Duration Modeling by Multi-Models based on Vowel Production characteristics

V Ramu Reddy
TCS, Innovation Labs,
Kolkata - 700156, West Bengal, India
ramu.vempada@tcs.com

K Sreenivasa Rao and Parakrant Sarkar
Indian Institute of Technology Kharagpur,
Kharagpur - 721302, West Bengal, India
ksrao@iitkgp.ac.in
parakrantsarkar@gmail.com

Abstract

An accurate estimation of segmental durations is needed for natural sounding text-to-speech (TTS) synthesis. This paper propose multi-models based on production aspects of vowels. In this work four multi-models are developed based on vowel length, vowel height, vowel frontness and vowel roundness. In each multi-model, syllables are divided into groups based on specific vowel articulation characteristics. In this study, (i) linguistic constraints represented by positional, contextual and phonological features and (ii) production constraints represented by articulatory features are used for predicting duration patterns. Feed-forward Neural Networks are used for developing duration models. From the results, it was observed that the average prediction error is reduced by 23.21% and correlation coefficient is improved by 9.64% using multi-model developed based on vowel length production characteristics, compared to single duration model.

1 Introduction

Naturalness and intelligibility of the synthetic speech generated by the text-to-speech synthesis (TTS) systems can be improved by means of accurate prediction of prosodic parameters. Prosody refers to duration, intonation and intensity patterns of speech for the sequence of syllables, words and phrases. In this work, we focus on modeling or predicting one of the important prosodic parameters i.e., duration. Duration plays an important role in human speech communication. Duration patterns of an utterance is defined as the sequence of segmental (phone) or supra-segmental (syllable) durations. Variation in duration patterns provide naturalness to speech. Human hearing system is highly sensitive to variations in duration patterns. Hence, while developing speech synthesis systems, acquisition and incorporation of the duration knowledge is very much essential.

In this work, we are modeling the syllable durations for Indian language Bengali. In speech signal, the duration of each unit is dictated by the linguistic and production constraints of the unit (Reddy and Rao, 2012) (Rao and Yegnanarayana, 2007). In (Reddy and Rao, 2012), Ramu *et al* have developed single duration model using linguistic constraints represented by positional, contextual and phonological (PCP) features, and production constraints represented by articulatory (A) features (Reddy and Rao, 2012). From here onward it is referred as PCPA features in the rest of the paper. In most of the existing Indian context TTS works (Kumar and Yegnanarayana, 1989) (Kumar, 1990) (Kumar, 2002) (Krishna and Murthy, 2004) (Rao and Yegnanarayana, 2007) (Kumar et al., 2002) single duration models are developed by considering all the available syllables present in the training dataset irrespective of syllable position or articulation aspects of syllables. The distribution plot of duration of syllables present in our database is shown in Fig. 1. From Fig. 1, it is observed that duration values of syllables in the database vary from 50 to 560 ms with mean and standard deviations 212.9 ms and 80.6 ms, respectively. It is also observed that most of the duration values of syllables are concentrated between 110 ms and 350 ms, respectively. Therefore, the single duration model will be more prone to erroneous by classifying the low and high duration values to-

wards mean values (central tendency) due to less frequency of low and high duration syllables in the training phase. This results in high average prediction error.

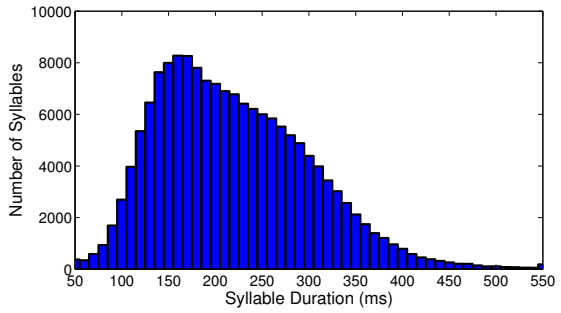

Figure 1: Distribution plot of Syllable durations

To improve the prediction accuracy by eliminating the biases of short durations of syllables towards long durations of syllables and vice-versa, Rao *et al* have developed two-stage duration model (Rao, 2005) (Rao and B.Yegnanarayana, 2004). However, the accuracy of second stage depends on the accuracy of first stage. Therefore, in this study, we have explored in a different way to improve the prediction accuracy of durations by separating or grouping the durations of syllables based on production aspects of vowel generation and thereby developing single stage multi-model based duration models rather than multi-stage duration model. The implicit knowledge of duration is usually captured by using modeling techniques. In this work, supervised learning is carried out using neural networks to capture the underlying interactions that exist between input and output features (Haykin, 1999). The main contributions of this paper are as follows:

1. Analysing the syllable durations based on the vowel production characteristics.

2. Separating the syllables based on production aspects of vowels and developing the duration models for each class of vowel articulations.

