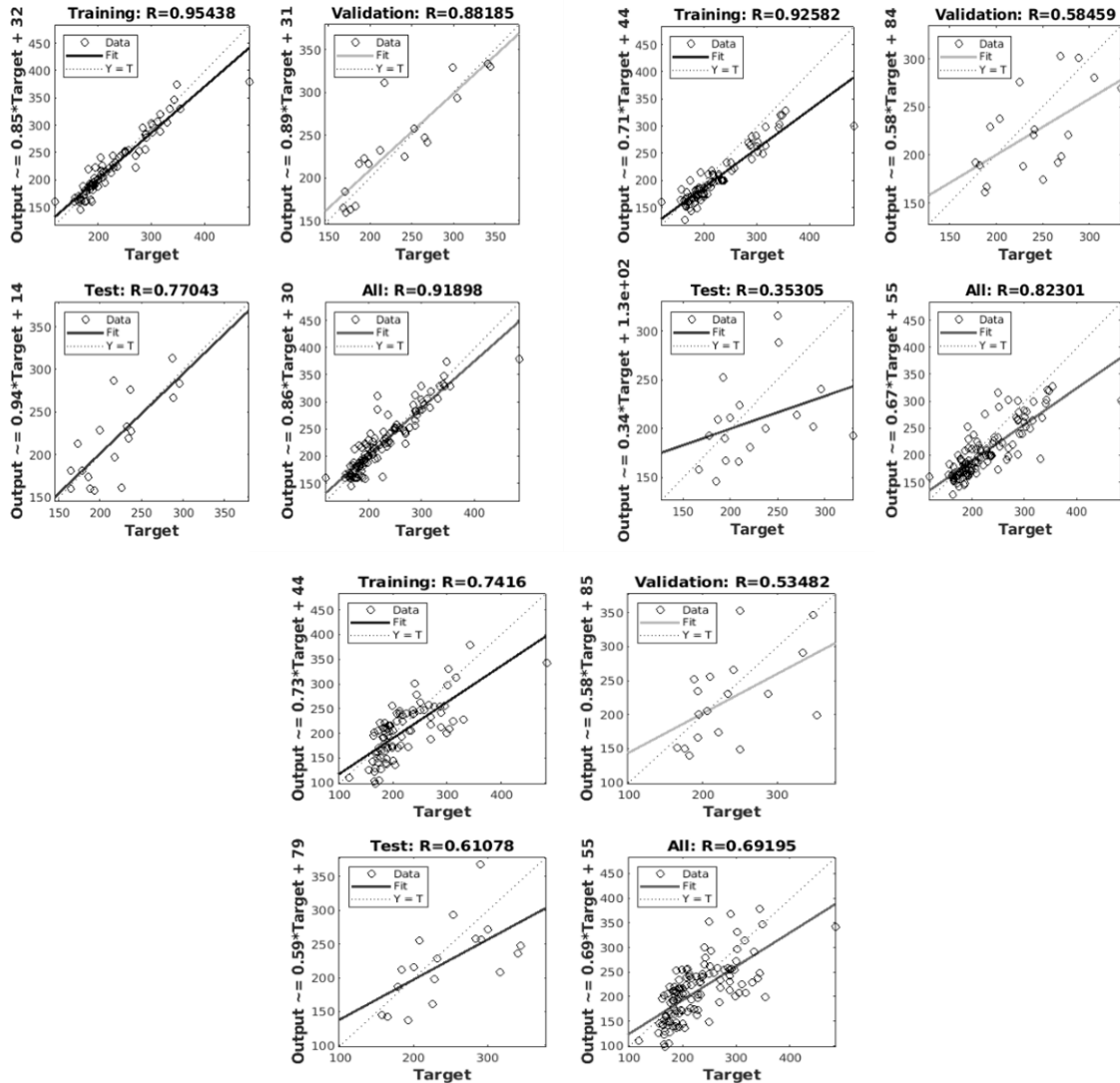
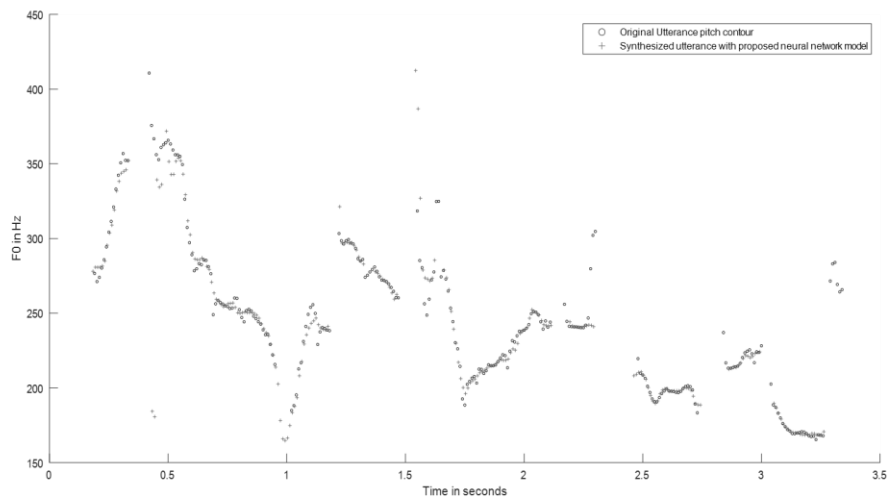


