Neurocomputing 171 (2016) 1323–1334

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Prosody modeling for syllable based text-to-speech synthesis using feedforward neural networks

V. Ramu Reddy, K. Sreenivasa Rao

School of Information Technology, Indian Institute of Technology Kharagpur, Kharagpur 721302, West Bengal, India

ARTICLE INFO

Article history:
Received 28 April 2014
Received in revised form
31 May 2015
Accepted 24 July 2015
Communicated by R. Capobianco Guido
Available online 4 August 2015

Keywords:
Prosody
Text-to-speech synthesis
Feed-forward neural networks
Phonological features
Positional and contextual features
Articulatory features

ABSTRACT

Prosody plays an important role in improving the quality of text-to-speech synthesis (TTS) system. In this paper, features related to the linguistic and the production constraints are proposed for modeling the prosodic parameters such as duration, intonation and intensities of the syllables. The linguistic constraints are represented by positional, contextual and phonological features, and the production constraints are represented by articulatory features. Neural network models are explored to capture the implicit duration, F_0 and intensity knowledge using above mentioned features. The prediction performance of the proposed neural network models is evaluated using objective measures such as average prediction error (μ), standard deviation (σ) and linear correlation coefficient ($\gamma_{X,Y}$). The prediction accuracy of the proposed neural network models is compared with other state-of-the-art prosody models used in TTS systems. The prediction accuracy of the proposed prosody models is also verified by conducting listening tests, after integrating the proposed prosody models to the baseline TTS system.

1. Introduction

Prosody modeling plays a vital role in developing a high quality text-to-speech synthesis (TTS) system. Prosody refers to duration, intonation and intensity patterns of speech associated to the sequence of syllables, words and phrases. These features are usually observed over longer segments of speech. The sequence of syllable durations is defined as duration pattern. Variation in duration patterns provides naturalness to speech. Intonation can be defined as the dynamics of fundamental frequency (F_0) contour over time, caused by vocal fold vibration. The intensity is considered to be closely related with the perceived loudness. The dynamic behavior of intensity pattern is known as intensity or energy contour.

A good prosody model should capture the duration, intonation and intensity patterns of natural speech. The objective of the present study is to determine whether non-linear models can capture the implicit knowledge of the prosodic patterns of syllables. In this work, neural network models were proposed for modeling the prosody. Neural networks are known for their ability to capture the complex non-linear relations present in data [1,2]. Neural networks have generalization ability to predict the values reasonably well for the patterns which are not present in the learning phase. Recently, an incremental self-organizing map integrated with hierarchical neural network (ISOM-HNN) has gained lot of attention among researchers working in the area of neural networks and pattern recognition [3]. The potential reasons for its popularity are due to (i) effective detection of unknown radio signals in highly ambiguous environments, (ii) suitability to real-time applications and (iii) improvement in prediction accuracy. The objective of incremental self-organizing map (ISOM) is to embed incremental capturing ability into SOM [4]. This will ensure the adaptability of ISOM, and provide the robust detection accuracy even in severely degraded environments. Hierarchy of neural networks (HNN) always provide superior performance compared to single large neural network [5]. Therefore, the combination of ISOM and HNN will certainly enhance the accuracy of prediction as well as ensure robust performance against degraded environments.

In speech signal, the duration, intonation and intensity of each unit is dictated by the linguistic and production constraints of the unit [6–8]. In this paper, linguistic and production constraints are proposed for predicting the prosodic patterns of the sequence of syllables. Linguistic constraints are represented by positional, contextual and phonological (PCP) features, and production constraints are represented by articulatory (A) features. From here onwards features representing linguistic and production constraints together are referred as PCPA features.

In this work, prosody models are developed using PCPA features extracted from speech corpus used for developing baseline Bengali TTS system [9]. Most of the existing works have explored only PCP features for modeling prosody. In this work, we want to explore

E-mail addresses: ramu.csc@gmail.com (V. Ramu Reddy), ksrao@iitkgp.ac.in (K. Sreenivasa Rao).

http://dx.doi.org/10.1016/j.neucom.2015.07.053
0925-2312/

combination of PCP and articulatory (A) features for improving the accuracy of prediction. Further, we also want to examine the contribution of individual features on prediction accuracy. The quality of the speech synthesized by the proposed prosody model based TTS system is compared with the baseline TTS system [9].

Rest of the paper is organized as follows: Following section presents an overview of the existing works on acquisition of prosodic knowledge using different models. Section 3 discusses the baseline TTS system and speech database used for building prosody models as well as evaluating the performance of the proposed prosody models. Section 4 describes the proposed features used for predicting the prosodic values. The details of proposed neural network model is given in Section 5. The evaluation of the proposed FFNN models along with the other state-of-the-art prosody models used in TTS systems is carried out in Section 6, using objective and subjective measures. Section 7 presents the effect of individual features in predicting the prosodic parameters of the sequence of syllables. Summary and conclusions of the work are laid in the final section.

2. Literature review

2.1. Related work on duration modeling

Different approaches have been proposed for modeling durations of sound units in the development of TTS systems [10,11]. Duration models range from rule-based methods to data-based methods. Rule-based models like Klatt model applies rules to lengthen or shorten the duration of the segments [12]. Umeda developed rule-based duration model [13] which is distinctly different from Klatt's model. Lee et al. have developed Chinese syllable based TTS system with simple rule-based duration model [14]. Jan van Santen proposed sums-of-products (SoP) model [15], which 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 that follow under this category are Classification and Regression Trees (CART) and Artificial Neural Networks (ANN). The self-configuration capability of CART has gained lot of popularity [16]; for instance the Festival TTS system uses 'wagon' tool to construct CART from the existing database. Riley used the CART based model for predicting the segmental durations [17]. The prediction of syllable durations using neural networks is first proposed by Campbell [18].

In Indian context, the rule based duration model is developed by Kumar and Yegnanarayana for Hindi TTS system [19–21]. CART based duration models are developed for languages like Hindi and Telugu [22]. Rao and Yegnanarayana have used statistical models such as neural networks and support vector machines for modeling the durations of syllables in Hindi, Telugu and Tamil [5,8,23,24].

2.2. Related work on intonation modeling

The phonological (tone sequence) models interpret the F_0 contour as a linear sequence of phonologically distinctive units (tones or pitch accents), which are local in nature. Popular phonological models include Pierreihumbert's model [25] and TOBI (Tone and Break Indices) [26]. Intonation model for Danish language was developed by Gronnum which is conceptually quite different from the tone sequence model [27]. Phonetic (superposition or overlay) models interpret F_0 contour as the result of superposition of several components of different temporal scopes. A classical *superpositional* intonational model for Japanese has been presented by Fujisaki and his colleagues [28]. Apart from these two prevalent intonation models, there are several important models that defy categorization as being either tone sequence or superpositional type. For example, *perception-based* IPO models explored some measurable acoustic prosodic properties of speech but not perceived by the listener all the time [29]. Finally, *acoustic stylization* models aim at efficient analysis and synthesis using *rise/fall/connection* (RFC) models [30]. Paul Taylor designed the tilt intonation model to provide a robust computational analysis and synthesis of intonation contours. This model is based on *rise/fall/connection* (RFC) model, and is a bi-directional model that gives an abstraction directly from the data [30].

In the literature, several models using neural networks are described for predicting the intonation patterns of syllables in continuous speech [31–34]. In [31], Scordilis and Gowdy have used neural networks in parallel and distributed manner to predict the average F_0 value for each segment, and also the temporal variations of F_0 within a segment. In [35], Traber used a neural network with two hidden layers to model eight target F_0 points for each syllable in German. Buhmann et al. have used a recurrent neural network for developing multi-lingual (six languages) intonation models [34].

In Indian context, a rule-based intonation model was proposed by Kumar et al. for Hindi TTS system [36,6]. Rao and Yegnanarayana have developed feedforward neural network and support vector machine models for capturing the intonation patterns of speech in Indian languages [7].