The paper is organized as follows: Section 2 presents an overview of the existing research on acquisition of duration knowledge using different models. Performance of single neural network model for predicting the duration values of syllables along with the details of database and feature is given in Section 3. Analysis of durations of syllables for different vowel production characteristics is discussed in section 4. Section 5 compares the performance results of the proposed multi-models with single duration model. Summary and conclusions of this paper is presented in Section 6.

2 Previous efforts

Different approaches have been proposed by many researchers for modeling durations of sound units in the development of TTS systems. Duration models range from rule-based methods to data-based methods (Mixdorff, 2002). In the rule-based models, some set of rules will be derived with the help of linguistic experts and phoneticians using limited amount of data. However, the state-of-art is dominated by data-based models which gain knowledge directly from the data. The data-based methods are generally dependent on the quality and quantity of available training data.

Rule-based models like Klatt model, which applies rules to lengthen or shorten the duration of the segments (Klatt, 1979). Umeda developed rule-based duration model (Umeda, 1976) which is distinctly different from Klatt's model. The Chinese and Japanese TTS systems emphasize more on pitch rather than durations, due to tonal nature of the languages. Lee *et al* have developed Chinese syllable based TTS system with simple rule-based duration model (Lee et al., 1989). Linear statistical model like sum-of-products models, which combine multiple features into single expression. Jan van Santen proposed sums-of-products (SoP) model (Santen, 1994). The model uses set of linear equations based on the prior phonetic and phonological information as well as the information obtained by analyzing the data. Due to availability of large speech corpora, many researchers have proposed non-linear statistical approaches for analyzing large data. The two major approaches follow under this category are Classification and Regression Trees (CART) and Artificial Neural Networks (ANN). The CART based models are typical data-based duration models that can be constructed automatically. The self-configuration capability of CART makes them very popular (Black and Lenzo, Beijing China 2000); for instance the Festival TTS system uses *'wagon'* tool to construct such trees from the existing databases. Riley used the CART based model for predicting the segmental durations (Ri-

ley, 1992). The prediction of syllable durations using neural networks is proposed by Campbell (Campbell, 1990). Neural network models also used by Barbosa and Bailly to predict the duration of unit, known as Inter Perceptual Center Group (IPCG) (Barbosa and Bailly, 1994). Neural network based duration models also exist for languages like Arabic (Hifny and Rashwan, 2002), Spanish (Cordoba et al., 1999), Portuguese (Teixeira and Freitas, 2003) and German (Sonntag et al., 1997).

In view of Indian context, the rule based duration model is developed by Kumar and Yegnanarayana for Hindi TTS system (Kumar and Yegnanarayana, 1989) (Kumar, 1990). The rules were derived by analyzing 500 sentences, considering contextual and positional information. About 31 rules were derived to predict the durations. Later the rules were upgraded by analysing the large broadcast news data in Hindi (Kumar, 2002) (Kumar et al., 2002). CART based duration models are developed for languages like Hindi and Telugu (Krishna and Murthy, 2004). Rao and Yegnanarayana have used statistical models such as neural networks and support vector machines for modeling the durations of syllables for Hindi, Telugu and Tamil (Rao and Yegnanarayana, 2007). The duration models were developed by using broadcast news data from the three languages. Linguistic constraints represented in the form of positional, contextual and phonological features were used to capture the durational phenomena. To improve the accuracy of prediction further, a two-stage duration model was developed. By using two-stage model prediction of short and long durations of syllables is better compared to single stage model.

3 Single duration model using feed-forward neural network

The details of experimental database, features, neural network and the evaluation results for single duration model is presented in the following subsections.

