Available online at www.sciencedirect.com
ScienceDirect
Computer Speech and Language 23 (2009) 240–256
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

Intonation modeling for Indian languages

K. Sreenivasa Rao [a,*], B. Yegnanarayana [b]

[a] School of Information Technology, Indian Institute of Technology Kharagpur, Kharagpur 721 302, West Bengal, India
[b] International Institute of Information Technology (IIIT), Gachibowli, Hyderabad 500 032, Andhra Pradesh, India

Received 24 June 2007; received in revised form 16 May 2008; accepted 29 June 2008
Available online 8 July 2008

Abstract

In this paper we propose models for predicting the intonation for the sequence of syllables present in the utterance. The term intonation refers to the temporal changes of the fundamental frequency (F_0). Neural networks are used to capture the implicit intonation knowledge in the sequence of syllables of an utterance. We focus on the development of intonation models for predicting the sequence of fundamental frequency values for a given sequence of syllables. Labeled broadcast news data in the languages Hindi, Telugu and Tamil is used to develop neural network models in order to predict the F_0 of syllables in these languages. The input to the neural network consists of a feature vector representing the positional, contextual and phonological constraints. The interaction between duration and intonation constraints can be exploited for improving the accuracy further. From the studies we find that 88% of the F_0 values (pitch) of the syllables could be predicted from the models within 15% of the actual F_0. The performance of the intonation models is evaluated using objective measures such as average prediction error (μ), standard deviation (σ) and correlation coefficient (γ). The prediction accuracy of the intonation models is further evaluated using listening tests. The prediction performance of the proposed intonation models using neural networks is compared with Classification and Regression Tree (CART) models.

Keywords: Intonation models; Feedforward neural network; Prediction accuracy; F_0 of syllable, and Classification and regression tree (CART) models

1. Introduction

During production of speech human beings seem to impose durational constraints and intonation patterns on the sequence of sound units to convey the intended message. For human beings the prosody (duration and intonation) knowledge is naturally acquired, and it is difficult to articulate this knowledge. Acoustic analysis and synthesis experiments have shown that duration and intonation patterns are the two most important prosodic features responsible for the quality of synthesized speech (Huang et al., 2001). The prosody constraints not only possess the characteristic of the speech message and the language, but they also possess characteristics of the speaker. Even in speech recognition, human beings seem to rely on the prosody cues to disambiguate

* Corresponding author. Tel.: +91 322 2282336.
E-mail addresses: ksrao@iitkgp.ac.in, ksrao1969@gmail.com (K. Sreenivasa Rao), yegna@iiit.ac.in (B. Yegnanarayana).

0885-2308/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2008.06.005

errors in the perceived sounds. Thus, acquisition and incorporation of prosody knowledge becomes important for developing speech systems. In this paper we focus on the models to capture the intonation knowledge.

Several studies on modeling the intonation patterns in different languages worldwide are observed in the literature (Benesty et al., 2008; Olive, 1975; Fujisaki et al., 1986; Pierrehumbert et al., 1980; Cosi et al., 2001; Vainio, 2001; Hwang and Chen, 1994; Buhmann et al., 2000). But, for Indian context, there is no systematic study on modeling the intonation patterns. Some studies were performed on modeling the intonation parameters using limited data (Kumar et al., 1993; Kumar, 1993, 1990). These studies with limited amount of data derived rules for predicting the intonation patterns. In the present study, we propose nonlinear models such as neural networks for better prediction of the intonation patterns using large amount of data.

The temporal changes of the fundamental frequency (F_0) i.e., intonation, depends on the information at various levels: (a) segmental coarticulation at the phonemic level; (b) emphasis of the words and phrases in a sentence; and (c) syntax, semantics, prominence and presence of new information in an utterance (Klatt, 1987). In addition to these, changes in the fundamental frequency are also influenced by age, gender, mood and the emotional state of the speaker.

The implicit knowledge of prosody is usually captured using modeling techniques. In a speech signal, the intonation pattern (F_0 contour) corresponding to a sequence of sound units is constrained by the linguistic context of the units, and the constraints in the production of these units. Intonation models can be either rule-based or data-based. Rule-based methods model the intonation patterns using a set of phonological rules (Olive, 1975; Fujisaki et al., 1986; Taylor, 2000; Madhukumar et al., 1991). These rules are inferred by observing the intonation patterns for a large set of utterances with the help of linguists and phoneticians. The relationship between the linguistic features of input text and the intonation (F_0 contour) pattern of an utterance is examined to derive the rules. Although this is done by induction, it is generally difficult to analyze the effect of mutual interaction of the linguistic features at different levels. Hence, the inferred phonological rules for intonation modeling are imprecise and incomplete.

Some of the prominent intonation models reported in the literature are: Tone sequence model, Fujisaki model, IPO (Institute of Perception Research) model and tilt model (Pierrehumbert et al., 1980; Taylor, 2000; Fujisaki, 1983, 1988; t'Hart et al., 1990). In the tone sequence model, the F_0 contour is generated from a sequence of phonologically distinctive tones, which are determined locally, and they do not interact with each other. Tone sequence models do not properly represent the actual pitch variations. No distinction is made on the differences in tempo or acceleration of the pitch moments. The Fujisaki model is hierarchically organized, and includes several components of different temporal scopes. In this model, the F_0 contour is generated by a slowly varying phrase component and a fast varying accent component. The main drawback of this model is that, it is difficult to deal with low accents and slowly rising section of the contour. In the IPO model, intonation analysis is performed in three steps: (1) Perceptually, relevant movements in F_0 contour are stylized by straight lines (known as *copy contour*). (2) Features of the copy contours are expressed in terms of duration and range of F_0 moments. (3) A grammar of possible and permissible combination of F_0 movements is derived. The major drawback observed in the IPO model is that the contours derived from the straight line approximations are quite different from the original. This is due to the difficulty in modeling curves with straight lines. The tilt model provides a robust analysis and synthesis of intonation contours. In the tilt model, intonation patterns are represented by a sequence of events. The event may be an accent, a boundary, silence or a connection between events. The events are described by the following parameters: Starting F_0, rise amplitude, fall amplitude, duration, peak position and tilt. The tilt model gives a good representation of natural F_0 contour and a better description of the type of accent. But the model does not provide the linguistic interpretation for different pitch accents.

Methods based on databases generally depend on the quality and quantity of the available data. Among several database driven methods, Classification and Regression Tree (CART) models and neural network models are more popular (Cosi et al., 2001; Breiman et al., 1984; Dusterhoff et al., 1999; Krishna et al., 2004; Goubanova and King, 2008; Tesser et al., 2004; Vainio and Altosaar, 1998; Vegnaduzzo, 2003). In the CART model a binary branching tree is constructed by feeding the attributes of the feature vectors from top node, and passing through the arcs representing the constraints. The feature vector of a segment represents the positional, contextual and phonological information. The F_0 of a segment is predicted by passing the feature vector through the tree so as to minimize the variance at each terminal node. The tree construction algo-

rithm usually guarantees that the tree fits the training data well. But there is no guarantee that the new and unseen data will be predicted properly. The prediction performance of the CART model depends on the coverage of the training data. Wagon is a popular tool, which is part of the Edinburgh Speech Tools Library is used to build a regression tree (Black et al., 2003). Wagon uses the data variance times the number of feature vectors (the "impurity") to determine the best question to add at each node of the tree, using a standard greedy algorithm to grow the tree. The algorithm stops growing a branch when one of the following is true: (1) all questions about all elements of the feature vector have been asked; (2) all the feature vectors at the current leaf are identical; (3) the number of data points after the next split would fall below a threshold (the "stop value") and (4) the improvement in impurity after the next split would fall below a threshold. In addition, in order to prune an over-trained tree with a small stop value, one can use held-out data. Held-out data is the subset of the data used for testing a tree. The tree is built using the training data, it is then pruned back to where it best matches the held-out data. The advantage of this approach is that it allows the stop value to vary through different parts of the tree depending on how good the prediction is when compared against the held-out data.

