Available online at www.sciencedirect.com

SciVerse ScienceDirect

ELSEVIER

Computer Speech and Language 27 (2013) 1105–1126

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Two-stage intonation modeling using feedforward neural networks for syllable based text-to-speech synthesis[☆]

V. Ramu Reddy [*], K. Sreenivasa Rao

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

Received 31 March 2012; received in revised form 8 February 2013; accepted 9 February 2013
Available online 21 February 2013

Abstract

This paper proposes a two-stage feedforward neural network (FFNN) based approach for modeling fundamental frequency (F_0) values of a sequence of syllables. In this study, (i) linguistic constraints represented by positional, contextual and phonological features, (ii) production constraints represented by articulatory features and (iii) linguistic relevance tilt parameters are proposed for predicting intonation patterns. In the first stage, tilt parameters are predicted using linguistic and production constraints. In the second stage, F_0 values of the syllables are predicted using the tilt parameters predicted from the first stage, and basic linguistic and production constraints. The prediction performance of the 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 two-stage FFNN model is compared with other statistical models such as Classification and Regression Tree (CART) and Linear Regression (LR) models. The prediction accuracy of the intonation models is also analyzed by conducting listening tests to evaluate the quality of synthesized speech obtained after incorporation of intonation models into the baseline system. From the evaluation, it is observed that prediction accuracy is better for two-stage FFNN models, compared to the other models.

Keywords: Intonation models; Prediction accuracy; Text-to-speech synthesis; Feedforward neural networks; Linguistic constraints; Production constraints; Positional; Contextual; Phonological; Articulatory; F_0 of syllable; Tilt

1. Introduction

Prosody plays an important role in improving the quality of text-to-speech synthesis (TTS) system both in terms of naturalness and intelligibility. Prosody refers to duration, intonation and intensity patterns of speech for the sequence of syllables, words and phrases. In this work, we focus on modeling one of the important prosodic parameters i.e., intonation. Intonation plays an important role in human speech communication. Intonation can be defined as the dynamics of fundamental frequency (F_0) contour over time, caused due to vocal folds vibration. The perceptual correlate of F_0 is pitch. Human hearing system is highly sensitive to variations in pitch (Moore, 1989). In speech synthesis, intonation directly affects the overall quality of the synthetic speech. From the speaker's view point, intonation can be used to convey pragmatic and emotional information. Different intonation patterns of syntactically similar sentences can convey dramatically distinct information. Intonation patterns are influenced by various factors, while analyzing

[☆] This paper has been recommended for acceptance by B. Moebius.

[*] Corresponding author. Tel.: +91 7872381435.
E-mail addresses: ramu.csc@gmail.com (V.R. Reddy), ksrao@iitkgp.ac.in (K.S. Rao).

0885-2308/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.csl.2013.02.003

the speech at different levels (O'Shaughnessy, 1984). At the lowest level, F_0 (micro-intonation) is affected by local segmental factors which are caused by dynamics of human speech production process. At a higher level, stress patterns, rhythm and melody affect the F_0 contour. The F_0 contour is also affected by attitude, gender, physical and emotional state of the speaker. From the listener's view point, intonation plays an important role in (i) resolving syntactic ambiguity, (ii) segmentation of utterances, (iii) speech perception during noisy environments, and (iv) perceiving the emotional state of the speaker (O'Shaughnessy, 1987). In addition to the above functions, pitch contours also carry the lexical meaning in tonal languages, such as Mandarin Chinese. The speech synthesized without intonation appears to be highly monotonous and robotic, and is not pleasant for listening over longer durations. Hence, while developing speech synthesis systems, acquisition and incorporation of the intonation knowledge is very much essential.

The implicit knowledge of intonation is usually captured by using modeling techniques. In this work, we propose a two-stage intonation model using feedforward neural networks (FFNN) for predicting the intonation patterns of the sequence of syllables. Neural networks are known for their ability to capture the underlying interactions that exist between input and output features (Haykin, 1999; Yegnanarayana, 1999). Neural networks also have the generalization ability to predict intonation patterns reasonably well for the patterns which are not present in the learning phase (Haykin, 1999). In speech signal, the intonation of each unit is dictated by the linguistic and production constraints of the unit (Kumar, 1993; Rao and Yegnanarayana, 2009). In this study, linguistic and production constraints are used to predict the F_0 values of the sequence of syllables. Linguistic constraints are represented by positional, contextual and phonological features, and production constraints are represented by articulatory features. In addition to the above features, tilt parameters are also used to capture the true shape of the intonation patterns. The prediction accuracy of the intonation models is further improved by imposing the other prosodic constraints represented by duration and intensity values of the syllables.