3.1 Experimental database

The text utterances of speech database used for this study are collected mainly from Anandabazar Patrika - a Bengali news paper. It consists of news from several domains like sports, politics, entertainment, and stories. The other sources include story and text books in various fields such as history, geography, travelogue, drama and science. The text corpus covers 7762 declarative sentences derived from 50,000 sentences through optimal text selection method (Narendra et al., 2011). The corpus covers 4372 unique syllables and 22382 unique words. The optimal text is recorded with a professional female artist in a noiseless chamber. The duration of total recorded speech is around 10 hrs. The speech signal was sampled at 16 kHz and represented as 16 bit numbers. The speech utterances are segmented and labeled into syllable-like units using ergodic hidden Markov models (EHMM) (Rabiner and Juang, 1993). For every utterance a label file is maintained, which consists of syllables of the utterance and their timing information. The percentage of different syllable structures present in the database are V(8.20%), VC(3.50%), VCC (0.20%), CV(50.41%), CVC(32.26%), CVCC(1.05%), CCV(2.50%), CCVC(1.77%) and CCCV(0.11%), where C is a consonant and V is a vowel.

3.2 Features

As we have developed syllable based TTS system, therefore we have used syllable specific features represented by linguistic and production constraints are represented by positional, contextual, phonological and articulatory (PCPA) features (Reddy and Rao, 2012)(Reddy and Rao, 2013). The features representing linguistic constraints are syllable position in the sentence, syllable position in the word, word position in the sentence, syllable identity, contextual information, syllable nucleus, whereas production constraints features are vowel length, vowel height, vowel frontness, vowel roundness, consonant type, consonant place, consonant voice, aspiration, nukta, first phone, last phone.

3.3 Feed-forward Neural Network

Feed-forward neural networks (FFNN) are used in this work for modeling the durations of sequence of syllables using PCPA features, since Ramu *et al* in (Reddy and Rao, 2012) had confirmed neural network model is outperformed compared to other models like classification and regression trees and linear regression trees. Therefore, in this work also a four layer feedforward neural network (FFNN) with the structure represented in Fig. 2 is used for predicting the duration values.

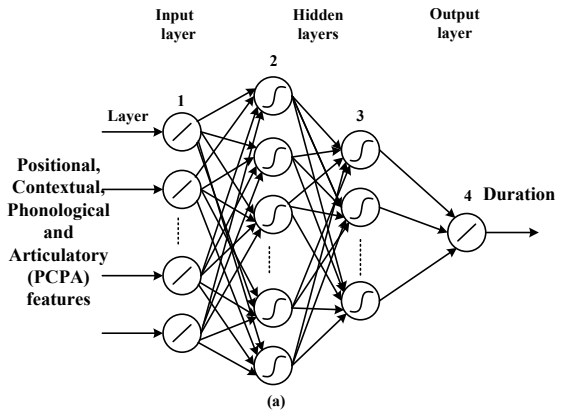

Figure 2: Architecture of four layer feedforward neural network for predicting the duration values of syllables.

In Fig. 2, the input layer which is the first layer consists of linear neuron units. The second and third layers are the hidden layers with non-linear neuron units. The last layer is the output layer with linear neuron units. The first hidden layer (second layer in Fig. 2) of the neural network consists of more units compared to the input layer (first layer in Fig. 2), so that network can capture local variations of features in the input space. The second hidden layer (third layer in Fig. 2) of the neural network has fewer units compared to the input layer, so that network can capture global variations of features in the input space (Haykin, 1999). The last layer (fourth layer in Fig. 2) is the output layer having one linear unit. The activation function for the units at the input and output layers is linear, whereas the activation function used at hidden layers is non-linear. The extracted PCPA feature vectors representing positional, contextual, phonological and articulatory features are presented as input, and the corresponding duration values are presented as desired outputs to the FFNN models.

The generalization by the network is basically influenced by three major factors : (1) the architecture of the network, (2) the amount of data used in the training phase of the network, and (3) the complexity of the problem. We have some control over the second factor but there is no control over the third factor. Different network structures were explored in this study to obtain the optimal performance, by incrementally varying the hidden layer neurons in between 5 and 100 as follows:

1. In the first iteration, the number of neurons in layer 2 (approximately 55) is considered to be greater than 1.5 times the number of neurons present in layer 1.

2. The number of neurons in layer 3 (approximately 25) considered in the first iteration is to be less than 0.75 times the number of neurons present in layer 1.

3. In this work, optimal structure is determined in 2 steps.the number of neurons in layer 2 are increased with an increment of 5 from 55 to 100, whereas in layer 3, the number of neurons are decreased with a decrement of 5 from 25 to 5. Based on the best performance (least training error) of all combinations (i.e., $10 \times 5 = 50$ combinations), layer 2 and layer 3 neurons are fixed.