2.3. Related work on intensity modeling

Over the years, most of the researchers have explored only intonation and duration patterns of speech segments. In most of the synthesis applications, intensity is often either completely neglected or is modeled concurrently with fundamental frequency because it has been felt that intensity features are embedded in intonation [37]. On the other hand, loudness has a relation with duration in the perception of prominence [38], which inevitably increases its significance. In [39], artificial neural networks were used for prediction of duration, average loudness-level and average pitch for a known phoneme context and the word context of the Finnish language. Lee et al have used artificial neural networks to predict syllable energy [40]. In [41], prosody models were developed for TTS systems in European languages. It was observed that there was a decreasing trend of intensity variations in the tonic syllables as its position changes from the beginning to the end of the word. Some relevant variations of F_0, durations and intensity patterns across tonic syllable as a function of its position in the word was analyzed. Variations of prosody patterns were also analyzed for words in initial, middle and final position in the phrase and isolated words. Mannel [42] modeled intensities for a sequence of diphones in diphone based formant speech synthesis system. Tesser developed intensity models in addition to intonation and duration models for diphone based emotional TTS system [43].

3. Syllable based TTS system

In [9], unit selection based baseline TTS system was developed with the speech corpus recorded by a professional female artist. The choice of basic unit for synthesis depends on the language. Indian languages have a well-defined syllable structure. Syllable is a speech unit having a vowel at the nucleus covered by optional consonants on both sides of the vowel. A character in Indian language is close to syllable. Moreover, syllables can preserve coarticulation effect better, compared to phones and diphones. So, for building TTS in Bengali, syllables are used as basic units of

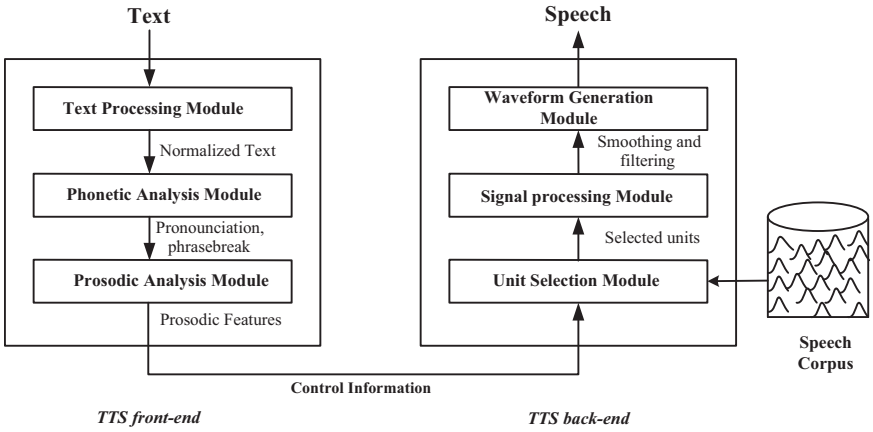

Fig. 1. Flow diagram indicating the sequence of steps for synthesizing the speech.

concatenation for synthesis. Festival framework was used in developing TTS system [16]. The sequence of steps followed for text-to-speech conversion is shown in Fig. 1.

The text processing module in TTS front-end will derive normalized text from the input text. In general, the text consists of special symbols, abbreviations and punctuation marks. The text processing module will replace them with appropriate spoken words. Phonetic analysis module derives the sequence of sounds for the sequence of words using language specific pronunciation rules or dictionary [44]. Prosodic analysis module consists of prosody models that are used to derive the prosodic information such as duration, intonation and intensity patterns associated to the sequence of syllables. The unit selection module present in TTS back-end, select the sequence of syllables from the speech corpus based on prosodic and phonetic features given by the text analysis module. Signal processing and waveform generation modules concatenate the sequence of syllables provided by the unit selection module and generate smooth speech waveform.

3.1. Speech database

The optimal text is recorded with a professional female artist in a noiseless chamber. The duration of total recorded speech is around 10 h. The speech signal was sampled at 16 kHz and represented as 16 bit numbers. For developing duration, intonation and intensity models, the durations, fundamental frequencies (F_0) and intensities of the syllables should be available in the database. The speech utterances are segmented and labeled into syllable-like units using ergodic hidden Markov models (EHMM). For every utterance a label file is maintained, which consists of syllables of the utterance and their timing (duration) information. Zero frequency filter method [45] is used to calculate the frame level fundamental frequencies of the utterance. Syllable boundaries are then marked based on the timing information provided by label files. Now, the frame level F_0 values which falls within the syllable boundaries are used to calculate syllable F_0 values. Syllable F_0 values are represented by three values namely, start F_0, middle F_0 and end F_0. Mean of 25% of beginning frame F_0 values of syllable are considered as start F_0, mean of 25% of end frame F_0 values of syllable are considered as end F_0, and mean of remaining frame F_0 values are considered as middle F_0. The average intensities of the syllables are computed as follows:

$$I = 10 \log_{10}\left(\frac{\sum_{i=1}^{N} x_i^2}{NP_0^2}\right) \quad (1)$$

where I is the average intensity of syllable expressed in decibel, N is number of speech samples present in a syllable, x_i is the amplitude of ith speech sample and P_0 is auditory threshold pressure expressed in Pascals ($P_0 = 2 \times 10^{-5}$ Pa). The syllable structures considered here are V, CV, CCV, CVCC and CCVC, where C is a consonant and V is a vowel. The minimum, maximum, mean and standard deviations of durations, intonation and intensities of syllables observed in the database are given in Table 1. The distribution plot of syllable durations, F_0 values located at the start, middle and end positions of syllables, and syllable intensities are shown in Fig. 2. From Fig. 2, it is observed that most of the durations, F_0 values and intensities of syllables are concentrated in between 110–350 ms, 130–270 Hz and 62–80 dB, respectively. It is observed, for some utterances in the database, the F_0 values of the syllables of end words are found to be around 150 Hz. This results in small F_0 peaks around 150 Hz (Fig. 2(b)). This is mainly attributed to speaking style of the speaker.

Table 1
Statistics of prosodic features of syllables present in the speech database.

Prosodic features	Minimum	Maximum	Mean	Standard deviation
Duration (ms)	50	560	212.9	80.6
F_0 (Hz)	80	290	223	47
Intensity (dB)	55	83	70.92	4.78

4. Features for developing Prosody models

In a speech signal, the prosodic pattern corresponding to a sequence of sound units is constrained by the linguistic and production constraints of the units [6,46]. Hence for modeling the prosodic parameters, we use features representing the linguistic and production constraints of the sound units. The linguistic and production constraints of syllables can be expressed using positional, contextual, phonological and articulatory (PCPA) features. In this study we use 35 dimensional feature vector representing the linguistic and production constraints of each syllable. Out of 35 features, 24 features represent the linguistic constraints in the form of positional, contextual and phonological information and the remaining 11 features represent articulatory information of production constraints of each syllable. The positional features are further classified based on syllable position in a word and sentence, and word position in a sentence.

4.1. Linguistic constraints

The detailed list of features representing linguistic constraints and the number of input nodes needed for the neural network to

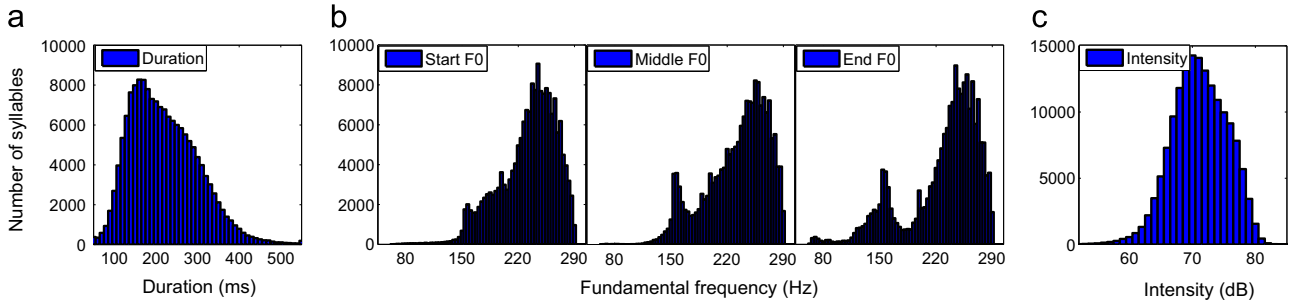

Fig. 2. Distribution plot of (a) durations of syllables, (b) F_0 values at start, middle and end positions of syllables and (c) intensities of syllables.

Table 2
List of linguistic constraints.