Neural network models are known for their ability to capture the functional relation between input–output pattern pairs (Haykin, 1999; Yegnanarayana, 1999). Several models based on neural network principles are described in the literature for predicting the intonation patterns of syllables in continuous speech (Vainio, 2001; Hwang and Chen, 1994; Buhmann et al., 2000; Vainio and Altosaar, 1998; Scordilis and Gowdy, 1989; Sonntag et al., 1997). Scordilis and Gowdy (1989) used neural networks in a parallel and distributed manner to predict the average F_0 value for each phoneme, and the temporal variations of F_0 within a phoneme. The network consists of two levels: macroscopic level and microscopic level. At the macroscopic level, a Feedforward Neural Network (FFNN) is used to predict the average F_0 value for each phoneme. The input to the FFNN consists of the set of phonemic symbols, which represents the contextual information. At the microphonemic level, a Recurrent Neural Network (RNN) is used to predict the temporal variations of F_0 within a phoneme. Marti Vainio and Toomas Altosaar used a three layer FFNN to predict F_0 values for a sequence of phonemes in Finnish language (Vainio and Altosaar, 1998). The features used for developing the models are phoneme class, length of a phoneme, identity of a phoneme, identities of previous and following syllables (context), length of a word and position of a word. Buhmann et al. (2000) used a RNN for developing multi-lingual intonation models. The features used in this work are the universal (language independent) linguistic features such as part-of-speech and type of punctuation, along with prosodic features such as word boundary, prominence of the word and duration of the phoneme.

In this paper, we use a four layer feedforward neural network to predict the F_0 contour for a sequence of syllables. The linguistic and production constraints associated with a sequence of syllables are represented with positional, contextual and phonological features. These features are used to train the neural network to capture the implicit intonation knowledge. Details of these features are discussed in Section 3. In addition to these features, we also examine the effect of intonation and duration constraints on the F_0 of a syllable.

The main objective of this study is to determine whether neural network models can capture the implicit knowledge of the intonation patterns of syllables in a language, in the Indian context. One way to infer this is to examine the error for the training data. If the error is reducing for successive training cycles, then one can infer that the network is indeed capturing the implicit relations in the input–output pairs. We propose to examine the ability of the neural network models to capture the intonation knowledge for speech from various speakers in different Indian languages. We consider three Indian languages (Hindi, Telugu and Tamil) using the syllable as the basic sound unit. The reason for choosing syllable as the basic unit is that, it is a natural and convenient unit for production and perception of speech in Indian languages. In Indian scripts characters generally correspond to syllables. A character in an Indian language script is typically in one of the following forms: V, CV, CCV, CCVC and CVCC, where C is a consonant and V is a vowel.

The paper is organized as follows: the database used for developing the intonation model is described in Section 2. Section 3 discusses the features used as input to the neural network for capturing the knowledge of the F_0 contour for a sequence of syllables. Section 4 gives the details of the neural network model, and discusses its performance in predicting the F_0 values of syllables. In this section, prediction accuracy is extensively analyzed using objective and subjective measures and pre-processing and post-processing methods. The performance of the proposed neural network models is compared with the performance of the CART models.

A summary of this work is given in the final section of the paper along with a discussion on some issues that need to be addressed further.

2. Speech database

The database for this study consists of 19 Hindi, 20 Telugu and 33 Tamil broadcast news bulletins (Khan et al., 2003). In each language these news bulletins were read by male and female speakers. Total durations of speech in Hindi, Telugu and Tamil are 3.5 h, 4.5 h and 5 h, respectively. The speech signal was sampled at 16 kHz and represented as 16 bit numbers. The speech utterances are manually transcribed into text using common transliteration code (ITRANS) for Indian languages (Chopde). The speech utterances are segmented and labeled manually into syllable-like units. Each bulletin is organized in the form of syllables, words and orthographic text representations of the utterances. Each syllable and word file contains the text transcriptions and timing information in number of samples. The fundamental frequencies of the syllables are computed using the autocorrelation of the Hilbert envelope of the linear prediction residual (Prasanna and Yegnanarayana, 2004). The average pitch (F_0) for male speakers and female speakers in the database was found to be 129 and 231 Hz, respectively.

3. Features for developing intonation models

In this study we use 25 features (which form a feature vector) for representing the linguistic context and production constraints of each syllable. These features represent positional, contextual and phonological information of each syllable. Features representing the positional information are further classified based on the position of a word in the phrase and the position of the syllable in a word and phrase.

Syllable position in the phrase: A phrase is delimited by the orthographic punctuation. The syllable position in a phrase is characterized by three features. The first one represents the distance of the syllable from the starting position of the phrase. It is measured in number of syllables, i.e., the number of syllables ahead of the present syllable in the phrase. The second feature indicates the distance of the syllable from the terminating position of the phrase. The third feature represents the total number of syllables in the phrase.

Syllable position in the word: In Indian languages words are identified by spacing between them. The syllable position in a word is characterized by three features similar to the phrase. The first two features are the positions of the syllable with respect to the word boundaries. The third feature is the number of syllables in a word.

Position of the word: The F_0 of a syllable may depend on the position of the word in an utterance. Therefore the word position is used for developing the intonation model. The word position in an utterance is represented by three features. They are the positions of the word with respect to the phrase boundaries, and the number of words in the phrase.

Syllable identity: A syllable is a combination of segments of consonants (C) and vowels (V). In this study, syllables with more than four segments (Cs or Vs) are ignored, since the number of such syllables present in the database is less than 1%. Each segment of a syllable is encoded separately, so that each syllable identity is represented by a 4-dimensional feature vector. Each of the C and V segments are uniquely coded based on their identity (see Appendix A for details).

Context of the syllable: The F_0 of a syllable may be influenced by its adjacent syllables. Hence, for modeling the F_0 of a syllable, the contextual information is represented by the previous and following syllables. Each of these syllables is represented by a 4-dimensional feature vector, representing the identity of the syllable.

Syllable nucleus: Another important feature is the vowel position in a syllable, and the number of segments before and after the vowel in a syllable. This feature is represented with a 3-dimensional feature vector specifying the consonant–vowel structure present in the syllable.

Pitch: The F_0 value of a syllable may be influenced by the pitch value of the preceding syllable. Therefore this information is used in the feature vector of the syllable.

The list of features and the number of input (feature) nodes for a neural network are given in Table 1. These features are coded and normalized before presenting to the neural network model. The details of coding the features are given in Appendix A.