4. In step 2, with the obtained neurons in step 1, fine tuning is carried out by incrementing the neurons of layer 2 and decrementing the neurons of layer 3 with the step count of 1.

5. For example in step 1, assume the best performance is obtained with h1 and h2 neurons in layer 2 and 3, respectively. In step 2, the neurons of layer 2 are varied from h1-4 to h1+4 excluding h1 (7 values) with step count of 1. Thus, in step 2, 49 possible combinations are tried out.

6. Overall, for finding the best optimal structure, we explored 99 possible combinations.

The structure of the network is represented by A*L* B*N* C*N* D*L*, where *L* denotes linear unit and *N* denotes non-linear unit. A, B, C and D are the integer values indicate the number of units used in different layers. The activation function used in the non-linear unit (*N*) is *tanh(s)* function, where '*s*' is activation value of that unit. The (*empirically arrived*) final optimal structures obtained with minimum generalization errors for predicting the durations is 35L 68N 17N 1L. The input and output features are normalized between [-1, 1], before giving to the neural network.

The training process of FFNN is carried out using Levenberg-Marquardt back-propagation algorithm to adjust the weights of the neural network, by back propagating the mean-squared error to the neural units and optimizes the free parameters (synaptic weights) to minimize the error (Yegnanarayana, 1999). The back-propagation

network learns by examples. So, we use input-output examples to show the network what type of behaviour is expected, and the back propagation algorithm allows the network to adapt. The back propagation learning process works in small iterative steps as follows:

- One of the example cases is applied to the network.
- The network produces some output based on the current state of it's synaptic weights (initially, the output will be random).
- The network output is then compared to the desired output and a mean-squared error signal is calculated.
- The error value is then propagated backwards through the network, and weights are updated to decrease the error in each layer.
- The whole process is repeated for each of the examples.

For each syllable a 35 dimensional feature vector is formed, representing the positional, contextual, phonological and articulatory information. In this work, the data consists of 177820 syllables are used for modeling the duration. The data is divided into two parts namely design data and test data. The design data is used to determine the network topology. The design data in turn is divided into two parts namely training data and validation data. Training data is used to estimate the weights (includes biases) of the neural network and validation data is used to minimize the over-fitting of network, to verify the performance error and to stop training once the non-training validation error estimate stops decreasing. The test data is used once and only once on the best design, to obtain an unbiased estimate for the predicted error of unseen non-training data. The amount of data used for training, validation and testing the network are 70%, 15% and 15%, respectively. The motivation here is to validate the model on a data set from the one used for parameter estimation. As generalization is the goal of the neural network, hence we used cross validation.The early stopping method is used to avoid over-fitting of the neural network.

3.4 Objective and subjective evaluation

The performance of duration model is evaluated by using objective measures such as percentage of syllables predicted within different deviations from their actual duration values, average prediction error (μ), standard deviation (σ) and linear correlation coefficient ($\gamma_{X,Y}$) between actual and predicted duration values. The computation of objective measures are as follows:

$$D_i = \frac{|x_i - y_i|}{x_i} \times 100, \mu = \frac{\sum_i |x_i - y_i|}{N}, \quad (1)$$

$$\sigma = \sqrt{\frac{\sum_i d_i^2}{N}}, \text{ and } \gamma_{X,Y} = \frac{V_{X,Y}}{\sigma_X . \sigma_Y} \quad (2)$$

$$\text{where } d_i = e_i - \mu, e_i = x_i - y_i \quad (3)$$

$$\text{and } V_{X,Y} = \frac{\sum_i |x_i - \bar{x}| . |y_i - \bar{y}|}{N} \quad (4)$$

where x_i, y_i are the actual and predicted duration values respectively, and e_i is the error between the actual and predicted duration values. The deviation in error is d_i, and N is the number of observed duration values of the syllables. σ_X, σ_Y are the standard deviations for the actual and predicted duration values respectively, and $V_{X,Y}$ is the correlation between the actual and predicted duration values.