Factors	Features	#Nodes
Syllable position in the sentence	Position of syllable from beginning of the sentence Position of syllable from end of the sentence Number of syllables in the sentence	3
Syllable position in the word	Position of syllable from beginning of the word Position of syllable from end of the word Number of syllables in the word	3
Word position in the sentence	Position of word from beginning of the sentence Position of word from end of the sentence Number of words in the sentence	3
Syllable identity	Segments of the syllable (consonants and vowels)	4
Context of the syllable	Previous syllable Following syllable	4 4
Syllable nucleus	Number of segments before the nucleus Number of segments after the nucleus Number of segments in a syllable	3

represent these features are given in Table 2. These features are coded and normalized before giving to neural network.

4.2. Production constraints

Each sound unit has specific articulatory movements and positions during its production. Therefore, in this study the features related to different position and manner of articulation of speech segments (consonants and vowels) are considered as production constraints. These production constraints are represented as articulatory features to predict the prosodic parameters of the syllables.

In this study, 11 dimensional feature vector representing the articulatory features is used. The features used to represent the articulatory information are vowel length, vowel height, vowel frontness, vowel roundness (lip rounding), consonant type, consonant place, consonant voicing, aspiration, nukta, type of first phone and type of last phone in a syllable. The quality of the vowel depends on the articulatory features that distinguish different vowel sounds [47]. Daniel Jones developed the cardinal vowel system to describe vowels in terms of the common articulatory features *height* (vertical dimension), *backness* (horizontal dimension) and *roundedness* (lip position) [47]. Height relates to position of tongue and to degree of lowering of jaw. Backness relates to the position of the body of the tongue in oral cavity. Rounding refers to the roundness of the lips.

The consonant related features like consonant type, consonant place, consonant voicing, aspiration for syllables are extracted as follows:

(i) If there is any consonant or consonants present after the vowel then consonant which follows immediately after vowel is considered. For example, if the syllable structure is CVCC then the consonant related features of syllable are extracted for the second consonant.
(ii) If consonants are absent after the vowel, then the consonant previous to vowel is considered. For example, if syllable structure is CCV, as there is no consonant after the vowel, hence the consonant prior to vowel (i.e., second consonant) is considered in this case.
(iii) In case of syllables with only vowel, dummy values are used for the consonant related features.

In Bengali language, there are some consonants which are formed by combining the existing consonants and nukta to generate different pronunciation [48]. The detailed list of production constraints represented in the form of articulatory features is given in Table 3. Each articulatory feature is uniquely coded and it is shown within brackets in Table 3.

5. Modeling the prosody using feedforward neural networks

In this work, four layer feedforward neural networks (FFNN) [49] with input layer, two hidden layers and output layer are used for modeling the prosodic parameters of the syllables. The structures of the FFNN for predicting the durations, intonation and intensities of the syllables using linguistic and production constraints as features are shown in Fig. 3. In Fig. 3, 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. 3) of the neural network consists of more units compared to the input layer (first layer in

Table 3
List of Production Constraints.

Features	Description
vlen	Length of the vowel in a syllable [short(1), long(2) and dipthong(3)].
vheight	Height of the vowel in a syllable [high(1), mid(2)and low(3)].
vfront	Frontness of the vowel in syllable [front(1), mid(2) and back(3)].
vrnd	Lip roundness [no rounding(1) and rounding(2)].
ctype	Consonant type [stop(1), fricative(2), affricative(3), nasal(4), and liquid(5)].
cplace	Place or position of the production of the consonant [labial(1), alveolar(2), palatal(3), labio-dental(4), dental(5) and velar(6)].
cvox	Consonant is voiced or unvoiced [voiced(1) and unvoiced(2)].
asp	Consonant is aspirated or unaspirated [aspirated(1) and unaspirated(2)].
nuk	Consonant with nukta or not nukta [withnukta(1) and withoutnukta(2)].
fph	Type of first phone in a syllable [vowel(1), voiced consonant(2), unvoiced consonant(3), nasal(4), semivowel(5), nukta(6) and fricative(7)].
lph	Type of last phone in a syllable [vowel(1), voiced consonant(2), unvoiced consonant(3), nasal(4), semivowel(5), nukta(6) and fricative(7)].

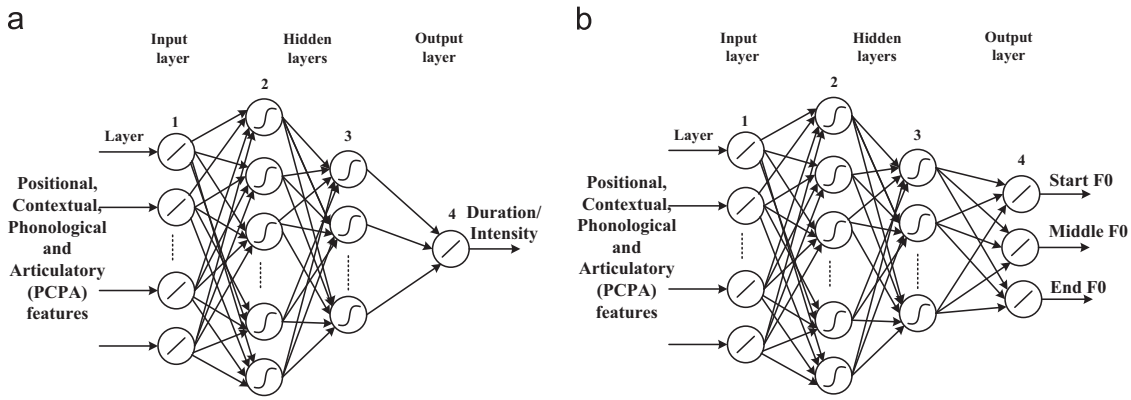

Fig. 3. Architectures of four layer feedforward neural network for predicting the (a) durations or intensities and (b) F_0 values of syllables.

Fig. 3), so that network can capture local variations of features in the input space. The second hidden layer (third layer in Fig. 3) 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 [1,2]. The last layer (fourth layer in Fig. 3) is the output layer having one or more linear units. The number of units in the output layer represent the dimension of the output feature vector. Fig. 3 (a) has one linear unit at output layer representing the duration or average intensity of syllable, whereas Fig. 3(b) has three linear units at output layer representing the three average F_0 values located at start, middle and end positions of syllables. Here, three F_0 values for each syllable were chosen to capture the broad shape of the intonation contour of syllables. The activation function for the units at the input and output layers is linear, whereas the activation function used at hidden layers is nonlinear. The extracted 35 dimensional input vectors representing positional, contextual, phonological and articulatory features are presented as input, and the corresponding prosodic parameters 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. 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 that 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 35*L* 68*N* 17*N* 1*L*, 35*L* 72*N* 19*N* 3*L* and 35*L* 71*N* 17*N* 1*L* are obtained for duration, intonation and intensity models with minimum generalization errors. 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. Gradient descent with momentum weight function is used as an adaptation learning function. The back-propagation network learns by examples. So, we use input–output examples to show the network what type of behavior is expected, and the back propagation algorithm allows the network to adapt. 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 177,820 syllables is used for modeling the duration, intonation and intensity. 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 overfitting 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 overfitting of the neural network. The mean-squared errors for training, validation and testing of the FFNN for modeling the duration, intonation and intensity patterns of the sequence of syllables are shown in Fig. 4. The mean-square error decreases with an increasing number of

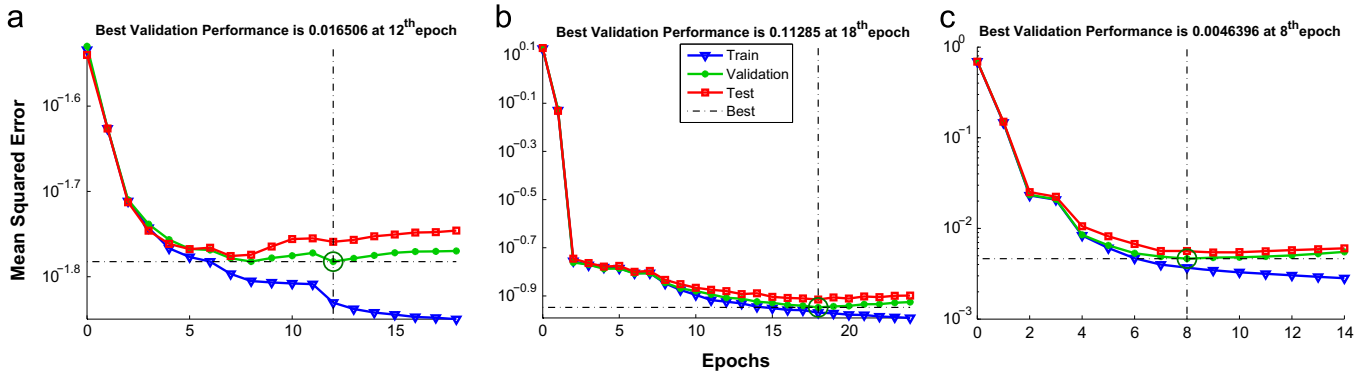

Fig. 4. Train, validation and test errors of the FFNN model developed for modeling the (a) durations, (b) intonation, and (c) intensities of syllables.