The paper is organized as follows: Section 2 presents an overview of the existing research on acquisition of intonation knowledge using different models. The speech database used for the development of baseline TTS system, and the performance of the intonation models used in the baseline TTS system are discussed in Section 3. Section 4 describes the proposed features used for predicting the F_0 values. Performance of the neural network models along with that of Linear Regression (LR) and Classification and Regression Tree (CART) models is given in Section 5. The performance of the proposed two-stage intonation model using tilt parameters is discussed in Section 6. Section 7 presents the prediction accuracy of the intonation models using individual features. The influence of the other prosodic constraints for predicting the intonation patterns is discussed in Section 8. Summary of this paper is presented in Section 9.

2. Literature review

Many methods have been developed for generation of F_0 contours to build successful TTS systems. In the last 20 years, two major approaches have emerged for modeling intonation: (i) the tone sequence approach which follows the traditional phonological description of intonation and (ii) the superposition approach (Botinis et al., 2001).

Phonological (tone sequence) models interpret F_0 contour as a linear sequence of phonologically distinctive units (tones or pitch accents), which are local in nature. There is no interaction of events in the F_0 contour with each other. Popular phonological models include Pierreihumbert's model and TOBI (Tone and Break Indices). Tone sequence intonation model was initially developed by Pierrehumbert (1980) for American English. The original model was extended by Beckman and Pierrehumbert (1986), and it evolved as Tone and Break Indices (ToBI) for transcribing intonation of American English (Silverman et al., 1992). Tone sequence models have been implemented for English, German, Chinese, Navajo and Japanese (Sproat, 1998; Jilka et al., 1999). Phonological models do not properly represent actual pitch variations. No distinction is made on the differences in tempo or acceleration of pitch movements. The temporal information is also not modeled. Phonological models are not easily ported from one language to another, since the inventory of categories must be thoroughly reviewed by linguistic experts (Buhmann et al., 2000).

Due to availability of large speech corpora, more acoustic-phonetic models have been proposed in recent years for text-to-speech systems. Acoustic-phonetic models are developed using acoustic data. Intonation model for Danish language was developed by Gronnum which is conceptually quite different from the tone sequence model (Gronnum, 1992, 1995). The model is hierarchically organized and includes several simultaneous, non-categorical components of different temporal scopes. The components are *layered*, i.e., a component of short temporal scope is superimposed on a component of a longer scope. Gårding (1983) also developed an intonation model which analyzes the intonation contour of an utterance as the result of the effects of several factors. Acoustic-phonetic (superposition or overlay) models interpret F_0 contour as a result of superposition of several components of different temporal scopes. A classical *superpositional* intonational model for Japanese has been proposed by Fujisaki and his colleagues (Fujisaki et al., 1971; Fujisaki, 1983, 1988). Fujisaki model is also widely applied for the languages such as German, English, Greek, Polish, Spanish and French (Fujisaki et al., 1986, 1994, 1997; Fujisaki and Ohno, 1995; Mixdorff and Fujisaki, 1994).

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 which are not perceived by the listener all the time. The IPO model developed at the Institute of Perception Research at Eindhoven is probably the best-known perception model of intonation t'Hart et al. (1990) originally developed for Dutch. Later it was applied to English, German and Russian (Terken, 1993; de Pijper, 1983; Adriaens, 1991; Ode, 1989). Finally, *acoustic stylization* models aim at efficient analysis and synthesis using *rise/fall/connection* (RFC) models (Taylor, 1995, 2000). 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 (Taylor, 2000, 1995). The abstraction can then be used to produce a close copy of the original contour.