The performance of method is also evaluated by means of subjective analysis. Naturalness and intelligibility are two important key features to measure the quality of the synthesized speech. Naturalness can be defined as, how close the synthesized speech to human speech, whereas intelligibility is defined as how well the message is understood from the speech. The perceptual evaluation is conducted by incorporating FFNN based duration model developed into the TTS system. In this work, 20 subjects within the age group of 23-35 were considered for perceptual evaluation of synthesized speech. After giving appropriate training to the subjects, evaluation of TTS system is carried out in a laboratory environment. Randomly 10 sentences were selected, and played the synthesized speech signals through headphones to evaluate the quality. Subjects have to assess the quality on a 5-point scale (Reddy and Rao, 2012) for each of the synthesized sentences. The subjective listening tests are carried out for the synthesized sentences generated by FFNN duration model developed using the PCPA features. The mean opinion scores (MOS) are calculated for both naturalness and intelligibility of the synthesized speech.

The objective and subjective evaluation results of single duration model is given in Table 1.

Table 1: Performance of single duration model

% Predicted syllables within deviation			**Objective measures**			**Subjective measures**	
10%	15%	25%	μ (ms)	σ (ms)	γ	*Naturalness*	*Intelligibility*
35.14	50.63	72.56	39.04	35.09	0.83	3.53	2.86

4 Analysis of duration of syllables based on vowel articulation factors

Production aspect of speech segments (vowels and consonants) is one of the major factor influencing the variation in the duration of syllables. In syllables, major contribution of the duration values is mainly from the vowels compared to consonants. This can be verified from the example shown in Fig. 3. The phrase "sArodAdebI" contains two words("sArodA","debI"), 5 syllables("sA","ro","dA","de","bI") and 10 phones(s,A,r,o,d,A,d,e,b,I). The speech signal of the phrase "sArodAdebI" with its consonant and vowel portions (duration) shown in Fig. 3, indicates that major portion of duration of syllables is mainly from the vowel region.

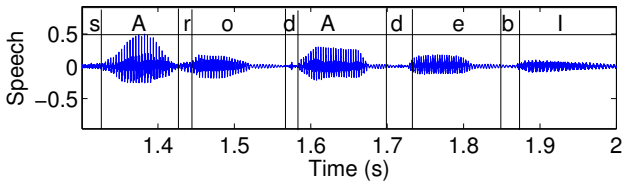

Figure 3: Durations of consonants and vowels of the syllables in the phrase "sArodAdebI"

Moreover, the duration values of vowels vary based on their place and manner of articulation while uttering. Therefore, in this study, analysis of duration is carried out based on production aspects of vowels. The features related to vowel are vowel length, vowel height, vowel frontness and vowel roundness. The distribution plot of syllables based on production characteristics of vowel is given in Fig. 4.

From Fig. 4, it is quite clear that there is a variations in the distributions of syllables of different vowel production characteristics. Therefore, if we separate the syllables based on these variations model and modeled each group separately, the prediction performance can be improved.

5 Proposed multi-model based approach

It was concluded that the durations of syllables depends on articulations of vowels from the analysis presented in Section 4. Therefore, in this study, multi-models are developed separately for each case to improve the prediction accuracy. The prediction performance of proposed FFNN multi-models is compared with the FFNN single duration model developed using PCPA features. The details of performance of multi-models are discussed in the following subsections.

In this study, multi-models are developed based on different vowel articulations described in the section 4. For the vowel length, durations of syllables are divided into 3 parts and hence 3 models are developed representing short, diphthongs and long vowels. For vowel height, syllables are divided into 3 parts such as high, mid and low vowels based on tongue height while articulating vowels, and hence 3 models are developed for vowel height. Similarly, based on tongue frontness, 3 models are developed representing front, mid and backness of tongue while articulating vowels. Lastly, 2 models based on vowel roundness is developed by categorizing syllables as lip roundness and no lip roundness while articulating vowels.

The input features used for developing multi-models is same as that of single model except the articulatory feature of the vowel type. For vowel length models, the articulatory feature vowel length is not used as it is redundant. This is applicable even for vowel height, vowel frontness and vowel roundness models. The performance of multi-models for each category of vowel articulations is given in Table 2. Column 1 of Table 2 indicate the vowel feature , column 2 indicate the models developed for each vowel feature, columns 3-5 indicates the percentage of syllables predicted within different deviations from their actual duration values and columns 6-8 indicates objective