epochs during training: it starts at a large value, decreases rapidly, and then continues to decrease slowly as the network makes its way to local minimum of error. The mean-square error of validation subset also decreases monotonically to a minimum, it then starts to increase as the training continues. This indicates that network learned beyond this point is essentially the noise contained in the training data. This heuristic suggests that the minimum point on the validation error curve can be used as a sensible criterion for stopping the training session. The number of epochs needed for training depends on the behavior of the validation error. The training of the network stops once the validation error stops decreasing continuously. The validation error is monitored by keeping regular validation checks at each epoch. In this work, 6 validation checks are used to monitor the validation error. The learning ability of the network from the training data can be observed from the training error. The decreasing nature of the errors shown in Fig. 4 represents, network is capturing the implicit relation between the input and the output.

6. Evaluation of prosody models

The prediction performance of the FFNN models is evaluated by means of objective and subjective tests. In this study, the durations, F_0 values and intensities of syllables are also modeled using LR and CART with the same PCPA features used for FFNN. LR and CART are well known prosody models used for developing TTS systems in Festival framework. The performance of the FFNN model is compared with LR and CART models developed by PCPA features, in predicting the durations, average F_0 values and intensities of syllables. In this work, CART and LR models are developed by using the codes available in Festival [16]. The prediction performance of the proposed FFNN models is also compared with state of the art machine learning approaches such as Support Vector Machine (SVM) and Extreme Learning Machine (ELM). Here, SVM and ELM models are developed using the codes from [50,51], respectively. The well-known Gaussian kernel function is used in SVM as well as ELM. For achieving good generalization performance, the cost (C) and kernel (γ) parameters have to be selected appropriately. We have explored 50 different values $\{2^{-24}, 2^{-23}, \ldots\ldots, 2^{24}, 2^{25}\}$ for each of the parameters C and γ. The values of (C, γ) used for the optimal performance of SVM for predicting the durations, F_O and intensity values of syllables are (1024, 1), (512, 1) and (512, 1), respectively. In case of ELM, the values of (C, γ) are (32, 2048), (32, 2048) and (64, 4096) respectively, for duration, intonation and intensity models. For achieving the optimal performance using FFNNs, we have explored various network structures by varying the number of hidden layers and number of neurons in each hidden layer. The network structures chosen for achieving the optimal performance are 35*L* 68*N* 17*N* 1*L*, 35*L* 72*N* 19*N* 3*L* and 35*L* 71*N* 17*N* 1*L* for predicting duration, intonation and intensity, respectively. In case of CART model, the prediction performance is optimized by properly choosing the set of questions. In LR model, the weights are chosen such that the mean squared error has to be minimized for the overall training data. Since, the objective of the paper is to develop efficient prosody models for syllable based unit selection text-to-speech synthesis system, we have carried out the comparison of detailed objective and subjective measures with respect to well known prosody models developed using CART and LR methods used in TTS framework. The details of objective and subjective tests used for evaluating the models are discussed in the following subsections.

Table 4
Performance of the LR, CART, FFNN, SVM and ELM models for predicting the duration values of syllables using PCPA features (PCPA: Positional, Contextual, Phonological and Articulatory features).

Model	% Predicted syllables within deviation					Objective measures		
	2%	5%	10%	15%	25%	μ (ms)	σ (ms)	γ
LR	6.98	15.77	29.99	43.07	63.11	48.07	38.49	0.74
CART	7.03	16.73	32.15	43.76	65.38	44.54	39.78	0.77
SVM	7.21	17.76	35.92	49.27	71.87	40.12	34.82	0.81
ELM	8.03	18.27	35.27	51.16	72.13	39.57	35.26	0.82
FFNN	7.96	18.88	35.14	50.63	72.56	39.04	35.09	0.83

6.1. Objective evaluation

The duration, intonation and intensity models are evaluated with the syllables in the test set. The durations, F_0 values and intensities of each syllable in the test set are predicted using FFNN, by presenting the feature vector of each syllable as input to the network. The performance of the FFNN models for predicting the durations, F_0 values and intensities of sequence of syllables with PCPA features are given in Tables 4–6. The percentage of syllables predicted within different deviations from their actual duration, F_0 and intensity values are shown in Tables 4–6.

The deviation (D_i) is calculated as follows:

$$D_i = \frac{|x_i - y_i|}{x_i} \times 100 \quad (2)$$

where $x_i$and y_i are the actual and predicted prosodic (duration, F_0, intensity) values, respectively.

The prediction accuracy is evaluated by means of objective measures such as average prediction error (μ), standard deviation (σ) and linear correlation coefficient ($\gamma_{X,Y}$) between actual and predicted durations, F_0 values and intensities of the syllables. The first column of Tables 4–6 indicate the models used for predicting the prosodic parameters. Columns 2–6 of Table 4 indicate the

percentage of syllables predicted within different deviations from their actual duration values. Columns 7–9 of Table 4 indicate the objective measures ((μ), (σ) and ($\gamma_{X,Y}$)). The formulas used to compute objective measures are given below

$$\mu = \frac{\sum_i |x_i - y_i|}{N} \tag{3}$$

$$\sigma = \sqrt{\frac{\sum_i d_i^2}{N}}; \quad d_i = e_i - \mu; \quad e_i = x_i - y_i \tag{4}$$

where x_i, y_i are the actual and predicted prosodic values respectively, and e_i is the error between the actual and predicted prosodic values. The deviation in error is d_i, and N is the number of observed prosodic values of the syllables. The correlation coefficient is given by

$$\gamma_{X,Y} = \frac{V_{X,Y}}{\sigma_X \cdot \sigma_Y} \tag{5}$$

$$\text{where } V_{X,Y} = \frac{\sum_i |x_i - \bar{x}| \cdot |y_i - \bar{y}|}{N} \tag{6}$$

where σ_X, σ_Y are the standard deviations for the actual and predicted prosodic values, respectively, and $V_{X,Y}$ is the linear correlation between the actual and predicted prosodic values.

The performance of FFNN models is compared with LR, CART, SVM and ELM models. From Tables 4–6, it is observed that among LR, CART, SVM, ELM and FFNN models, the prediction performance of LR model is very low compared to other models. The lower performance of the linear models can be attributed to their inability to capture the nonlinear (complex) relations present in the data. This indicates that the PCPA features are nonlinearly related to durations, intonation and intensities of syllables rather than simple linear combination of features. Among all models FFNN model performs better in predicting the prosodic patterns of sequence of syllables. From this we can hypothesize, that the neural network model captures the inherent complex relationships between PCPA features, and durations, intonation and intensities of syllables reasonably well compared to other models. Among CART, SVM, ELM and FFNN models, performance of CART models is slightly lower, compared to remaining 3 models. Performance of SVM and ELM models is comparable to FFNN models.

In this work, we have also explored FFNN models with single hidden layer. The prediction performance of duration, intonation and intensity models developed using single hidden layer FFNNs are given in Table 7. From the results (see Tables 4–7), it is observed that the performance of single hidden layer FFNN models is low, compared to two hidden layer FFNN models. The main reason for better performance of two hidden layer FFNN models is due to the ability to capture the complex non-linear relations present at different levels. The proposed PCPA features represent highly complex non-linear information at various levels: (i) positional information of syllables at word and phrase levels, (ii) positional information of words at sentence level, (iii) co-articulation knowledge is represented by preceding and following syllables, (iv) microlevel information of syllables is represented by phonetic segments and (v) highly non-linear movements of articulators are represented by articulatory features. For modeling the above mentioned complex non-linear relations across various features, two hidden layers may be more appropriate, compared to single layer. It is known that, a two hidden layer neural network has the ability to capture global and local nonlinear relations effectively [1,2].