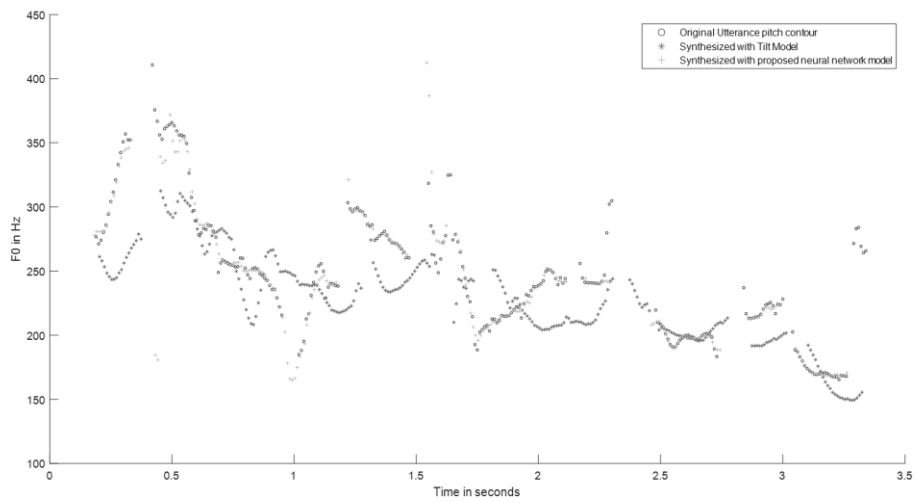
Figure 4. Regression plots of trained neural network with hidden neurons=5, 10 and 25

#### 4.2. Synthesis and MCD score analysis

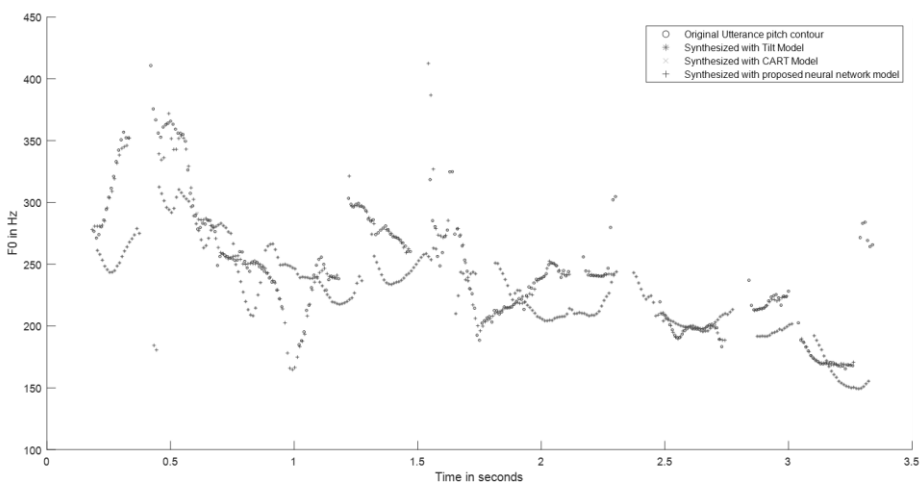
Synthesis of simple Kannada declarative sentences are synthesized using festival framework [16]–[20] by considering three different prosodic models, namely Tilt [21], [22], classification and regression tree (CART) [23]–[25] and proposed neural network model. In this work we have done qualitative analysis. Qualitative analysis involves subjective analysis in which persons of different age group are asked to listen original and synthesized utterances and noted their opinions, which is indicated in Table 5, whereas another qualitative analysis involves plotting the F0 contours of original and synthesized utterances. From the observations of pitch contours of the synthesized utterances with Tilt model is having high variations in comparison to original utterance, whereas CART model developed based on Fujisaki parameters is matching very close with original utterance. This is also reflected in another qualitative analysis done using MCD score. It is very clear that the pitch contours of the synthesized speech with the proposed neural network model is highly matching with the pitch contours of the original utterance. The comparison of the F0 contours of the synthesized utterances using different models are shown in Figures 5(a)–(c). But the proposed model is lacking in naturalness and intelligibility though the F0 contour is almost following the original one. The main reason for poor performance of proposed neural network model is due to, the model is only able to predict pitch frequency of the phoneme but unable to predict duration of the phoneme, whereas CART model with Fujisaki parameters is very accurate in predicting F0 and duration both.