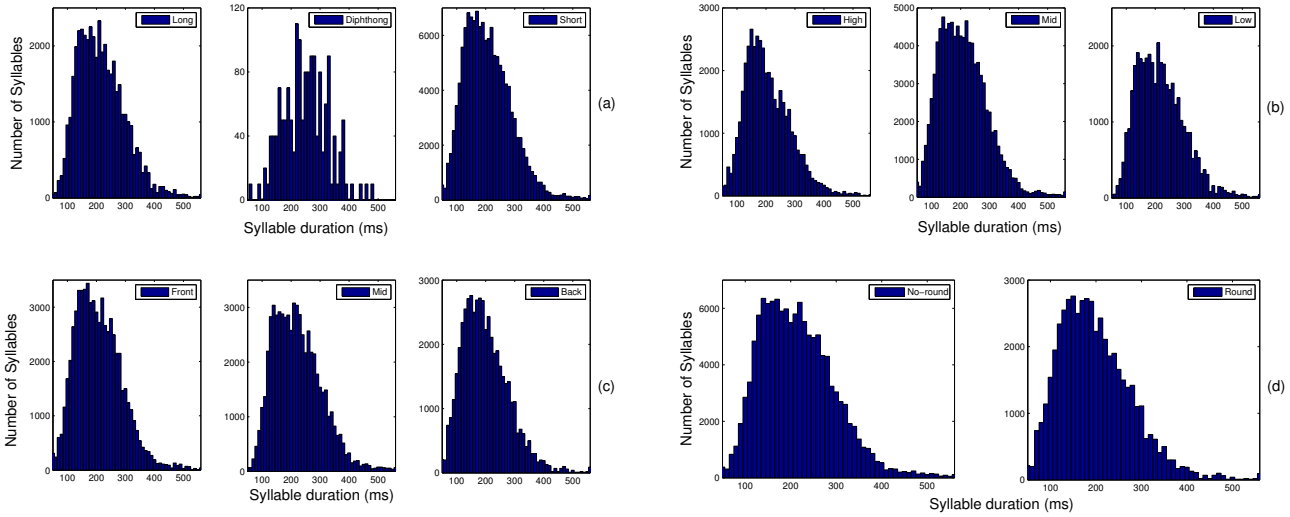

Figure 4: Distribution plots of the syllable durations based on different articulations of vowels related to (a) vowel length (b) vowel height, (c) vowel frontness, and (d) vowel roundness

Table 2: Performance of FFNN based multi-models based on vowel articulations for predicting the duration values of the syllables.

Vowel features	**Models**	**% Predicted syllables within deviation**			**Objective measures**		
		10%	15%	25%	μ (ms)	σ (ms)	γ
Vowel Length	Long	36.82 (35.19)	54.87 (51.25)	73.08 (72.31)	36.50 (39.87)	29.09 (35.50)	0.86 (0.80)
	Dipthong	32.16 (48.39)	45.59 (63.40)	66.83 (88.56)	40.05 (33.63)	23.91 (30.66)	0.85 (0.78)
	Short	43.03 (34.99)	60.00 (50.30)	81.46 (72.47)	27.76 (38.83)	25.28 (35.00)	0.93 (0.84)
	AVERAGE	**41.44 (35.14)**	**58.64 (50.63)**	**79.31 (72.56)**	**29.98 (39.04)**	**26.20 (35.09)**	**0.91 (0.83)**
Vowel Height	High	40.63(35.11)	55.25(51.99)	77.30 (72.68)	31.33 (37.13)	28.72 (33.85)	0.84 (0.82)
	Mid	39.15 (34.77)	56.48 (49.61)	79.79 (72.61)	31.19 (39.66)	27.49 (35.34)	0.91 (0.85)
	Low	35.28 (36.11)	52.14 (51.75)	75.54 (72.29)	35.37 (39.54)	28.18 (35.78)	0.87 (0.80)
	AVERAGE	**38.66 (35.14)**	**55.25 (50.63)**	**78.29 (72.56)**	**32.13 (39.04)**	**27.93 (35.09)**	**0.89 (0.83)**
Vowel Frontness	Front	39.21 (34.61)	55.22 (51.01)	76.89 (73.64)	30.82 (38.36)	25.34 (32.89)	0.90 (0.83)
	Mid	41.13(37.01)	58.92 (53.23)	80.57 (74.70)	31.53 (39.12)	27.25 (35.57)	0.91 (0.83)
	Back	33.10 (33.52)	47.38 (46.89)	70.60 (68.47)	36.67 (39.84)	30.72 (37.41)	0.87 (0.83)
	AVERAGE	**38.17 (35.14)**	**54.31 (50.63)**	**76.41 (72.56)**	**32.71 (39.04)**	**27.51 (35.09)**	**0.90 (0.83)**
Vowel Roundness	No-round	33.50 (35.77)	50.14 (52.09)	72.43 (74.15)	37.88 (38.73)	30.53 (34.19)	0.85 (0.83)
	Round	34.71 (33.52)	51.02 (46.89)	74.68 (68.47)	33.24 (39.84)	28.61 (37.40)	0.89 (0.83)
	AVERAGE	**33.84 (35.14)**	**50.39 (50.63)**	**73.06 (72.56)**	**36.58 (39.04)**	**29.99 (35.09)**	**0.87 (0.83)**

measures.