In literature, several models using neural networks are described for predicting the intonation patterns of syllables in continuous speech (Scordilis and Gowdy, 1989; Traber, 1992; Hwang and Chen, 1994; Sonntag et al., 1997; Vainio and Altosaar, 1998; Vainio, 2001; Buhmann et al., 2000). In Scordilis and Gowdy (1989), Scordilis and Gowdy have used neural networks in parallel and distributed manner to predict the average F_0 value for each phoneme, and also the temporal variations of F_0 within a phoneme. In Traber (1992), Traber used a neural network with two hidden layers to model eight target F_0 points for each syllable in German. Hwang and Chen (1994) have used neural networks for F_0 prediction for Mandarin TTS system. Two multilayer perceptrons are used separately to synthesize the mean and shape of F_0 contour using linguistic features extracted from the text. Sontagg et al. have used neural networks as an alternative to synthesize-by-rule for prosody generation. Intonational, durational, positional and perceptual parameters were used as input to the neural networks, and the contribution of each parameter was analysed. A three layer FFNN was used for predicting the F_0 values for a sequence of phonemes in Finnish language (Vainio and Altosaar, 1998; Vainio, 2001). The features used for developing the models were: phoneme class, length of a phoneme, identity of a phoneme, identities of previous and following syllables (context), length of a word and place of a word. Buhmann et al. (2000) have used a recurrent neural network (RNN) for developing multi-lingual (six languages) intonation models. In their work, an RNN model of the Elman type was trained with backpropagation-through-time algorithm. The output targets of the network were five equidistant F_0 points selected from each syllable. These points were transformed into F_0 contours by interpolation. The features used in their work are the universal (language independent) linguistic features such as part-of-speech and type of punctuation, combined with prosody features such as word boundary, prominence of the word and duration of the phoneme.

In Odéjobí et al. (2006), Relational Tree technique was used for modeling the intonation patterns for the language Yorùbá. Using tone phonological rules, Skeletal Tree (S-Tree) was generated and then using a fuzzy logic based model, numerical values of the perceptually significant peaks and valleys on the S-Tree were computed. Interpolation technique was applied to join the resulting points and PSOLA technique was used to synthesize the actual intonation contour. Gu et al. (2006) have analyzed the intonation patterns for emphasis and question related utterances of Cantonese language. In Campillo and Banga (2011), multiple intonation contours using parallel Viterbi search are proposed to improve the quality of unit selection based TTS system. Recently, efficient intonation models have been developed for languages like Chinese (Peng et al., 2008), Romanian (Bodo et al., 2009), Turkish (Uslu and Ilk, 2009), Croatian (Nacinovic et al., 2010), Azeri (Damadi et al., 2010) and Macedonian (Gerazov and Ivanovski, 2012) for building TTS systems.

In the Indian context, a rule-based intonation model was proposed by Kumar et al. (1993) and Kumar (1993) for Hindi TTS system. The analysis of intonational phenomena was carried out on read speech. A corpus of 500 sentences was used for deriving the rules. Rao and Yegnanarayana have developed feedforward neural network and support vector machine models for capturing the intonation and duration patterns of speech in Indian languages (Rao and Yegnanarayana, 2009, 2007). The intonation models were developed by using broadcast news data from Hindi, Telugu and Tamil. Linguistic constraints represented by positional, contextual and phonological features were used to capture the intonation patterns. A four layer feedforward neural network and SVMs were used to map the linguistic constraints to intonation patterns of the sequence of syllables. In Reddy and Rao (2011), positional, contextual, phonological and prosodic features are used for predicting the F_0 values of the syllables for developing syllable based TTS systems.

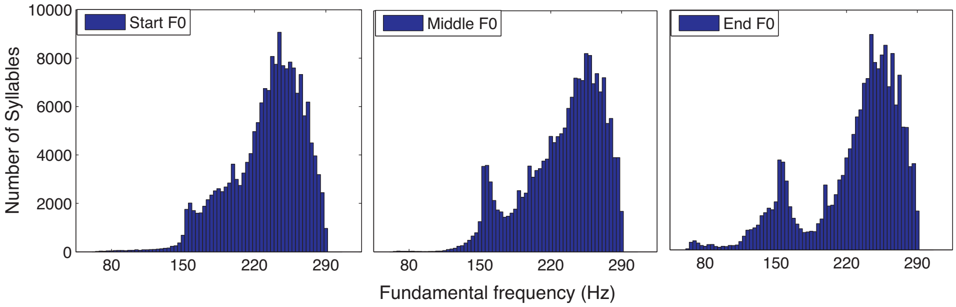

Fig. 1. Distribution plots of F_0 values at start, middle and end positions of syllables.