The prediction performance of the neural network model depends on the nature of the training data used. Distributions of

Table 5
Performance of the LR, CART, SVM, ELM and FFNN models for predicting the F_0 values of syllables using PCPA features.

Model	F_0 Position in syllable	% Predicted syllables within deviation					Objective measures		
		2%	5%	10%	15%	25%	μ (Hz)	σ (Hz)	γ
LR	Start	10.11	21.32	45.39	61.50	80.91	43.72	39.10	0.74
	Middle	13.93	29.17	52.65	78.93	87.56	33.96	28.19	0.78
	End	9.01	19.97	45.32	59.99	77.60	44.20	34.87	0.73
CART	Start	12.56	27.15	51.97	70.50	82.19	41.01	33.93	0.77
	Middle	16.93	35.74	60.09	77.73	88.92	32.53	28.74	0.80
	End	9.13	23.99	47.75	68.41	79.97	42.15	37.73	0.76
SVM	Start	13.86	30.87	55.72	72.62	87.52	32.49	28.03	0.81
	Middle	18.62	39.93	64.98	79.53	89.73	29.17	24.52	0.81
	End	10.13	22.50	49.98	68.21	84.16	36.11	30.59	0.79
ELM	Start	14.06	31.38	55.93	72.79	87.46	32.11	27.23	0.82
	Middle	18.26	40.14	65.82	80.53	90.41	28.84	25.02	0.82
	End	10.42	22.79	50.84	69.49	85.17	35.29	31.06	0.79
FFNN	Start	14.35	31.26	55.25	73.16	87.34	32.19	27.13	0.82
	Middle	19.17	40.58	66.35	80.81	90.12	28.31	24.74	0.83
	End	10.57	23.18	51.72	69.89	84.99	35.52	30.91	0.79

Table 6
Performance of the LR, CART, SVM, ELM and FFNN models for predicting the intensity values of syllables using PCPA features.

Model	% Predicted syllables within deviation				Objective measures		
	1%	3%	5%	7%	μ (dB)	σ (dB)	γ
LR	13.97	37.96	60.89	76.2	3.40	2.66	0.69
CART	16.67	48.22	71.63	85.53	2.80	2.38	0.80
SVM	17.48	48.16	72.67	86.95	2.63	2.04	0.80
ELM	18.05	48.97	72.88	87.03	2.61	2.02	0.81
FFNN	18.02	49.24	73.23	87.16	2.60	2.03	0.81

Table 7
Prediction performance of duration, intonation and intensity models developed using single hidden layer FFNNs.

Model	% Predicted syllables within deviation					Objective measures		
	2%	5%	10%	15%	25%	μ	σ	γ
Dur	7.28	17.97	33.86	48.23	68.56	40.86	35.93	0.81
Intonation	13.52	29.96	53.42	70.93	85.18	33.49	28.12	0.80
	17.96	38.42	65.12	77.63	88.37	29.83	25.17	0.80
	8.93	21.86	48.27	68.18	82.19	36.85	31.79	0.77

Intensity model	% Predicted syllables within deviation				Objective measures		
	1%	3%	5%	7%	μ	σ	γ
Intensity	16.97	48.92	72.12	86.52	2.67	2.27	0.79

the durations, F_0 values and intensities of syllables (see Fig. 1) indicate that majority of the prosodic values are concentrated around mean of the distribution. This kind of training data forces the model to be biased towards mean of the distribution. This results in high prediction error at extreme values (lower duration, F_0 and intensity values are overestimated and higher values are underestimated). This problem can be handled in two ways: using (1) Preprocessing methods and (2) Postprocessing methods. In this work, we proposed a preprocessing methodology based on histogram equalization, with which the training data is prepared to yield approximately uniform distribution. We have also explored a postprocessing method, which will modify the predicted values by imposing the piecewise linear transformation to overcome the biasing of the network due to the implicit distribution of training data.

Here, we have developed duration, intonation and intensity models using preprocessing method mentioned above, i.e., the training data is prepared to yield approximately uniform distribution using histogram equalization technique. Fig. 5 illustrates the nature of the original distribution of syllable duration values and the modified distribution of duration values using histogram equalization. For improving the accuracy of prediction, we have also incorporated a postprocessing method based on piecewise linear transformation [52–54,7]. In this work, we have incorporated preprocessing and postprocessing methods separately. The prediction performance of FFNN models after incorporating preprocessing and postprocessing methods is given in Table 8. The first two rows of the table indicates prediction accuracy of FFNNs after incorporating preprocessing method mentioned above. Third and fourth rows of the table indicate the prediction accuracy of FFNNs after incorporating postprocessing method. The last row indicates the prediction accuracy of intensity models after incorporating preprocessing and postprocessing methods. The first 4 rows of the table contains the performance measures related to duration and intonation models. From the results it is observed that the accuracy of prediction has been improved for all prosodic parameters in both cases (see Tables 4–6, and 8).

From these two studies, it is observed that the overall accuracy of prediction is better by using preprocessing method compared to postprocessing method (see Table 8). From the philosophical point of view also, preprocessing approach seems to be more logical, because for avoiding the bias in the prediction, the training data is modified (restructured) such that the bias in the prediction is minimized. Whereas in the case of postprocessing method, the bias in the prediction is corrected after the erroneous prediction due to the implicit biased distribution of the training data.

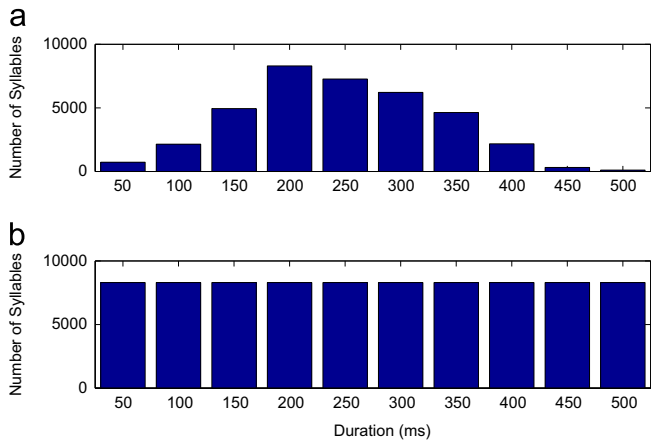

Fig. 5. Histogram representation of (a) original distribution of syllable duration values and (b) modified distribution of syllable duration values using histogram equalization.

For improving the accuracy of prediction, in addition to preprocessing and postprocessing methods mentioned above, multimodels may be explored. In the context of neural networks, multimodels indicate usage of more than one neural network for carrying out the desired task. In this work, we may use more than one neural network for predicting the individual prosodic parameter values. Since, the range of prediction is large and distribution of training data is nonuniform, the accuracy of prediction will be poor. This problem can be addressed by adopting multiple models. Here, by using multiple models, the range of prediction will be reduced and the distribution of data within smaller intervals will be uniform to some extent. While adopting the multiple models, dividing the training data blindly into sub-groups based on prosodic value (output to be predicted) may not be logical. Instead of dividing the data into subgroups based on prosodic value, it will be more meaningful and logical, if we divide or group the syllables based on production and articulation constraints. In this work, the syllables may be divided into groups based on inherent production characteristics of vowel such as height of the tongue hump, position of the tongue hump and roundness of lips. Similarly, we may group the syllables based on the production characteristics of consonants present in syllable.

6.2. Subjective evaluation

Naturalness and intelligibility are the 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 LR, CART and FFNN based duration, intonation and intensity models into the TTS system. The overall architecture representing the incorporation of prosodic models in TTS system is given in Fig. 6.