4. Intonation modeling with feedforward neural networks

A four layer feedforward neural network is used for modeling the intonation patterns of syllables. The general structure of the FFNN is shown in Fig. 1. Here, the FFNN model is expected to capture the functional relationship between the input and output feature vectors of the given training data. The mapping function is between the 25-dimensional input vector and the 1-dimensional output. It is known that a neural network with two hidden layers can realize any continuous vector-valued function (Sontag, 1992). The first layer is the input layer with linear units. The second and third layers are hidden layers. The second layer (first hidden layer) of the network has more units than the input layer, and it can be interpreted as capturing some local features in

Table 1
List of the factors affecting the F_0 of syllable, features representing the factors and the number of nodes needed for neural network to represent the features

Factors	Features	# Nodes
Syllable position in the phrase	Position of syllable from beginning of the phrase Position of syllable from end of the phrase Number of syllables in the phrase	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 phrase	Position of word from beginning of the phrase Position of word from end of the phrase Number of words in the phrase	3
Syllable identity	Segments of the syllable (consonants and vowels)	4
Context of the syllable	Previous syllable Following syllable	4 4
Syllable nucleus	Position of the nucleus Number of segments before the nucleus Number of segments after the nucleus	3
Pitch	F_0 of the previous syllable	1

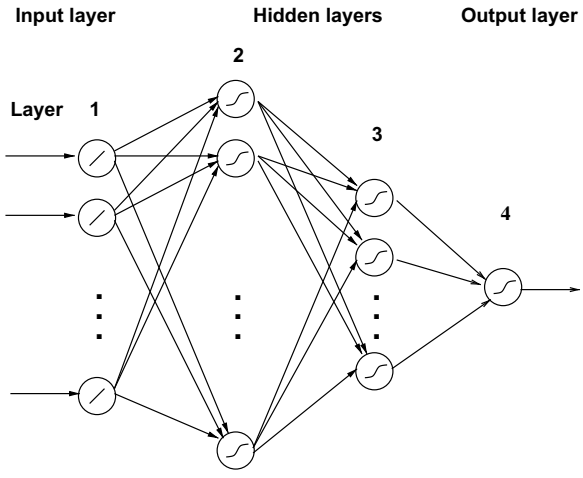

Fig. 1. A four layer feedforward neural network.

the input space. The third layer (second hidden layer) has fewer units than the first layer, and can be interpreted as capturing some global features (Haykin, 1999; Yegnanarayana, 1999). The fourth layer is the output layer having one unit representing the F_0 of a syllable. The activation function for the units at the input layer is linear, and for the units at the hidden layers, it is nonlinear. Generalization by the network is influenced by three factors: the size of the training set, the architecture of the neural network, and the complexity of the problem. We have no control over the first and last factors. Several network structures are explored in this study. The (empirically arrived) final structure of the network is 25L 50N 12N 1N, where L denotes a linear unit, and N denotes a nonlinear unit. The integer value indicates the number of units used in that layer. Note that the structure in terms of # units in the hidden layer is not critical. The # units in the two hidden layers is guided by the heuristic arguments given above. The nonlinear units use $tanh(s)$ as the activation function, where s is the activation value of that unit. All the input and output features are normalized to the range $[-1, +1]$ before presenting to the neural network. The backpropagation learning algorithm is used for adjusting the weights of the network to minimize the mean squared error at the output for each F_0 value of a syllable (Yegnanarayana, 1999).

A separate model is developed for each of the speakers in three languages. The distribution of F_0 values for each of the speakers is not uniform. Majority of the F_0 values are around the mean of the distribution. If the neural network model is build with this kind of data, the model will be biased towards the mean of the distribution. This results in high prediction error at extreme values (lower F_0 values are overestimated and higher F_0 values are underestimated). This problem can be handled in two ways: using (1) pre-processing methods and (2) post-processing methods. In the pre-processing methodology, the training data is prepared to yield approximately uniform distribution using histogram equivalization. In the post-processing method, the predicted values are modified by imposing the piecewise linear transformation to overcome the biasing of the network due to the implicit distribution of training data.

From these two studies, it is observed that the overall accuracy of prediction is better by using pre-processing method compared to post-processing method (see Tables 2 and 9). From the philosophical point of view also, pre-processing 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 post-processing method, the bias in the prediction is corrected after the erroneous prediction due to the implicit biased distribution of the training data. The details of the intonation prediction using post-processing method are given at the end of this section.

In this paper, the intonation models are developed using pre-processing methodology, i.e., the training data is prepared to yield approximately uniform distribution using histogram equalization technique. Fig. 2 illustrates the nature of the original distribution of F_0 values and the modified distribution of F_0 values using histogram equalization. The data used in Fig. 2 corresponds to a male speaker in Hindi.

For each syllable a 25-dimensional input vector is formed, representing the positional, contextual and phonological features. The fundamental frequency of each syllable is obtained from the database. The number of epochs needed for training depends on the behavior of the training error. It was found that about 500 epochs are adequate for this study. The learning ability of the network from training data can be observed from the trend of the training error with # epochs. The training errors for neural network models for one speaker in

Table 2
Performance of the FFNN models for predicting the F_0 values of syllables for different languages

Language	Gender (# syllables)	% Predicted syllables within deviation					Objective measures		
		2%	5%	10%	15%	25%	μ (Hz)	σ (Hz)	γ
Hindi	F(660)	19	42	70	84	97	19.27	17.82	0.79
	M(1143)	21	45	76	93	98	11.72	9.42	0.81
Telugu	F(1276)	20	41	74	92	99	16.02	13.04	0.80
	M(984)	18	39	68	85	97	9.93	9.59	0.82
Tamil	F(741)	23	46	78	92	99	17.26	13.73	0.87
	M(1267)	24	43	79	91	97	12.41	11.54	0.81

The results for female (F) and male (M) speakers are given separately.

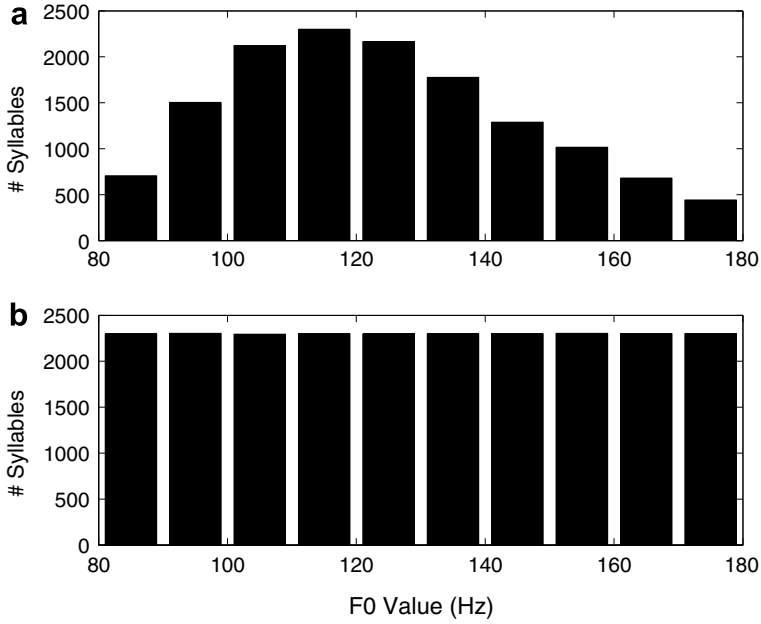

Fig. 2. Histogram representation of: (a) original distribution of F_0 values and (b) modified distribution of F_0 values using histogram equalization for a male speaker in Hindi.

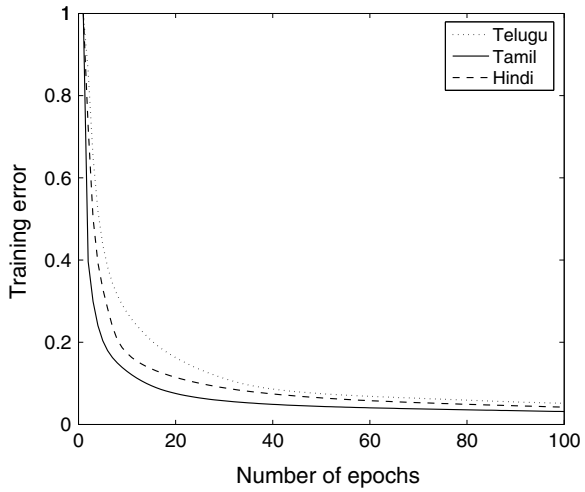

Fig. 3. Training errors for the neural network models for predicting the intonation patterns in three languages Hindi, Telugu and Tamil.

each of the three languages are shown in Fig. 3. The decreasing trend in the training error indicate that the network is capturing the implicit relation between the input and output.