3. Development of syllable based Bengali TTS system

3.1. Speech database

Text utterances of the speech database used for this study are mainly collected from Anandabazar Patrika – a Bengali newspaper. It consists of news from several domains such as politics, entertainment, sports and stories. The other sources include story and text books in various fields such as history, geography, travelogue, drama and science. The text corpus covers 7762 declarative sentences derived from 50,000 sentences through optimal text selection method. The corpus covers 4372 unique syllables and 22382 unique words. The optimal text is recorded with a professional female artist in a noiseless chamber. The total duration of recorded speech is around 10 h. The speech signal was sampled at 16 kHz and represented as 16 bit numbers. The speech utterances are segmented and labeled into syllable-like units using ergodic hidden Markov models (EHMM). For every utterance a label file is maintained, which consists of syllables of the utterance and their timing information. The syllable structures considered here are V, CV, CCV, CVCC and CCVC, where C is a consonant and V is a vowel. For developing intonation models, the fundamental frequencies (F_0) of the syllables should be available in the database. The fundamental frequencies of the syllables in the database are computed by using zero frequency filter method (Murty and Yegnanarayana, 2008).

From the database, it is observed that the F_0 values vary from 80 Hz to 290 Hz. The average and standard deviation of F_0 of female speaker in the database was found to be 223 Hz and 47 Hz, respectively. The distribution plots of F_0 at start, middle and end positions of syllables is shown in Fig. 1. From the figure it is observed that most of the F_0 values of syllables are concentrated between 130 Hz and 270 Hz, respectively.

3.2. Baseline TTS system

The baseline TTS system (Narendra et al., 2011) is developed using recorded speech corpus described in the above subsection. Festival framework is used for developing TTS system (Black and Lanzo, 2009). Festival offers general tools for building unit selection synthesizer. Festival offers general architecture for easy experimentation and evaluation of different algorithms and modules (Black et al., 2009). Major issues considered in developing TTS are text corpus collection, recording and labeling the speech corpus, deriving letter to sound rules and prosody modeling. The sequence of steps followed for text-to-speech conversion is shown in Fig. 2.

Letter to sound (LTS) rules are used to derive the pronunciation for the given sequence of words. LTS rules are developed in two ways: (1) deriving the pronunciation dictionary and (2) deriving rules using linguistic knowledge. Pronunciation dictionary is a look-up table where for each unique word, there exists a corresponding sequence of phones. When a given word is not present in the pronunciation dictionary, rules are applied based on linguistic knowledge which includes orthographic rules, schwa deletion rules, and some specific rules for special cases (Ghosh et al., 2010). At the time of synthesis, the input text is divided into a sequence of sound units. For each of the sound units, target specifications are denoted by linguistic and prosodic features. Depending on the target specification, unit selection module selects an optimal sequence of units from the speech corpus by minimizing two costs, namely, concatenation cost and target cost. The details of the TTS system are given in Narendra et al. (2011).

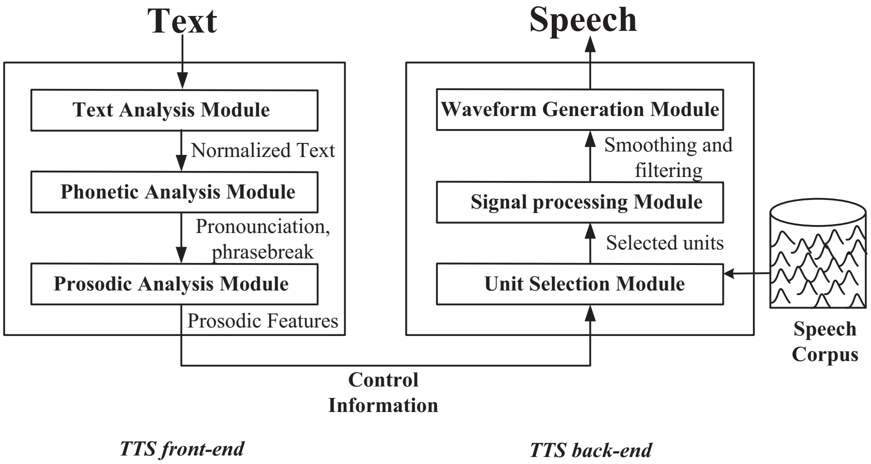

Fig. 2. Block diagram of text-to-speech system.

Table 1
Features used by Festival for modeling intonation patterns.

Features used in Festival for modeling F_0		
Start F_0	Middle F_0	End F_0
syl_in	stress	ssyl_out
stress	n.stress	stress
n.stress	syl_out	n.stress
pos_in_word	syl_in	syl_break
p.stress	ssyl_in	syl_in
nn.stress		n.syl_break
syl_out		p.syl_break
ssyl_out		
pp.stress		

3.3. Intonation modeling in Festival

Festival provides different rule-based and trained intonation modules for predicting the target F_0 values. In developing baseline TTS system, trained models like Classification and Regression Trees (CART) are used for predicting the F_0 values of syllables. The features used by Festival for modeling the intonation patterns are given in Table 1.