In this study, prosody models (duration, intonation and intensity models) were developed in offline and integrated to TTS system as shown in Fig. 6. The TTS system used in this study is syllable based concatenative speech synthesis, where syllables are chosen using unit selection framework. The speech database consists of natural waveforms and it is manually labeled in terms syllable identities. Each syllable also carries its prosodic information such as duration, F_O and intensity. Using feature extraction module, PCPA features are derived for each syllable. Derived PCPA features and prosodic values associated to syllables present in database are used for building prosody models. In this work, we have explored FFNNs, CART and LR models for developing the prosody models and then integrated them to syllable based Festival TTS system. During synthesis, text correspond to speech to be spoken is given as input to TTS system. Then, text analysis module derive the sequence of syllable-labels and feature extraction module derive the PCPA features associated to syllable-labels. The derived PCPA features are given as input to the integrated prosody models, and the required prosodic parameters are predicted by these prosody models. Now, the waveforms associated to sequence of syllable labels are picked up and concatenated using unit selection module based appropriate target and concatenation costs. The quality of synthesized speech is evaluated using subjective listening tests.

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 (1-Unsatisfactory, 2-Poor, 3-Fair, 4-Good and 5-Excellent) for each of the synthesized sentences. The subjective listening tests are carried out for the

Table 8
Prediction performance of FFNN models after incorporating preprocessing and postprocessing methods.

Performance of FFNN using preprocessing method

Prosody parameter	% Predicted syllables within deviation					Objective measures		
	2%	5%	10%	15%	25%	μ	σ	γ
Dur	9.07	22.18	39.48	56.87	77.59	37.31	34.16	0.78
F_0	16.56	34.82	59.13	75.82	89.28	30.86	26.02	0.84
	22.12	43.86	68.92	82.36	91.84	27.13	23.86	0.84
	11.93	27.82	54.36	73.18	88.67	33.21	29.27	0.81

Performance of FFNN using postprocessing method

Prosody parameter	% Predicted syllables within deviation					Objective measures		
	2%	5%	10%	15%	25%	μ	σ	γ
Dur	8.34	20.03	36.79	52.47	74.83	38.82	34.76	0.83
F_0	14.96	32.87	57.12	73.92	87.96	31.87	26.89	0.82
	20.04	41.29	67.58	81.19	90.75	27.97	24.52	0.83
	10.93	23.76	52.28	70.53	85.78	34.79	30.47	0.80

Performance of intensity model with pre and postprocessing methods

Intensity model	% Predicted syllables within deviation				Objective measures		
	1%	3%	5%	7%	μ	σ	γ
Preproces	19.15	52.16	77.52	88.93	2.42	1.94	0.83
Postproces	18.69	49.78	74.62	87.53	2.52	1.98	0.82

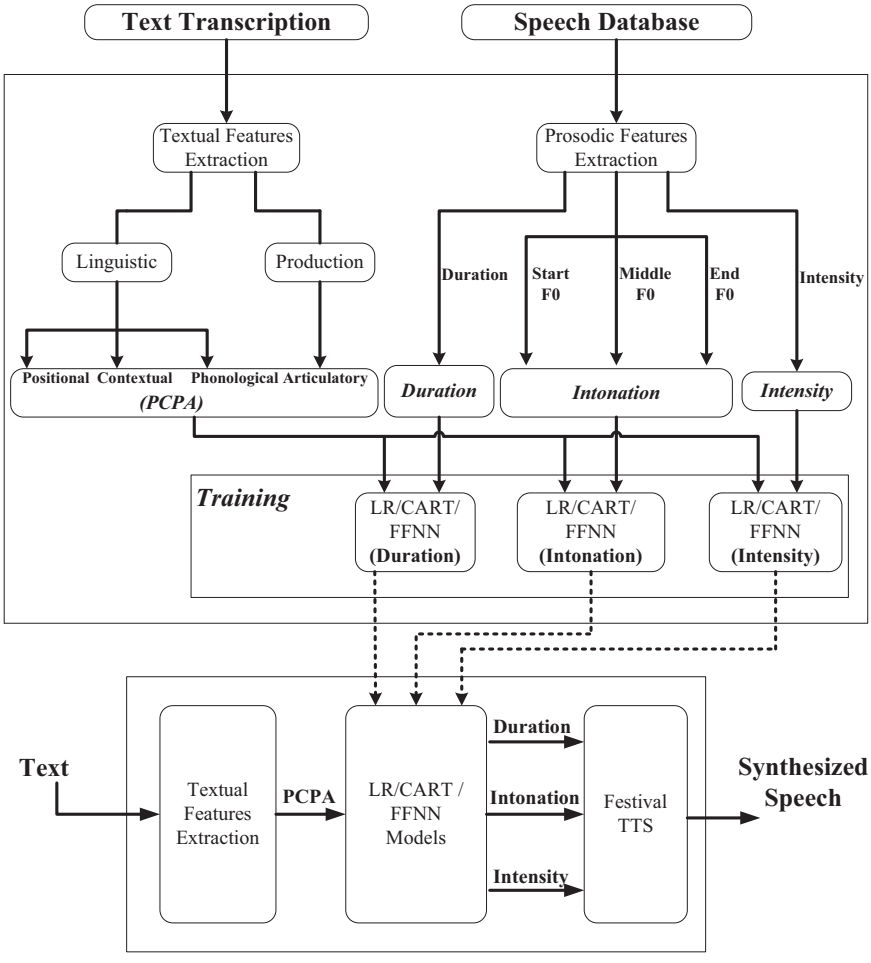

Fig. 6. Incorporation of duration, intonation and intensity models in TTS system.

Table 9
Mean opinion scores for the quality of synthesized speech of TTS after incorporating the Prosodic models.

Prosodic feature	Model (input features)	MOS (confidence level)	
		Naturalness	*Intelligibility*
Duration	CART (Festival)	3.23 (> 99.5)	2.58 (> 99.5)
	LR (PCPA)	3.31 (> 99.5)	2.67 (> 99.5)
	CART (PCPA)	3.46 (> 95.0)	2.71 (> 99.0)
	FFNN (PCPA)	**3.53**	**2.86**
Intonation	CART (Festival)	3.24 (> 99.0)	2.60 (> 99.5)
	LR (PCPA)	3.29 (> 97.5)	2.64 (> 99.0)
	CART (PCPA)	3.37 (> 90.0)	2.70 (> 97.5)
	FFNN (PCPA)	**3.41**	**2.79**
Intensity	LR (PCPA)	3.23 (> 99.0)	2.56 (> 99.0)
	CART (PCPA)	3.36 (> 95.0)	2.66 (> 95.0)
	FFNN (PCPA)	**3.41**	**2.73**

synthesized sentences generated by LR, CART and FFNN models developed for duration, intonation and intensity models using the PCPA features, and also duration and intonation models using the Festival default features. The mean opinion scores (MOS) are calculated for both naturalness and intelligibility of the synthesized speech. Table 9 shows the MOS values for the synthesized speech correspond to different duration, intonation and intensity models. From Table 9, it is observed that the MOS values for naturalness and intelligibility of the proposed FFNN model seems to be 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. Naturalness can be improved to some extent by combining all the proposed models developed in this work.

The significance of the differences in the pairs of the mean opinion scores for intelligibility and naturalness is tested using hypothesis testing. The level of confidence for the observed differences in the pairs of MOSs between FFNN model and other models are given in brackets in Table 9. The numbers in brackets represent the confidence levels of FFNN model compared against respective systems. The level of confidence is computed as follows:

$$t = \frac{(m1 - m2)}{\sqrt{\frac{((n_1-1)\times s_1 \times s_1)+((n_2-1)\times s_2 \times s_2)}{(n_1+n_2-2)} \times \left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \quad (7)$$

where m_1 and m_2 are mean opinion score values, n_1 and n_2 are number of scores, and s_1 and s_2 are standard deviations of the scores of FFNN and LR or CART models for both naturalness and intelligibility, respectively. From the obtained 't' values confidence value is calculated using student's-t test.

From Table 9, it is observed that the level of confidence is high (> 90.0) in all cases. This indicates that the differences in the pairs of MOS in each case is significant. From this study, we conclude that the quality of speech using proposed FFNN model is significantly better than all other models at the perceptual level, and also the proposed syllable specific features are significantly better compared to default phone based features used by the Festival in predicting the durations and F_0 values.

The subjective evaluation is also carried out with incorporation of all the proposed FFNN based prosodic models. The MOS values of individual models along with different combinations is shown in Table 10. From the MOS values given in Table 10, it is observed that the durations, intonation and intensities of the syllables are important to bring naturalness in synthesized speech, but durations and intonation have more influence at the perceptual level

Table 10
Mean opinion scores for the quality of synthesized speech of TTS after incorporating the duration, intonation and intensity models.