(a)



(b)



(c)

Figure 5. Comparison of pitch contour of original utterance and synthesized utterance, (a) with Tilt model, (b) CART model, and (c) proposed neural network model

Table 5. Subjective analysis of synthesized utterances

Person	Synthesized utterances														
	Tilt model					CART Model					neural network model				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
1	1	1	2	1	1	4	4	4	4	4	2	2	2	2	2
2	1	1	1	1	1	3	4	4	4	4	2	2	2	2	2
3	1	1	1	1	1	4	4	4	4	4	1	1	1	2	1
4	2	1	2	1	1	4	4	4	4	4	2	2	2	2	2
5	1	1	1	1	1	4	4	4	3	4	2	2	2	1	2
6	1	2	1	1	1	4	3	4	4	4	2	2	2	2	2
7	1	1	1	1	1	4	4	4	4	4	2	1	1	2	1
8	1	1	1	2	1	4	4	4	4	4	1	1	2	2	1
9	1	1	1	1	1	4	4	4	4	4	1	2	1	1	2
10	1	1	1	1	1	4	4	4	4	4	2	2	2	2	2

1. Poor, 2. Average, 3. Good, and 4. Very good

Mel cepstral distortion (MCD score): A vital entity that is very much necessary in deciding the quality of the speech which has been synthesized is the MCD Score. The mathematical formal that is commonly used to arrive at the MCD is as shown in:

$$MCD = \frac{10\sqrt{2}}{\ln 10} \sqrt{\sum_d (mc_{(d)}^{(t)} - mc_{(d)}^{(e)})^2}$$

where,

$mc_{(d)}^{(t)}$  : Mel - Cepstral parameters of training utterance

$mc_{(d)}^{(e)}$  : Mel - Cepstral parameters of synthesized utterance

d : index of Mel - Cepstral parameters array

Since MCD is the yardstick to measure the proximity of the synthesized speech to the actual speech, it is normally considered that a MCD value between 4.5 to 6 dB is acceptable. In a work carried out earlier by us where synthesizing using the Tilt model was carried out, we were able to achieve MCD value between 3.52 and 5.02 dB. Despite the fact that value could be considered fairly good, we found that in the Tilt method the synthesized speech was rather unnatural and shaky. On working with the CART model where the Fujisaki parameters were used for synthesizing speech we achieved MCD values in the range of 1.62 dB and 2.43 dB while a value in the range of 3 to 4.5 dB was achieved applying the Neural Network technique. All these details are shown in Table 6 for different 10 utterances. There is a slight improvement seen in MCD score for synthesis using neural network model, but still this model lacks in generating quality synthesized speech though the F0 contour is very close to original utterances.

Table 6. Performance analysis of different models based on MCD score

Utterance	MCD Score of Synthesized Speech in dB			Utterance	MCD Score of Synthesized Speech in dB		
	Tilt Model	CART Model	NN Model		Tilt Model	CART Model	NN Model
1	3.52	1.68	3.21	1	4.65	1.89	3.19
2	4.16	2.10	3.48	2	4.80	2.37	4.02
3	3.87	1.78	3.07	3	4.21	1.65	3.63
4	4.32	2.43	3.62	4	5.02	2.29	3.94
5	4.34	1.62	3.87	5	4.56	1.67	3.54

## 5. CONCLUSION

From the performance analysis it is very clear that F0 contours of utterances can be predicted with high accuracy in the feedback neural network with 5 neurons in the hidden layer. But prediction of F0 contour alone is not sufficient to generate quality speech as it requires additional factors such as duration models. However neural network model is able to perform better than Tilt model. Naturalness is very high in speech synthesized using CART model built using Fujisaki parameters.