The prediction performance of the CART model based on the features used by Festival (shown in Table 1) is given in Table 2. Here, the prediction accuracy of the model is computed using objective measures such as average prediction error (μ), standard deviation (σ) and linear correlation coefficient ($\gamma_{X,Y}$). The number of predicted syllables within different deviations is also shown in Table 2. From Table 2 it is observed that the average prediction error is large. This is mainly due to poor prediction of F_0 values by CART model, and this in turn depends upon the features

Table 2
Performance of CART based intonation model using Festival features.

F_0 position in the syllable	% predicted syllables within the deviation					Objective measures		
	2%	5%	10%	15%	25%	μ (Hz)	σ (Hz)	$\gamma_{X,Y}$
Start	6.31	14.96	29.87	43.29	62.10	54.21	47.17	0.65
Middle	8.65	17.67	34.13	47.10	67.32	44.47	38.13	0.67
End	6.02	10.93	25.59	41.16	58.35	56.82	48.19	0.64

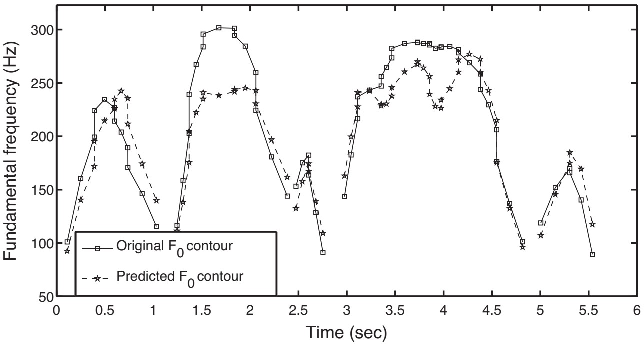

Fig. 3. Comparison of original and predicted F_0 contours with respect to time for the Bengali utterance *"bArAmotI eArsTriper pAsh die kud∼i pon∼chish kimi gele gaond"*.

used to model the intonation patterns of syllables. The features used by Festival to model the intonation patterns are more stress related features. It ignores the features related to positional and phonological information, and also the features related to production aspects of sound units. The stress related features used by the Festival intonation model are more suitable for stress-timed languages such as English, German, Danish, Swedish, Norwegian, Faroese, Dutch and Portuguese. However, Bengali is syllable-timed language having equal stress on each syllable. Therefore, more deviation is observed between predicted and actual F_0 values. The performance of CART model is also analyzed for an utterance *"bArAmotI eArsTriper pAsh die kud∼i pon∼chish kimi gele gaond"*, by plotting the actual and predicted F_0 values for the sequence of syllables of the utterance. The actual and predicted F_0 contours for the utterance mentioned above is shown in Fig. 3. From Fig. 3, it is observed that F_0 contour of an utterance consists of sequence of rise and fall patterns. Fig. 4 shows the actual and predicted F_0 contours correspond to start, middle and end F_0 values of the syllables. From Fig. 4, it is observed that there exists large deviation between actual and predicted F_0 values of the sequence of syllables. Therefore, there is a need for appropriate model with the features which are more relevant for syllable based TTS systems. Hence, in this work for predicting the intonation patterns, we have proposed neural network models with syllable-specific features.

4. Features for modeling intonation patterns

Intonation knowledge is not explicitly taught or learned when we learn to speak. Therefore, it is difficult to state the rules governing the intonation patterns of an utterance. Extracting the implicit rules related to intonation patterns of the sound units present in the speech signal is a difficult task. In a speech signal, the intonation pattern (F_0 contour) corresponding to a sequence of sound units is constrained by the linguistic and production constraints of the units (Kumar, 1993; Rao, 2005). The linguistic and production constraints of syllables can be expressed using positional, contextual, phonological and articulatory (PCPA) features. In this study we use a 35 dimensional feature vector to represent the linguistic and production constraints of each syllable. Out of 35 features, the first 24 features represent the linguistic constraints in the form of positional, contextual and phonological information and the remaining 11 features represent articulatory information of the 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.