The intonation model is evaluated using syllables in the test set. For each syllable in the test set, the F_0 value is predicted using the FFNN by presenting the feature vector of each syllable as input to the neural network. The deviation of the predicted F_0 from the actual F_0 is obtained. The prediction performance of the intonation model for a male speaker in Hindi is shown in Fig. 4.

The thick solid line in Fig. 4 represents the average predicted F_0 *vs.* the average F_0 of the syllables. These F_0 values are derived as follows: the range of the available F_0 values is uniformly divided into a finite number of bins. In this work, the range of F_0 for each bin is 10 Hz. The average F_0 corresponds to the mean of the F_0 values that fall into a particular bin. The average predicted F_0 corresponds to the mean of the predicted F_0

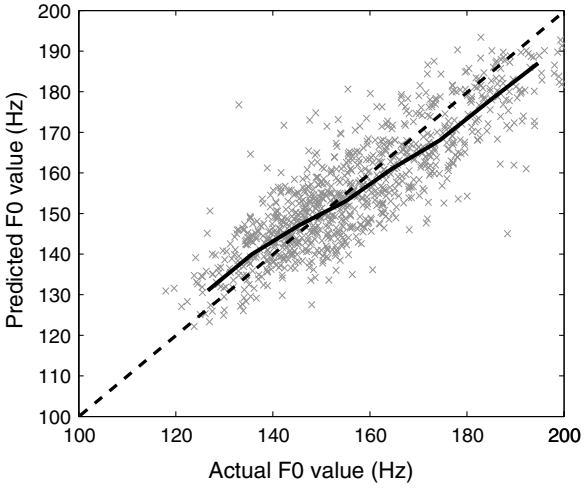

Fig. 4. Prediction performance of the neural network model for a male speaker in Hindi.

values for the syllables whose actual F_0 values fall into a particular bin. The average performance plot (thick solid line) in Fig. 4 is slightly deviated from the dashed line (diagonal), indicating that there is some prediction error at lower and higher F_0 values. For illustrating the actual prediction performance, the actual and the predicted F_0 values are jointly plotted (as 'x') on the same figure. The figure shows that the prediction is better in the range of 140–170 Hz, the same can be observed in the average performance plot (thick solid line) in Fig. 4.

The prediction performance of the intonation models is illustrated in Table 2 for one male (M) and one female (F) speaker for each language. The percentages of syllables predicted for different deviations from their actual F_0 values are given in Table 2. For each syllable the deviation (D_i) is computed as follows:

$$D_i = \frac{|x_i - y_i|}{x_i} \times 100,$$

where x_i and y_i are the actual and predicted F_0 values, respectively.

In order to evaluate the prediction accuracy, the average prediction error (μ), the standard deviation (σ), and the correlation coefficient $(\gamma_{X,Y})$ are computed using actual and predicted F_0 values. These results are given in Table 2.

The first column indicates different languages used in the analysis. The second column indicates the number of syllables specific to a speaker used in testing. Columns 3–7 indicate the percentage of syllables predicted within different deviations from their actual F_0 values. Columns 8–10 indicate the objective measures $(\mu, \sigma$ and $\gamma_{X,Y})$. The definitions of average prediction error (μ), standard deviation (σ) and linear correlation coefficient $(\gamma_{X,Y})$ are given below:

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

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

where x_i, y_i are the actual and predicted F_0 values, respectively, and e_i is the error between the actual and predicted F_0 values. The deviation in error is d_i, and N is the number of observed F_0 values of the syllables. The correlation coefficient is given by

$$\gamma_{X,Y} = \frac{V_{X,Y}}{\sigma_X \cdot \sigma_Y}, \quad \text{where } V_{X,Y} = \frac{\sum_i |(x_i - \bar{x})| \cdot |(y_i - \bar{y})|}{N}.$$

The quantities σ_X, σ_Y are the standard deviations for the actual and predicted F_0 values, respectively, and $V_{X,Y}$ is the correlation between the actual and predicted F_0 values.

The contribution of each component in the feature vector is analyzed by developing neural network models by omitting the particular component from the feature vector. For illustrating this, speech data of a male speaker in Hindi is considered. The results corresponding to the impact of each component in the feature vector are given in Table 3. The first column in the table indicates different components that constitute the feature vector. The percentage of syllables predicted within different deviations from their actual F_0 values are given in the columns 2–6. Columns 7–9 indicate the objective measures. From the analysis it is observed that the syllable position in the phrase (SPP) and the context of a syllable (CS) have major contribution in predicting the F_0 value of the syllable (see 1st and 4th rows of Table 3: in the 1st row the component SPP was omitted, and in the 4th row the component CS was omitted). The highest performance is achieved by combining all the components of the feature vector (see the last row of Table 3).

For illustrating the performance of the models in predicting the sequence of F_0 values of an utterance, both the predicted and the actual pitch contours of an utterance are plotted in Fig. 5. The utterance considered for the illustration is "*pradhAn mantri ne kahA ki niyantran rekhA se lekar*" in Hindi spoken by a male speaker. It shows that the predicted pitch contour is close to the original contour. This indicates that the neural network predicts the F_0 values reasonably well for the sequence of syllables in the given text.

The accuracy of prediction is also analyzed by perceptual tests. These tests are conducted on the speech utterances, which are synthesized by incorporating the intonation characteristics derived from the models. In this study speech utterances are synthesized using concatenative synthesis. Syllable is considered as a basic unit for synthesis. The database used for concatenative synthesis consists of syllables, which are extracted from

Table 3
Contribution of different components of feature vector for predicting the F_0 values of syllables

Components of a feature vector	% Predicted syllables within deviation					Objective measures		
	2%	5%	10%	15%	25%	μ (Hz)	σ (Hz)	γ
SPW + WPP + SI + CS + SN + P	17	38	66	85	90	16.87	14.82	0.73
SPP + WPP + SI + CS + SN + P	19	42	73	87	94	12.93	11.42	0.76
SPP + SPW+SI + CS + SN + P	20	44	74	92	96	12.57	12.04	0.77
SPP + SPW + WPP + SI+SN + P	16	36	64	80	88	17.93	15.59	0.72
SPP + SPW + WPP + SI + CS+P	20	46	75	90	96	12.26	10.73	0.78
SPP + SPW + WPP + SI + CS + SN	19	42	72	86	93	13.41	11.54	0.76
SPP + SPW + WPP + SI + CS + SN + P	21	45	76	93	98	11.72	9.42	0.81

The components of a feature vector are: SPP – syllable position in the phrase; SPW – syllable position in the word; WPP – word position in the phrase; SI – syllable identity; CS – context of the syllable; SN – syllable nucleus; P – pitch.

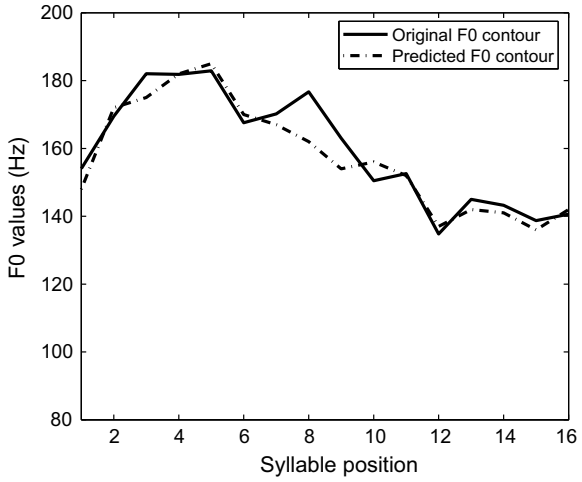

Fig. 5. Comparison of predicted F_0 contour with original contour for the utterance "*pradhAn mantri ne kahA ki niyantran rekhA se lekar*" in Hindi spoken by a male speaker.

a carrier words spoken in isolation. We have selected nonsense words as carrier words, because they offer minimum bias towards any of the factors that affect on the basic characteristics of the syllable. These syllables are considered as neutral syllables. Using some guidelines the carrier words can be formed for different categories of syllables. Some of the guidelines are prepared while developing the text-to-speech (TTS) system for Hindi at Indian Institute of Technology (IIT) Madras (Srikanth et al., 1989). For ensuring the syllables to be completely free from the implicit intonation, the pitch contour (intonation) of the syllables derived from the carrier words is made flat with respect to mean pitch value. With this the pitch contour of the utterance will become flat, and it is equivalent to the mean pitch of the speaker.