Model	Mean opinion score (confidence level)	
	Intelligibility	*Naturalness*
Without prosody	3.16 (> 99.5)	2.52 (> 99.5)
Duration model	3.53 (> 99.5)	2.86 (> 99.5)
Intonation model	3.41 (> 99.5)	2.79 (> 99.5)
Intensity model	3.41 (> 99.5)	2.73 (> 99.5)
Duration and intonation	3.73(> 99.5)	3.01 (> 99.5)
Intonation and intensity	3.62(> 99.5)	2.96 (> 99.5)
Duration and intensity	3.68(> 99.5)	3.13 (> 99.0)
All models	**4.01**	**3.29**

compared to intensities. From Table 10, it is observed that there is an improvement in the naturalness with the inclusion of intensity model to the duration and intonation models. From this we can hypothesize, that the prosodic parameter intensity also plays a major role in addition to duration and intonation in improving the quality of the TTS system. The highest quality is observed when all these models are incorporated into the baseline TTS system.

For analyzing the accuracy of the prediction of prosody (duration, intonation and intensity) models, we also conducted the listening tests for assessing the intelligibility and naturalness on the speech without incorporating the prosody. Here, the synthesized waveform will be generated at the time of synthesis using only linguistic features provided by linguistic module, without using prosodic information from prosodic module for picking the units from the speech database. The MOS values of these listening tests are given in the first row in Table 10. It is observed, that the MOS values without considering prosodic information is low compared to the MOS values using predicted durations, intonation and intensity features for selecting the units.

The level of confidence for the observed differences in the pairs of MOSs between all FFNN models and other models are given in bracket in Table 10. From Table 10, it is observed that the level of confidence is high (> 99.0) for both naturalness and intelligibility in all cases. From this study, we conclude that in view of perception, prosody characteristics will play a major role in synthesizing the intelligible and natural speech. From Table 10, it is also observed that the combination of different models is significant than individual models. Objective and subjective evaluation results indicate that the proposed neural network models have shown to be successful in predicting the appropriate prosody characteristics suitable for speech synthesis applications.

7. Influence of individual features for predicting the prosodic parameters of syllables

For studying the effect of positional, contextual, phonological and articulatory features on durations, intonation and intensities of syllables, separate models were developed. The features representing the positional factors are: (a) Syllable position in the sentence (three-dimensional feature), (b) syllable position in the word (three-dimensional feature), (c) word position in the sentence (three-dimensional feature) and (d) syllable identity (four-dimensional feature). Features representing the contextual factors are the identities of the present syllable, its previous and following syllables. Features representing the phonological factors are the syllable nucleus (three-dimensional feature) and syllable identity (four-dimensional feature). Features representing the articulatory factors are 11 dimensional features described in Table 3 and syllable identity (four-dimensional feature).

From the results, it is observed that the durations, F_0 values and intensities of the syllables depend on positional, contextual, phonological and articulatory features. However, positional features seem to perform slightly better compared to remaining features for predicting the durations and F_0 values of syllables, whereas contextual features seem to perform slightly better compared to remaining features for predicting the syllable intensities. But, the combination of all features outperform the individual features for predicting the durations, F_0 values and intensities of syllables.

8. Summary and conclusions

The baseline TTS system was developed with Festival framework using syllable as the basic unit. The prediction accuracy of the duration and intonation modules present in the baseline TTS is found to be poor. Therefore, we proposed syllable specific features for modeling the durations, F_0 values and intensities of syllables. The syllable specific features are represented by linguistic and production constraints. The linguistic constraints are represented by positional, contextual and phonological features, and the production constraints are represented by articulatory features. Feedforward neural networks were proposed for predicting the durations, F_0 values and intensities of the syllables. The performance of the neural network models is compared with LR, CART, SVM and ELM models. Preprocessing and postprocessing methods were proposed for improving the prediction accuracy by minimizing the bias towards mean of the distribution. The effect of individual features was examined by modeling the durations, F_0 values and intensities of the syllables with these features as input to the models. The prediction accuracy of the models is further improved by imposing the prosodic constraints from the developed models. The evaluation of quality of TTS system is carried out by integrating the proposed duration, intonation and intensity models into the baseline TTS system. In this work, all prediction models are developed using single feedforward neural network. Hierarchical neural networks (HNN) may be explored in future for improving the accuracy of prediction.

Acknowledgments

This work is carried out as a part of the project “Development of TTS systems for Indian languages” sponsored by Department of Information Technology (DIT), Government of India (Grant Number: 11(8)/2008-HCC(TDIL), Dt. 13-02-2009). We also acknowledge the research scholars of IIT Kharagpur for their voluntary participation in listening tests to evaluate the quality of Bengali TTS system developed in this work with the integration of proposed prosodic models.