## REFERENCES




- [1] S. V. Chakrasali, K. Indira, S. B. Sharma, N. M. Srinivas, and S. S. Varun, "HMM based Kannada speech synthesis using festvox," *International Journal of Recent Technology and Engineering*, vol. 8, no. 3, pp. 2635–2639, Sep. 2019, doi: 10.35940/ijrte.C4934.098319.






- [2] H. Mixdorff and D. Mehnert, "Exploring the naturalness of several German high-quality-text-to-speech systems.," *Eurospeech*, 1999, Accessed: Feb. 17, 2022. [Online]. Available: [http://public.beuth-hochschule.de/~mixdorff/thesis/files/mixdorff\\_mehnert\\_eurosp1999.pdf](http://public.beuth-hochschule.de/~mixdorff/thesis/files/mixdorff_mehnert_eurosp1999.pdf)
- [3] H. Mixdorff and O. Jokisch, "Evaluating the quality of an integrated model of German prosody," *International Journal of Speech Technology*, vol. 6, no. 1, pp. 45–55, 2003, doi: 10.1023/A:1021099922328.
- [4] A. S. Madhukumar, S. Rajendran, and B. Yegnanarayana, "Intonation component of a text-to-speech system for hindi," *Computer Speech and Language*, vol. 7, no. 3, pp. 283–301, Jul. 1993, doi: 10.1006/csla.1993.1015.
- [5] T. B. Patel and H. A. Patil, "Analysis of natural and synthetic speech using Fujisaki model," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, Mar. 2016, vol. 2016-May, pp. 5250–5254, doi: 10.1109/ICASSP.2016.7472679.
- [6] K. Hirose, *Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis*, February. Springer Berlin Heidelberg, 2015.
- [7] H. Fujisaki and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *Journal of the Acoustical Society of Japan (E) (English translation of Nippon Onkyo Gakkaishi)*, vol. 5, no. 4, pp. 233–242, 1984, doi: 10.1250/ast.5.233.
- [8] S. Narusawa, N. Minematsu, K. Hirose, and H. Fujisaki, "Automatic extraction of model parameters from fundamental frequency contours of English utterances," *7th International Conference on Spoken Language Processing, ICSLP 2002*, pp. 1725–1728, 2002, Accessed: Feb. 17, 2022. [Online]. Available: <http://www.iscaaspeech.org/archive>
- [9] H. Mixdorff, "A novel approach to the fully automatic extraction of Fujisaki model parameters," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing-Proceedings*, 2000, vol. 3, pp. 1281–1284, doi: 10.1109/ICASSP.2000.861811.
- [10] H. Zen *et al.*, "The HMM-based speech synthesis system (HTS) Version 2.0," *6th ISCA Workshop on Speech Synthesis*.
- [11] S. Chakrasali, S. Y. N, and C. M. Patil, "Performance analysis of fujisaki intonation model in Kannada speech synthesis," *Turkish Journal of Physiotherapy and Rehabilitation*, vol. 32, no. 3, Accessed: Feb. 17, 2022. [Online]. Available: [www.turkjphysiotherrehabil.org](http://www.turkjphysiotherrehabil.org)
- [12] Z. Mnasri, F. Boukadida, and N. Ellouze, "F<sub>0</sub> contour parametric modeling using multivariate adaptive regression splines for arabic text-to-speech synthesis," in *Eighth International Multi-Conference on Systems, Signals & Devices*, Mar. 2011, pp. 1–6, doi: 10.1109/ssd.2011.5981479.
- [13] K. Sreenivasa Rao and B. Yegnanarayana, "Intonation modeling for Indian languages," *Computer Speech and Language*, vol. 23, no. 2, pp. 240–256, Apr. 2009, doi: 10.1016/j.csl.2008.06.005.
- [14] T. F. Quatieri, "Discrete-time speech signal processing: principles and practice," p. 816, 2001, Accessed: Feb. 17, 2022. [Online]. Available: <http://www.amazon.com/Discrete-Time-Speech-Signal-Processing-Principles/dp/013242942X>
- [15] D. O'Shaughnessy, *Speech Communications: Human and Machine*, 2nd Editio. University Press, 2004.
- [16] S. King, "An introduction to statistical parametric speech synthesis," *Sadhana - Academy Proceedings in Engineering Sciences*, vol. 36, no. 5, pp. 837–852, Oct. 2011, doi: 10.1007/s12046-011-0048-y.
- [17] H. Zen, H. Zen, H. Zen, M. J. F. Gales, Y. Nankaku, and K. Tokuda, "Product of experts for statistical parametric speech synthesis," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 3, pp. 794–805, Mar. 2012, doi: 10.1109/TASL.2011.2165280.
- [18] T. Koriyama, T. Nose, and T. Kobayashi, "Statistical parametric speech synthesis based on gaussian process regression," *IEEE Journal on Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 173–183, Apr. 2014, doi: 10.1109/JSTSP.2013.2283461.
- [19] J. Tao, K. Hirose, K. Tokuda, A. W. Black, and S. King, "Introduction to the issue on statistical parametric speech synthesis," *IEEE Journal on Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 170–172, Apr. 2014, doi: 10.1109/JSTSP.2014.2309416.
- [20] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, May 2013, doi: 10.1109/JPROC.2013.2251852.
- [21] J. Yamagishi, "An introduction to HMM-based speech synthesis," *October*, no. October, 2006, Accessed: Feb. 17, 2022. [Online]. Available: <https://wiki.inf.ed.ac.uk/wiki/pub/CSTR/TrajectoryModelling/HTS-Introduction.pdf>
- [22] H. G. Assistant, "Acoustic phonetic characteristics of Kannada language," *International Journal of Computer Science Issues*, vol. 8, no. 6, pp. 332–339, 2011, Accessed: Feb. 17, 2022. [Online]. Available: [www.IJCSI.org](http://www.IJCSI.org)
- [23] P. Bhaskararao, "Salient phonetic features of Indian languages in speech technology," *Sadhana - Academy Proceedings in Engineering Sciences*, vol. 36, no. 5, pp. 587–599, Oct. 2011, doi: 10.1007/s12046-011-0039-z.
- [24] Z. H. Ling *et al.*, "Deep learning for acoustic modeling in parametric speech generation: a systematic review of existing techniques and future trends," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 35–52, May 2015, doi: 10.1109/MSP.2014.2359987.
- [25] S. Imai, K. Sumita, and C. Furuichi, "Mel log spectrum approximation (MLSA) filter for speech synthesis," *Electronics and Communications in Japan (Part I: Communications)*, vol. 66, no. 2, pp. 10–18, 1983, doi: 10.1002/ecja.4400660203.