For demonstrating the impact of the accuracy of the intonation prediction, speech utterances are synthesized by incorporating the derived intonation characteristics on the sequence of neutral syllables. The proposed intonation models provide the sequence of F_0 values for the given sequence of syllables. The pitch contour (F_0 contour) for the whole utterance is derived by using interpolation and smoothing. The flat pitch contour of the sequence of neutral syllables is replaced by the predicted pitch contour using prosody modification methods. In this work, the intonation parameters are modified using the instants of significant excitation of the vocal tract system during the production of speech (Rao and Yegnanarayana, 2003). Instants of significant excitation are computed from the Linear Prediction (LP) residual of the speech signals by using the average group-delay of minimum phase signals (Smits and Yegnanarayana, 1995; Murthy and Yegnanarayana, 1999).

The quality (intelligibility and naturalness) of synthesized speech is evaluated by perceptual studies. Perceptual evaluation is performed by conducting subjective tests with 25 research scholars in the age group of 25–35. The subjects have sufficient speech knowledge for proper assessment of the speech signals. Five sentences are synthesized from text for each of the languages Hindi, Telugu and Tamil. Each of the subjects were given a pilot test about perception of speech signals with respect to intelligibility and naturalness. Once they are comfortable with judging, they were asked to take the tests. The tests were conducted in a laboratory environment by playing the speech signals through headphones. In the test, the subjects were asked to judge the intelligibility and naturalness of the speech. Subjects have to assess the quality on a 5-point scale for each of the sentences. The 5-point scale for representing the quality of speech is given in Table 4 (Deller et al., 1993).

The mean opinion scores (MOS) for assessing the intelligibility and naturalness of the synthesized speech in each of the languages Hindi, Telugu and Tamil are given in columns 2 and 3 of Table 5. The scores indicate that the intelligibility of the synthesized speech is fairly acceptable for all the languages, whereas the naturalness seems to be poor. Naturalness is mainly attributed to individual perception. Naturalness can be improved to some extent by incorporating the appropriate duration, coarticulation and stress information along with intonation.

For analyzing the accuracy of the prediction of intonation models, we also conducted the listening tests for assessing the intelligibility and naturalness on the speech without incorporating the intonation. In this case, speech samples are derived by concatenating the neutral syllables without incorporating the intonation. The mean opinion scores of these listening tests are given in columns 4 and 5 of Table 5. The MOS of the quality of the speech without incorporating the intonation have been observed to be low compared to the speech synthesized by incorporating the derived intonation characteristics from the proposed models. 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 is high ($> 99.5\%$) in all cases (shown in columns 6 and 7 of Table 5). This indicates that the differences in the pairs of MOS in each case is significant. From this study, we conclude that in view of perception, intonation characteristics will play a major role in synthesizing the

Table 4
Ranking used for judging the quality of the speech signal

Rating	Speech quality
1	Unsatisfactory
2	Poor
3	Fair
4	Good
5	Excellent

Table 5
Mean opinion scores for the quality of synthesized speech in the languages Hindi, Telugu and Tamil

Language	Mean opinion score (MOS)					
	Speech with intonation		Speech without intonation		Level of confidence (%)	
	Intl	Nat	Intl	Nat	Intl	Nat
Hindi	3.52	2.72	2.63	1.89	>99.5	>99.5
Telugu	3.95	3.12	3.12	2.11	>99.5	>99.5
Tamil	3.92	3.06	2.98	2.07	>99.5	>99.5

Intl: intelligibility; Nat: naturalness.

intelligible and natural speech. The neural network models proposed in this paper have shown to be successful in predicting the appropriate intonation characteristics suitable for speech synthesis applications.

For demonstrating the necessity of the nonlinear models for capturing the complex relations between the linguistic and production constraints of the sound units to their associated intonation patterns, a simple linear regression model is developed using the same features that are used for developing the FFNN model. The prediction performance of the linear regression model for male and female speakers of Hindi is given in Table 6. The performance of the linear regression models is observed to be very low compared to FFNN models. The performance of the FFNN models for the same data is shown within brackets in Table 6. The lower performance of the linear models can be attributed to their inability to capture the nonlinear (complex) relations present in the data.

The prediction performance of the neural network models can be compared with the results obtained by CART models. The CART models are developed using Festival TTS framework speech-tool *wagon* (Black et al., 2003). The data and the features used in developing the CART models are same as those used for developing neural network models. Here, we used the z-score of the F_0 values. Held-out data was used to choose the stop value and balance factor. We also experimented with different amounts of held-out data between 5% and 20%. The local maxima of the correlation, or minima of the mean absolute error on the held-out set were used to select the best stop value, balance factor and amount of held-out data. From the results, we observed that a stop value of 15–25, held-out data of 10–20% and a balance factor of about 10–15% are the optimal values across the models. We have also explored the CART models using the fixed stop value. Since, the models using the held-out data are performing better compared to the models using fixed stop value, we have considered the models developed using held-out data are used for the comparison with neural network models. The performance of the CART models is given in Table 7. The performance of FFNN models (shown within brackets in Table 7) seems to be better than for the CART models. The significance of the differences in the pairs of the mean prediction errors is tested using hypothesis testing (Hogg and Ledolter, 1987.). The level of confidence for the observed differences in the sample means was obtained from each model using the sample variances and values of Student's-t distribution. The level of confidence is high ($> 97.5\%$) in all cases. This indicates that the differences in the pairs of the mean prediction error for each model is significant.

In speech signal, the duration and intonation patterns of the sequence of sound units are interrelated at some higher level through emphasis (stress) and prominence of the words and phrases. But it is difficult to represent the feature vector to capture these dependencies. In this study, the intonation patterns (average pitch values) of the adjacent syllables are considered as intonation constraints, and the durations of the present syllable and its adjacent syllables are considered as duration constraints, for estimating the F_0 of the syllable. For

Table 6
Performance of the linear regression models for predicting the F_0 values of syllables in Hindi

Gender	% Predicted syllables within deviation					Objective measures		
	2%	5%	10%	15%	25%	μ (Hz)	σ (Hz)	γ
F	10(19)	22(42)	45(70)	58(84)	73(97)	31(19)	28(18)	.58(.79)
M	9(21)	27(45)	48(76)	61(93)	77(98)	23(12)	21(9)	.52(.81)

The numbers within brackets indicate the performance of neural network models.