References

[1] S. Haykin, Neural Networks: A Comprehensive Foundation, Pearson Education Asia, Inc., New Delhi, India, 1999.
[2] B. Yegnanarayana, Artificial Neural Networks, Prentice-Hall, New Delhi, India, 1999.
[3] Qiao Cai, Sheng Chen, Xiaochen Li, Nansai Hu, Haibo He, Yu-Dong Yao, Joseph Mitola, An Integrated Incremental Self-organizing Map and Hierarchical Neural Network Approach for Cognitive Radio Learning, in: International Joint Conference on Neural Networks (IJCNN-2010), July 2010, pp. 1–6.
[4] K. Yamauchi, N. Yamaguchi, N. Ishii, Incremental learning methods with retrieving of interfered patterns, IEEE Trans. Neural Netw. 10 (November (6)) (1999) 1351–1365.
[5] K.S. Rao, B.Yegnanarayana, Two-stage duration model for Indian languages using neural networks, in: Lecture Notes in Computer Science : Neural Information Processing, vol. 3316, 2004, pp. 1179–1185.
[6] A.S.M. Kumar, Intonation knowledge for speech systems for an Indian language (PhD thesis), Department of Computer Science and Engineering, Indian Institute of Technology, Madras, Chennai, India, January 1993.
[7] K.S. Rao, B. Yegnanarayana, Intonation modeling for Indian languages, Comput. Speech Lang. 23 (April) (2009) 240–256.
[8] K.S. Rao, B. Yegnanarayana, Modeling durations of syllables using neural networks, Comput Speech Lang. 21 (April) (2007) 282–295.
[9] N.P. Narendra, K.S. Rao, K. Ghosh, V.R. Reddy, S. Maity, Development of syllable-based text to speech synthesis system in Bengali, Int. J. Speech Technol. Springer, 14 (3), (2011) 167–181.
[10] K.S. Rao, B. Yegnanarayana, Neural network models for text-to-speech synthesis, in: The Fifth International Conference on Knowledge Based Computer Systems (KBCS-2004), December 2004, pp. 520–530.
[11] K.S. Rao, B. Yegnanarayana, Impact of constraints on prosody modeling for Indian languages, in: The Third International Conference on Natural Language Processing (ICON-2004), December 2004, pp. 225–236.
[12] D.H. Klatt, Synthesis by rule of segmental durations in English sentences, in: B. Lindblom, S. Ohman (Eds.), Frontiers of Speech Communication Research, Academic Press, New York, 1979, pp. 287–300.
[13] N. Umeda, Linguistic rules for text-to-speech synthesis, Proc. IEEE 64 (April) (1976) 443–451.
[14] L. Lee, C. Tseng, M. Ouh-Young, The synthesis rules in a Chinese text-to-speech system, IEEE Trans. Acoust. Speech Signal Process. 9 (4) (1989) 1309–1320.
[15] J.P.H.V. Santen, Assignment of segment duration in text-to-speech synthesis, Comput. Speech Lang. 8 (April) (1994) 95–128.
[16] A.W. Black, K. Lanzo, Building synthetic voices in the festival speech synthesis system, December 2009. ⟨http://www.festvox.org⟩.
[17] M. Riley, Tree-based modeling for speech synthesis, in: G. Bailly, C. Benoit, T. Sawallis (Eds.), Talking Machines: Theories, Models and Designs1992, pp. 265–273.
[18] W.N. Campbell, Analog i/o nets for syllable timing, Speech Commun. 9 (February) (1990) 57–61.
[19] S.R.R. Kumar, B. Yegnanarayana, Significance of durational knowledge for speech synthesis in Indian languages, in: Proceedings of IEEE Region 10 Conference on Convergent Technologies for the Asia-Pacific, Bombay, India, November 1989, pp. 486–489.
[20] S.R.R. Kumar, Significance of durational knowledge for a text-to-speech system in an Indian language (Master's thesis), Department of Computer science and Engineering, Indian Institute of Technology Madras, March 1990.
[21] K.K. Kumar, Duration and intonation knowledge for text-to-speech conversion system for Telugu and Hindi (Master's thesis), Department of Computer Science and Engineering, Indian Institute of Technology Madras, May 2002.
[22] N.S. Krishna, H.A. Murthy, Duration modeling of Indian languages Hindi and Telugu, in: The 5th ISCA Speech Synthesis Workshop, Pittsburgh, USA, May 2004, pp. 197–202.
[23] K.S. Rao, B. Yegnanarayana, Modeling syllable duration in Indian languages using neural networks, in: IEEE International Conference on Acoustics, Speech Signal Processing (ICASSP), May 2004, pp. 1179–1185.
[24] L. Mary, K.S. Rao, S. Gangashetty, B. Yegnanarayana, Neural network models for capturing duration and intonation knowledge for language and speaker identification, in: The Eighth International Conference on Cognitive and Neural Systems (ICCNS), May 2004.
[25] J.B. Pierrehumbert, The Phonology and Phonetics of English Intonation (PhD thesis), MIT, MA, USA, 1980.
[26] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, J. Hirschberg, ToBI: a standard for labeling English prosody, in: Proceedings of International Conference on Spoken Language Processing, Banff, Alberta, 1992, pp. 867–870.
[27] N. Gronnum, The Groundworks of Danish Intonation: An Introduction, Museum Tusculanum Press, Copenhagen, 1992.
[28] H. Fujisaki, Dynamic characteristics of voice fundamental frequency in speech and singing, in: P.F. MacNeilage (Ed.), The Production of Speech, Springer-Verlag, New York, USA, 1983, pp. 39–55.
[29] J. t'Hart, R. Collier, A. Cohen, A Perceptual Study of Intonation, Cambridge University Press, Cambridge, 1990.
[30] P.A. Taylor, Analysis and synthesis of intonation using the tilt model, J. Acoust. Soc. Am. 107 (March) (2000) 1697–1714.
[31] M.S. Scordilis, J.N. Gowdy, Neural network based generation of fundamental frequency contours, in Proceedings of IEEE International Conference Acoustics, Speech, Signal Processing, Glasgow, Scotland, vol. 1, May 1989, pp. 219–222.
[32] M. Vainio, T. Altosaar, Modeling the microprosody of pitch and loudness for speech synthesis with neural networks, in: Proceedings of International Conference on Spoken Language Processing, Sydney, Australia, December 1998.
[33] S.H. Hwang, S.H. Chen, Neural network-based F0 text-to-speech synthesizer for Mandarin, IEEE Proc. Image Signal Process. 141 (December) (1994) 384–390.
[34] J. Buhmann, H. Vereecken, J. Fackrell, J.P. Martens, B.V. Coile, Data driven intonation modeling of 6 languages, in: Proceedings of International Conference on Spoken Language Processing, Beijing, China, vol. 3, October 2000, pp. 179–183.
[35] C. Traber, F0 generation with a database of natural f0 patterns and with a neural network, in: C. Benoit, T.R. Wawallis, G. Bailey (Eds.), Talking Machines: Theories, Models, and Designs, Elsevier Science, Amsterdam, The Netherlands, 1992, pp. 287–304.
[36] A.S.M. Kumar, S. Rajendran, B. Yegnanarayana, Intonation component of text-to-speech system for Hindi, Comput. Speech Lang. 7 (1993) 283–301.

[37] E. Kidder, Tone, Intonation, Stress and Duration, Coyote Papers: Working Papers in Linguistics, Linguistic Theory at the University of Arizona, vol. 16, pp. 55–66, 2008.
[38] A.E. Turk, J.R. Sawusch, The processing of duration and intensity cues to prominence, J. Acoust. Soc. Am. 99 (June) (1996) 3782–3790.
[39] M. Vainio, T. Atosaar, Pitch, loudness, and segmental duration correlates: towards a model for the phonetic aspects of Finnish prosody, in Proceedings of International Conference on Spoken Language Processing, Philadelphia, PA , USA, vol. 4, August 1996, pp. 2052–2055.
[40] J. Lee, J. Kang, D. Kim, S. Sung, Energy contour generation for a sentence using a neural network learning method, In" Proceedings of International Conference on Spoken Language Processing, 1998, pp. 1991–1994.
[41] J.P.R. Teixeira, Modeling of intonation for speech synthesis (PhD thesis), Faculdade de Engenharia da Universidade do Porto, Electrotechnical and Computer Engineering, May 2004.
[42] R.H. Mannel, Modeling of the segmental and prosodic aspects of speech intensity in synthetic speech, In: Proceedings of International Conference on Speech Science and Technology, Melbourne, December 2002, pp. 538–543.
[43] F. Tesser, Emotional speech synthesis: from theory to application (PhD thesis), International Doctorate School in Information and Communication Technologies, DIT - University of Trento, Italy, February 2005.
[44] K. Ghosh, R.V. Reddy, N.P. Narendra, S. Maity, S.G. Koolagudi, K.S. Rao, Grapheme to Phoneme Conversion in Bengali for Festival based TTS Framework, in: The Eighth International Conference on Natural Language Processing (ICON), Macmillan Publishers, India, 2010.
[45] K. Murty, B. Yegnanarayana, Epoch extraction from speech signals, IEEE Trans. Audio Speech Lang. Process. 16 (2008) 1602–1613.
[46] K. S. Rao, Acquisiton and incorporation prosody knowledge for speech systems in Indian languages (PhD thesis), Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai, India, May 2005.
[47] I.P. Association, Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet, University of Cambridge, Cambridge, United Kingdom, 1999.
[48] R. Ishida, Bengali Script Notes, ⟨http://people.w3.org/rishida/scripts/bengali/⟩.
[49] V.R. Reddy, K.S. Rao, Intonation modeling using FFNN for syllable based Bengali text to speech synthesis, In: Proceedings of International Conference Computer and Communication Technology, MNNIT, Allahabad, 2011, pp. 334–339.
[50] C.-C. Chang, C.-J. Lin, Libsvm: a library for support vector machines, ACM Trans. Intell. Syst. Technol. (2011).
[51] G.-B. Haung, H. Zhou, X. Ding, R. Zhang, Extreme learning machine for regression and multiclass classification, IEEE Trans. Syst. Man Cybern.—Part B: Cybern. 42 (April) (2012) 513–529.
[52] J.R. Bellegarda, K.E.A. Silverman, Improved duration modeling of English phonemes using a root sinusoidal transformation, In: Proceedings of International Conference Spoken Language Processing, December 1998, pp. 21–24.
[53] J.R. Bellegarda, K.E.A. Silverman, K. Lenzo, V. Anderson, Statistical prosodic modeling: from corpus design to parameter estimation, IEEE Trans. Speech Audio Process. 9 (January) (2001) 52–66.
[54] K.E.A. Silverman, J.R. Bellegarda, Using a sigmoid transformation for improved modeling of phoneme duration, in: Proceedings of IEEE International Conference on Acoustics, Speech, Signal Processing, Phoenix, AZ, USA, March 1999, pp. 385–388.

V. Ramu Reddy received the MS (by research) from the School of Information Technology, Indian Institute of Technology Kharagpur, Kharagpur, India, in 2012.

He is currently working as a Systems Engineer in TCS Innovation Lab, Kolkata, India. His research interests are speech signal processing, pattern recognition, machine learning and neural networks. He has published over 30 papers in international journals and conference proceedings in these areas.

K. Sreenivasa Rao received the Ph.D. degree from the Department of Computer Science and Engineering, Indian Institute of Technology (IIT), Chennai, in 2005.

He is currently working as an Associate Professor in the School of Information Technology, Indian Institute of Technology, Kharagpur. His research interests are speech signal processing, multimedia, pattern recognition, machine learning and neural networks. He has published over 200 papers in international journals and conference proceedings in these areas.