## BIOGRAPHIES OF AUTHORS






**Sadashiva Veerappa Chakrasali**    is currently working as an assistant professor in E & C department, M S Ramaiah Institute of Technology, Bangalore. His research interests are statistical signal processing and communication. He has completed BE and M.Tech from Kuvempu university and VTU Belgaum in the year 2000 and 2004 respectively. He can be contacted at email: [sadashiva.c@msrit.edu](mailto:sadashiva.c@msrit.edu).






**Dr. Krishnappa Indira**    is currently working as a Professor in E & C Dept., M S Ramaiah Institute of Technology, Bangalore. Her research interests are in the field of Image and speech processing. She has published several technical papers in peer reviewed journals. She has completed B. E and M.E from Bangalore University in the year 1988 and 1992 respectively. She has done her PhD from VTU, Belagavi in the year 2012. She is a senior member IEEE. Currently she is working on Machine Learning, Neural Networks, Pattern recognition and speech processing. She can be contacted at email: indira@msrit.edu.



**Dr. Sunitha Yariyur Narasimhaiah**    is currently working as an assistant professor in E & C Department, SJB Institute of Technology, Bangalore. Her research interests are Image processing and communication. She has completed B.E. from Mysore University in the year 2000, MTech and PhD from VTU Belgaum in the year 2003 and 2021, respectively. She can be contacted at email: sunithayn@sjbit.edu.in.



**Shadaksharaiah Chandraiah**    is working as an assistant Professor in Electrical and Electronics Engineering department of BIET, Davangere. Prior to this he has worked in Information science department of the same institute. He is a well known programmer in LINUX and C++. His current research are in the area of fault detections in computer networking using Artificial Intelligence techniques. He has done his M. Tech in Computer science and engineering from PDA college of Engineering, Gulbarga in the year 2004. He can be contacted at email: cshadaksharaiah@gmail.com.