Table 7
Performance of the CART models for predicting the F_0 values of syllables for different languages

Language (gender)	% Predicted syllables within deviation					Objective measures			Level of confidence (%)
	2%	5%	10%	15%	25%	μ (Hz)	σ (Hz)	γ	
Hi (F)	17(19)	36(42)	58(70)	73(84)	90(97)	21(19)	20(18)	.76(.79)	>99.5
Hi (M)	20(21)	39(45)	68(76)	90(93)	98(98)	14(12)	11(9)	.77(.81)	>99.5
Te (F)	19(20)	37(41)	68(74)	87(92)	99(99)	18(16)	13(13)	.74(.80)	>99.5
Te (M)	16(18)	34(39)	64(68)	83(85)	96(97)	11(10)	11(10)	.74(.82)	>97.5
Ta (F)	21(23)	40(46)	69(78)	90(92)	99(99)	21(17)	15(14)	.80(.87)	>99.5
Ta (M)	21(24)	42(43)	72(79)	88(91)	95(97)	15(12)	13(12)	.77(.81)	>99.5

The numbers within brackets indicate the performance of neural network models.

studying the influence of duration and intonation constraints in predicting the average F_0 of the syllables, separate models are developed with respect to different constraints. They are (a) model with linguistic features; (b) model with linguistic features and duration constraints; (c) model with linguistic features and intonation constraints and (d) model with linguistic features, duration and intonation constraints. The performance of these models is given in Table 8. The first column indicates the features used for developing the neural network models. The second column shows the gender of the speaker specific to the particular language used in testing. The columns 3–7 indicate the percentage of syllables having predicted F_0 values within the specified deviation with respect to their actual F_0 values. The columns 8–10 indicate the objective measures. The results indicate that the prediction performance has improved by imposing the constraints. From the table, it is observed that the percentage of syllables predicted within 2–15% is increased by including the features related to different constraints. A similar phenomenon is observed in objective measures also. Better performance is observed when all the constraints are applied together (last three rows of Table 8).

Modeling the intonation using original training data is biased towards the mean of the training data. Better prediction is observed around the mean of the training data. Syllables with high F_0 values tend to be underestimated, and low F_0 values to be overestimated. For improving the accuracy of prediction, the post-processing method suggests to modify the predicted values by imposing the piecewise linear transformation on the predicted F_0 values (Bellegarda et al., 2001; Bellegarda and Silverman, 1998; Silverman and Bellegarda, 1999). The piecewise linear transformation is defined by

$$F(x) = \begin{cases} \alpha x + a(1-\alpha), & 150 \leqslant x < a; \\ x, & a \leqslant x \leqslant b; \\ \beta x + b(1-\beta), & b \leqslant x < 350. \end{cases}$$

Table 8
Performance of the neural network models for prediction of the F_0 values of syllables using different constraints

Features	Language (gender)	% Predicted syllables within deviation					Objective measures		
		2%	5%	10%	15%	25%	μ (Hz)	σ (Hz)	γ
Linguistic	Hi(M)	16	48	65	82	98	14	12	0.69
	Te(F)	16	40	64	83	98	20	15	0.70
	Ta(F)	17	39	68	87	99	21	17	0.74
Linguistic and duration	Hi(M)	19	42	71	86	98	13	12	0.73
	Te(F)	20	44	73	89	99	17	15	0.73
	Ta(F)	21	42	73	90	98	19	17	0.77
Linguistic and intonation	Hi(M)	23	45	78	91	99	11	11	0.76
	Te(F)	22	44	79	92	99	16	14	0.79
	Ta(F)	20	47	78	93	99	18	14	0.81
Linguistic, duration and intonation	Hi(M)	26	51	82	94	99	9	9	0.80
	Te(F)	25	52	81	94	100	16	12	0.82
	Ta(F)	27	54	83	95	99	17	13	0.84

Hi: Hindi, Te: Telugu and Ta: Tamil.

Here, $F(x)$ denotes the transformed predicted value, x is the predicted value obtained from the models, and the parameters a, b, α, and β (all nonnegative) help to control the shape of the function. The interval $[a, b]$ defines the identity portion of the transformation, while α and β control the amount of compression/expansion in the intervals $[150, a]$ and $[b, 350]$, respectively. The bounds 150 and 350 represent the minimum and maximum values of F_0 observed for the news data read by a female speaker in Hindi. The values $\alpha, \beta < 1$ correspond to compression, and the values $\alpha, \beta > 1$ correspond to expansion. Fig. 6 shows the shape of the transformation function for two sets of values:

(1) $a = 200, b = 300, \alpha = 0.2, \beta = 0.7$ and
(2) $a = 200, b = 300, \alpha = 1.7, \beta = 1.8$.

The intonation models developed using original training data require the transformation to provide expansion at both the extremes (short and long duration syllables) for improving the accuracy of prediction. The limits a and b are derived from the distribution of the F_0 values. The optimum parameters of the transformation for the predicted values of a male speaker data in Hindi were found to be $a = \mu - 0.55 * \sigma$, $b = \mu + 0.65 * \sigma, \alpha = 1.2, \beta = 1.4$, where μ and σ are mean and standard deviation of the distribution of the F_0 values of the syllables, respectively. The F_0 values are recomputed from the predicted values using the transformation. Likewise the transformation is performed on the predicted F_0 values from other models. The performance measures derived from the transformed values and the actual F_0 values are given in Table 9. The numbers within brackets indicate the performance measures derived from the predicted (without transformation) and actual F_0 values. Here, the models used for predicting the intonation are developed using original training data. The prediction performance has improved slightly after imposing the transformation (see Table 9). The prediction accuracy using the post-processing method for correcting the bias in the prediction due to the inherent biased distribution of the training data is observed to be less effective compared to the prediction using pre-processing methodology (histogram equalization of training data). This can be observed from Tables 2 and 9. The entries in the columns 3–5 of Tables 2 and 9 indicate that there are more number of syllables predicted close to the actual F_0 value (2–10% deviation) in the case of prediction using pre-processing method compared to post-processing method.

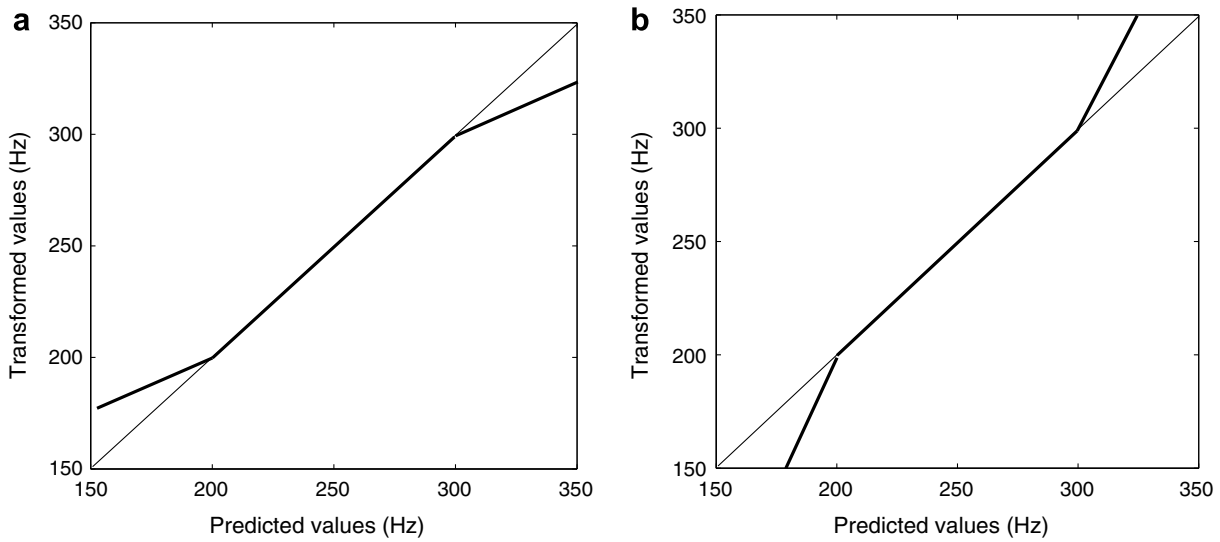

Fig. 6. Transformation with (a) compression at both the ends ($a = 200, b = 300, \alpha = 0.2, \beta = 0.7$), and (b) expansion at both the ends ($a = 200, b = 300, \alpha = 1.7, \beta = 1.8$).