Table 1 represents the average performance of all the syllables in the test set of the single duration model. The prediction performance of the syllables corresponding to different vowel production characteristics is also computed which is shown in brackets in Table 2 and the average performance of multi-models is shown in bold. From Table 2, it is quite clear that the performance of multi-models developed based on articulation of vowels is better compared to the performance of single duration model (the values in brackets). From this hypothesis, we concluded that the prediction accuracy of durations depends on the articulations of vowels,

and it can be captured by separating the syllables and developing different models based on the articulations. Among all the multi-models based on different vowel articulations. The multi-model developed based on vowel length production characteristics is outperformed compared to other multi-models. It is found that the average prediction error is reduced by 23.21% and correlation coefficient is improved by 9.64% using multi-model developed by separating syllables based on vowel length production characteristic compared to single duration model. However, we can notice that the average error for the model developed using duration of syllables having vowel diphthongs is more compared to single duration model. This is expected because the amount of syllables having vowel diphthongs are quite less for training the neural network (0.89% in training set and 0.77% in test set). This error can be minimized by taking average duration value of diphthongs.

The subjective listening tests are also carried out for the synthesized sentences generated by FFNN multi-models as shown in Table 3. The mean opinion scores (MOS) are calculated for both naturalness and intelligibility of the synthesized speech. For comparing the quality of synthesized speech based on incorporation of specific duration models, we have also derived the mean opinion scores for the synthesized speech generated in the absence of duration model. From Table 3, it is observed that the MOS values for naturalness and intelligibility of FFNN multi-model developed based on vowel length production characteristics is better compared to other models. The scores indicate that the intelligibility of the synthesized speech is fairly acceptable, whereas the naturalness seems to be poor. Naturalness is mainly attributed to individual perception.

6 Summary and conclusions

In this work, novel multi-models are developed based on the articulation characteristics of vowels. Among all multi-models, the multi-model developed based on vowel length category is performed better compared to other multi-models developed based on vowel height, vowel frontness and vowel roundness. The prediction accuracy of multi-model is outperformed compared with single duration model. The prediction performance of diphthongs model in vowel length multi-model is dropped compared to single duration model. The error can be minimized or prediction accuracy can be further improved by taking average duration of diphthongs rather than modeling using networks due to less frequency of diphthongs or it can be included in the model of long vowels as long vowel durations are quite close to duration values of diphthongs. The prediction accuracy can be further improved by analogizing and constructing the multi-models based on position aspect of syllables in sentence and in words, as well as multi-models based on consonant production characteristics.

Table 3: Mean opinion scores for the quality of synthesized speech of TTS after incorporating the multi-model based duration models

Models	**Mean Opinion Scores**	
	Intelligibility	***Naturalness***
Without Duration model	3.10	2.62
Single model	3.53	2.86
Vowel Length	3.70	3.01
Vowel Height	3.65	2.89
Vowel Frontness	3.62	2.88
Vowel Roundness	3.56	2.87

References

P. A. Barbosa and G. Bailly. 1994. Characterization of rhythmic patterns for text-to-speech synthesis. *Speech Communication*, 15:127–137.

Alan W Black and Kevin A. Lenzo. Beijing, China, 2000. Limited domain synthesis. In *ICSLP*.

W. N. Campbell. 1990. Analog I/O nets for syllable timing. *Speech Communication*, 9(1):57–61, Feb.

R. Cordoba, J. A. Vallejo, J. M. Montero, J. Gutierrezarriola, M. A. Lopez, and J. M. Pardo. 1999. Automatic modeling of duration in a Spanish text-to-speech system using neural networks. In *Proc. European Conf. Speech Communication and Technology*, pages 1619–1622, Budapest, Hungary, Sept.

Simon Haykin, 1999. *Neural Networks: A Comprehensive Foundation*. Pearson Education Aisa, Inc., New Delhi, India.