Table 9
Performance measures after imposing the piecewise linear transformation

Language	Gender	% Predicted syllables within deviation					Objective measures		
		2%	5%	10%	15%	25%	μ (Hz)	σ (Hz)	γ
Hindi	F	12(13)	38(36)	66(67)	84(82)	99(96)	20(20)	18(19)	.79(.78)
	M	18(17)	40(38)	75(74)	93(92)	98(98)	12(13)	10(9)	.80(.79)
Telugu	F	17(16)	37(34)	73(72)	91(91)	98(99)	16(17)	13(13)	.78(.78)
	M	11(12)	34(33)	63(64)	83(82)	97(96)	10(10)	10(9)	.80(.79)
Tamil	F	17(18)	42(40)	76(77)	93(91)	99(99)	18(18)	14(14)	.84(.85)
	M	19(18)	40(39)	78(77)	92(90)	97(97)	13(14)	12(12)	.81(.80)

The numbers within brackets indicate the performance measures derived from the predicted (without transformation) and actual F_0 values. Here, intonation models are developed using original training data.

5. Summary and conclusions

Feedforward Neural Network (FFNN) models were proposed for predicting the intonation patterns of the syllables. Intonation patterns of the syllables are constrained by their linguistic context and production constraints. These constraints were represented with positional, contextual and phonological features, and used for developing the models. Suitable neural network structures were arrived at empirically. The models were evaluated by computing the average prediction error (μ), the standard deviation (σ) and the correlation coefficient (γ) between the predicted and actual F_0 values of the syllables. The accuracy in prediction was also analyzed in terms of percentage of syllables predicted within different deviations with respect to their actual F_0. The significance of the individual components of the feature vector was analyzed by developing the neural network models with different combinations of the components of the feature vector. From the analysis, it was observed that positional information of the syllable in the phrase and the information about the linguistic context were significant compared to other components. The performance of the neural network models was compared with the performance of the CART models. The prediction performance of the intonation models was also verified by conducting the perceptual tests on the synthesized speech utterances by incorporating the derived intonation from the models. The mean opinion scores of the perceptual tests indicated that the quality of speech synthesized with the predicted intonation was better compared to the speech synthesized without the intonation information. The accuracy of prediction was also analyzed using pre-processing and post-processing methods. The performance of the models using pre-processing method was observed to be better over the models using post-processing method. The performance of the intonation models can be further improved by including the accent and prominence of the syllable in the feature vector. Weighting the constituents of the input feature vectors based on linguistic and phonetic importance may further improve the performance. The accuracy of labeling, diversity of data in the database, and fine tuning of neural network parameters, all of these factors may also play a role in improving the prediction of the intonation patterns of the syllables. The unique properties of the intonation patterns with respect to the message content, speaker and language can be exploited in developing the robust speech recognition, speaker recognition and language identification systems respectively. The proposed intonation models may also be useful in applications such as expressive speech processing and voice conversion.

Appendix A. Coding scheme used to represent linguistic and production constraints

A syllable is a combination of segments of consonants (C) and vowels (V). Each segment of the syllable is encoded separately. Codes for representing the segments (consonants or vowels) of the syllables are given in Table A.1. In this study, neural networks are used to model the prosody parameters from linguistic features. For developing these models, fixed length feature vectors are needed. Therefore we represented syllable with four segments. In this study, syllables with more than four segments are ignored. Syllables with less than four segments are represented by dummy segments. These dummy segments indicate the absence of segments, and are uniquely coded by a specific code "55". The context of the syllable is represented by its adjacent (preceding

Table A.1
Codes for representing the segments (consonants or vowels) of the syllables

b	11	g	16	L	21	q	26	t	31	y	36	sh	41	ph	46	bh	51	u	62
c	12	h	17	m	22	r	27	T	32	z	37	Sh	42	gh	47	ai	58	e	63
d	13	j	18	n	23	R	28	v	33	ñ	38	kh	43	dh	48	au	59	o	64
D	14	k	19	N	24	s	29	w	34	$\widetilde{N}$	39	th	44	Dh	49	a	60	A	65
f	15	l	20	p	25	S	30	x	35	ch	40	Th	45	jh	50	i	61	I	66
U	67	E	68	O	69						Absence of a segment: 55								

and following) syllables. For the syllables at the word boundary (initial and final) have only one adjacent syllable is present as its context. In this case, the other syllable (missing) is represented by four dummy segments. These details can be observed in the following illustration.

A.1. Illustration of feature extraction from the sequence of syllables in the text

Text: "*pAkistAn ke pradhAn mantrI navAz sharIph*"
Table A.2 gives the following details:

(1) number of syllables in the given utterance;
(2) number of words in the utterance;
(3) position of the word in the utterance;
(4) position of the syllable in the utterance;
(5) position of the syllable in each word.

Table A.3 illustrates the codes for 25 features for the syllable/*pA*/ in the utterance "*pAkistAn ke pradhAn mantrI navAz sharIph*", which represent the positional, contextual and phonological information. They are as follows: 1, 3, 3, 1, 12, 12, 1, 6, 6, 55, 55, 55, 55, 19, 61, 29, 55, 25, 65, 55, 55, 1, 0, 2, 145.
Table A.4 illustrates the codes for 25 features for the syllables in the utterance "*pAkistAn ke pradhAn mantrI navAz sharIph*", which represent the positional, contextual and phonological information.

Table A.2
Syllable and word boundaries present in the utterance

	First word			Second word	Third word		Fourth word		Fifth word		Sixth word	
Syllables:	pA	kis	tAn	ke	pra	dhAn	man	trI	na	vAj	sha	rIph
Syllable position w.r.t. word	1	2	3	1	1	2	1	2	1	2	1	2
Syllable position w.r.t. phrase	1	2	3	4	5	6	7	8	9	10	11	12

Table A.3
Features for the syllable/*pA*/:

Linguistic and production constraints		Codes used for representing features			
Syllable position	Word level	1	3	3	
	Phrase level	1	12	12	
Word position	Phrase level	1	6	6	
Syllable context	Previous syllable	55	55	55	55
	Following syllable	19	61	29	55
Syllable identity		25	65	55	55
Syllable nucleus		1	0	2	
Pitch		145			

Table A.4
Features for the syllables present in the utterance *pAkistAn ke pradhAn mantrI navAj sharIph*

Syllable	Syllable position						Word position			Syllable context												Syllable nucleus			Pitch
	Word			Phrase						Previous syllable				Present syllable				Following syllable							
pA	1	3	3	1	12	12	1	6	6	55	55	55	55	25	65	55	55	19	61	29	55	1	0	2	145
kis	2	2	3	2	11	12	1	6	6	25	65	55	55	19	61	29	55	31	65	23	55	1	1	3	163
tAn	3	1	3	3	10	12	1	6	6	19	61	29	55	31	65	23	55	55	55	55	55	1	1	3	165
ke	1	1	1	4	9	12	2	5	6	55	55	55	55	19	63	55	55	55	55	55	55	1	0	2	168
pra	1	2	2	5	8	12	3	4	6	55	55	55	55	25	27	60	55	48	65	23	55	2	0	3	158
dhAn	2	1	2	6	7	12	3	4	6	25	27	60	55	48	65	23	55	55	55	55	55	1	1	3	152
man	1	2	2	7	6	12	4	3	6	55	55	55	55	22	60	23	55	31	27	66	55	1	1	3	150
trI	2	1	2	8	5	12	4	3	6	22	60	23	55	31	27	66	55	55	55	55	55	2	0	3	154
na	1	2	2	9	4	12	5	2	6	55	55	55	55	23	60	55	55	33	65	18	55	1	0	2	149
vAj	2	1	2	10	3	12	5	2	6	23	60	55	55	33	65	18	55	55	55	55	55	1	1	3	151
sha	1	2	2	11	2	12	6	1	6	55	55	55	55	41	60	55	55	27	66	46	55	1	0	2	146
rIph	2	1	2	12	1	12	6	1	6	41	60	55	55	27	66	46	55	55	55	55	55	1	1	3	143

References

Bellegarda, J.R., Silverman, K.E.A., December 1998. Improved duration modeling of English phonemes using a root sinusoidal transformation. In: Proceedings of the International Conference on Spoken Language Processing. pp. 21–24.