Y. Hifny and M. Rashwan. 2002. Duration modeling of Arabic text-to-speech synthesis. In *Proc. Int. Conf. Spoken Language Processing*, pages 1773–1776, Denver, Colorado, USA, Sept.

D. H. Klatt. 1979. Synthesis by rule of segmental durations in English sentences. In B. Lindblom and S. Ohman, editors, *Frontiers of Speech Communication Research*, pages 287–300. Academic Press, New York.

N. Sridhar Krishna and Hema A. Murthy. 2004. Duration modeling of Indian languages Hindi and Telugu. In *5th ISCA Speech Synthesis Workshop*, pages 197–202, Pittsburgh, USA, May.

S. R. Rajesh Kumar and B. Yegnanarayana. 1989. Significance of durational knowledge for speech synthesis in Indian languages. In *Proc. IEEE Region 10 Conf. Convergent Technologies for the Asia-Pacific*, pages 486–489, Bombay, India, Nov.

K. Kiran Kumar, K. Sreenivasa Rao, and B. Yegnanarayana. 2002. Duration knowledge for text-to-speech system for Telugu. In *Proc. Int. Conf. Knowledge Based Computer Systems*, pages 563–571, Mumbai, India, Dec.

S. R. Rajesh Kumar. 1990. Significance of durational knowledge for a text-to-speech system in an Indian language. Master's thesis, Dept. of Computer science and Engineering, Indian Institute of Technology Madras, Mar.

K. Kiran Kumar. 2002. Duration and intonation knowledge for text-to-speech conversion system for Telugu and Hindi. Master's thesis, Dept. of Computer science and Engineering, Indian Institute of Technology Madras, May.

L. Lee, C. Tseng, and M. Ouh-Young. 1989. The synthesis rules in a Chinese text-to-speech system. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 9(4):1309–1320.

Hansjorg Mixdorff. 2002. *An integrated approach to modeling German prosody*. Ph.D. thesis, Technical University, Dresden, Germany, July.

N. P. Narendra, K. Sreenivasa Rao, Krishnendu Ghosh, Vempada Ramu Reddy, and Sudhamay Maity. 2011. Development of syllable-based text to speech synthesis system in Bengali. *Int. J. of Speech Technology, Springer*, 14(3):167–181.

L. Rabiner and B. H. Juang. 1993. *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ.

K. Sreenivasa Rao and B.Yegnanarayana. 2004. Two-stage duration model for Indian languages using neural networks. *Lecture Notes in Computer Science : Neural Information Processing, Springer*, 3316:1179–1185.

K. Sreenivasa Rao and B. Yegnanarayana. 2007. Modeling durations of syllables using neural networks. *Computer Speech and Language*, 21:282–295, Apr.

K. Sreenivasa Rao. 2005. *Acquisition and incorporation prosody knowledge for speech systems in Indian languages*. Ph.D. thesis, Dept. of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai, India, May.

V. Ramu Reddy and K. Sreenivasa Rao. 2012. Better human computer interaction by enhancing the quality of text-to-speech synthesis. In *Proc. Int. Conf. Intelligent Human Computer Interaction (IHCI)*, pages 1–6, IIT Kharagpur, India, Dec.

V. Ramu Reddy and K. Sreenivasa Rao. 2013. Two-Stage Intonation Modeling Using Feedforward Neural Networks for Syllable based Text-to-Speech Synthesis. *Computer Speech and Language*, 27(5):1105–1126, Aug.

M. Riley. 1992. Tree-based modeling for speech synthesis. *in G. Bailly, C. Benoit, and T. Sawallis (Eds.), Talking machines: Theories, models and designs*, pages 265–273.

J. P. H. Van Santen. 1994. Assignment of segment duration in text-to-speech synthesis. *Computer Speech and Language*, 8:95–128, Apr.

G. P. Sonntag, Thomas Portele, and Barbara Heuft. 1997. Prosody generation with a neural network: Weighing the importance of input parameters. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 931–934, Munich, Germany, Apr.

Joao Paulo Teixeira and Diamantino Freitas. 2003. Segmental durations predicted with a neural network. In *Proc. European Conf. Speech Communication and Technology*, pages 169–172, Geneva, Switzerland, Sept.

Noriko Umeda. 1976. Linguistic rules for text-to-speech synthesis. *Proc. IEEE*, 64:443–451, April.

B. Yegnanarayana, 1999. *Artificial Neural Networks*. Prentice-Hall, New Delhi, India.