Bellegarda, J.R., Silverman, K.E.A., Lenzo, K., Anderson, V., 2001. Statistical prosodic modeling: from corpus design to parameter estimation. IEEE Transactions on Speech and Audio Processing 9, 52–66.

Benesty, J., Sondhi, M.M., Huang, Y. (Eds.), 2008. Springer Handbook on Speech Processing. Springer, New York.

Black, A.W., Taylor, P., Caley, R., King, S., 2003. Edinburgh speech tools library version 1.2.3. <http://www.cstr.ed.ac.uk/projects/speech-tools/>.

Breiman, L., Friedman, N., Olshen, R., 1984. Classification and Regression Trees. Wadsworth and Brooks, Pacific Grove, CA.

Buhmann, J., Vereecken, H., Fackrell, J., Martens, J.P., Coile, B.V., October 2000. Data driven intonation modeling of 6 languages. In: Proceedings of International Conference on Spoken Language Processing, Beijing, China. vol. 3. pp. 179–183.

Chopde, A., Itrans Indian Language Transliteration Package Version 5.2 Source. <http://www.aczone.con/itrans/>.

Cosi, P., Tesser, F., Gretter, R., September 2001. Festival speaks Italian. In: Proceedings of EUROSPEECH 2001, Aalborg, Denmark. pp. 509–512.

Deller, J.R., Proakis, J.G., Hansen, J.H.L., 1993. Discrete-Time Processing of Speech Signals. Macmillan Publishing Company, New York.

Dusterhoff, K.E., Black, A.W., Taylor, P.A., 1999. Using decision trees within the Tilt intonation model to predict F_0 contour. In: Proceeding of the Eurospeech, Budapest, Hungary.

Fujisaki, H., 1983. Dynamic characteristics of voice fundamental frequency in speech and singing. In: MacNeilage, P.F. (Ed.), The Production of Speech. Springer, New York, pp. 39–55.

Fujisaki, H., 1988. A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour. In: Fujimura, O. (Ed.), Vocal Physiology: Voice Production, Mechanisms and Functions. Raven Press, New York, pp. 347–355.

Fujisaki, H., Hirose, K., Takahashi, N., 1986. Acoustic characteristics and the underlying rules of the intonation of the common Japanese used by radio and TV announcers. In: Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing. pp. 2039–2042.

Goubanova, O., King, S., 2008. Bayesian networks for phone duration prediction. Speech Communication 50, 301–311.

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

Hogg, R.V., Ledolter, J., 1987. Engineering Statistics. Macmillan Publishing Company, New York.

Huang, X., Acero, A., Hon, H.W., 2001. Spoken Language Processing. Prentice-Hall, New Jersey.

Hwang, S.H., Chen, S.H., 1994. Neural-network-based F_0 text-to-speech synthesizer for Mandarin. IEE Proceedings of Image Signal Processing 141 (December), 384–390.

Khan, A.N., Gangashetty, S.V., Yegnanarayana, B., December 2003. Syllabic properties of three Indian languages: implications for speech recognition and language identification. In: International Conference on Natural Language Processing, Mysore, India. pp. 125–134.

Klatt, D.H., 1987. Review of text-to-speech conversion for English. Journal of Acoustic Society of America 82 (3), 737–793.

Krishna, N., Tulukdar, P., Bali, K., Ramakrishnan, A., 2004. Duration modeling for Hindi text-to-speech synthesis system. In: Proceedings of the International Conference on Spoken Language Processing, Denver, USA.

Kumar, S.R.R., March 1990. 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.

Kumar, A.S.M., January 1993. Intonation knowledge for speech systems for an Indian language. PhD thesis. Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai, India.

Kumar, A.S.M., Rajendran, S., Yegnanarayana, B., 1993. Intonation component of text-to-speech system for Hindi. Computer Speech and Language 7, 283–301.

Madhukumar, A.S., Rajendran, S., Sekhar, C.C., Yegnanarayana, B., 1991. Synthesizing intonation for speech in Hindi. In: Proceedings of the Second European Conference on Speech Communication and Technology, Genoa, Italy. vol. 3. pp. 1153–1156.

Murthy, P.S., Yegnanarayana, B., 1999. Robustness of group-delay-based method for extraction of significant excitation from speech signals. IEEE Transactions on Speech and Audio Processing 7 (November), 609–619.

Olive, J.P., 1975. Fundamental frequency rules for the synthesis of simple declarative English sentences. Journal of Acoustic Society of America (57), 476–482.

Pierrehumbert, J.B., 1980. The Phonology and Phonetics of English Intonation. PhD thesis. MIT, MA, USA.

Prasanna, S.R.M., Yegnanarayana, B., May 2004. Extraction of pitch in adverse conditions. In: Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing, Montreal, Canada.

Rao, K.S., Yegnanarayana, B., July 2003. Prosodic manipulation using instants of significant excitation. In: Proceedings of the IEEE International Conference on Multimedia and Expo, Baltimore, MD, USA. pp. 389–392.

Scordilis, M.S., Gowdy, J.N., May 1989. Neural network based generation of fundamental frequency contours. In: Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing, Glasgow, Scotland. vol. 1. pp. 219–222.

Silverman, K.E.A., Bellegarda, J.R., March 1999. Using a sigmoid transformation for improved modeling of phoneme duration. In: Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing, Phoenix, AZ, USA. pp. 385–388.

Smits, R., Yegnanarayana, B., 1995. Determination of instants of significant excitation in speech using group delay function. IEEE Transactions on Speech and Audio Processing 3 (September), 325–333.

Sonntag, G.P., Portele, T., Heuft, B., April 1997. Prosody generation with a neural network: weighing the importance of input parameters. In: Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing, Munich, Germany. pp. 931–934.

Sontag, E.D., 1992. Feedback stabilization using two hidden layer nets. IEEE Transactions on Neural Networks 3 (November), 981–990.

Srikanth, S., Kumar, S.R.R., Sundar, R., Yegnanarayana, B., March 1989. A text-to-speech conversion system for Indian languages based on waveform concatenation model. Technical report no. 11, Project VOIS, Department of Computer Science and Engineering, Indian Institute of Technology Madras.

Taylor, P.A., 2000. Analysis and synthesis of intonation using the Tilt model. Journal of Acoustic Society of America 107 (March), 1697–1714.

Tesser, F., Cosi, P., Drioli, C., Tisato, G., May 2004. Prosodic data driven modeling of a narrative style in festival TTS. In: Fifth ESCA Speech Synthesis Workshop, Pittsburgh, USA. pp. 185–190.

t'Hart, J., Collier, R., Cohen, A., 1990. A Perceptual Study of Intonation. Cambridge University Press, Cambridge.

Vainio, M., 2001. Artificial neural network based prosody models for Finnish text-to-speech synthesis. PhD thesis. Department of Phonetics, University of Helsinki, Finland.

Vainio, M., Altosaar, T., December 1998. Modeling the microprosody of pitch and loudness for speech synthesis with neural networks. In: Proceedings of the International Conference on Spoken Language Processing, Vainio, M. 2001. Artificial neural network based prosody models for Finnish text-to-speech synthesis. PhD thesis, Dept. of Phonetics, University of Helsinki, Finland.Sidney, Australia.

Vegnaduzzo, M., 2003. Modeling intonation for the Italian festival TTS using linear regression. Master's thesis. Department of Linguistics, University of Edinburgh.

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