However, it was observed that the dynamics of F_0 follows rise-fall patterns. The rise-fall pattern of F_0 is mainly due to the linguistic structure of the utterances. The prediction performance may be further improved by exploring features related to the shape of the intonation event in addition to linguistic and production constraints. As we have seen in Section 2, tilt model has the ability to capture both phonological and phonetic aspects of intonation, and it derives the linguistic representation automatically from the utterance's acoustics (Taylor, 1995). Hence, for modeling the intonation, we use features representing linguistic constraints, production constraints and dynamic tilt parameters of the sound units. The details of the proposed features for predicting the F_0 values of the sequence of syllables are discussed in the following subsections.

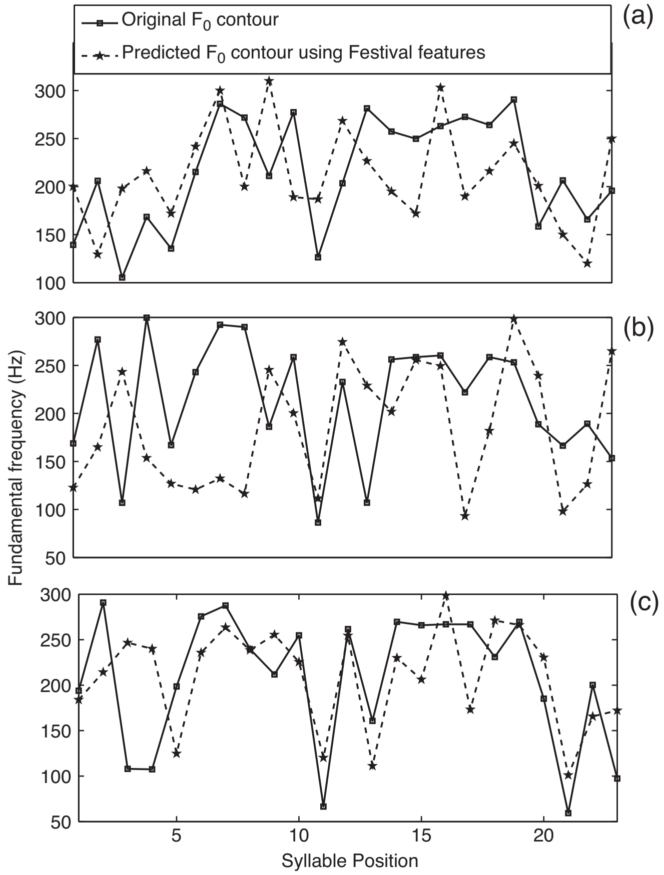

Fig. 4. Comparison of predicted (a) start F_0, (b) middle F_0, and (c) end F_0 patterns from CART models using Festival features with original F_0 patterns of the Bengali utterance *"bArAmotI eArsTriper pAsh die kud~i pon~chish kimi gele gaond"*.

4.1. Linguistic constraints

The linguistic constraints used in this work are represented by the following features:

Syllable position in the sentence: The position of the syllable in a sentence is represented by three features. The first feature represents distance of the syllable from the starting position of the sentence. It is measured in number of syllables which are ahead of the present syllable. The second feature indicates distance of the syllable from the end of the sentence. The third feature represents the total number of syllables present in the sentence.

Syllable position in the word: Words are separated by delimiter space. The syllable position in a word is characterized by three features, similar to syllable position in the sentence. Here, syllable position in a word from the starting and ending positions are considered as two features. The third feature indicates the total number of syllables in the word.

Word position in the sentence: The position of a word in a sentence is represented by three features. The first feature represents distance of the word from starting position of the sentence. It is measured in number of words which are ahead of the present word. The second feature indicates the distance of the word from the end of the sentence. The third feature represents total number of words present in the sentence.

Syllable identity: Each syllable consists of a combination of consonants (C) and a vowel (V) representing phonological information. In this work, syllable is represented by four segments, where each segment represents either vowel or consonant or empty. The segments are then encoded, so that each syllable is represented by four features indicating its identity. Each of the C and V segments are uniquely coded based on their identity.

Context of a syllable: Contextual information is represented by the previous syllable and the following syllable. Each of these syllables is represented by a four dimensional feature vector, representing the identity of the syllable.

Table 3
List of factors affecting intonation patterns of syllables, features representing the factors and number of nodes needed for neural network to represent the features.

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	4
	Following syllable	4
Syllable nucleus	Number of segments before the nucleus Number of segments after the nucleus Number of segments in a syllable	3

Syllable nucleus: In a syllable, vowel is treated as a nucleus. Within each syllable, the number of segments before and after the vowel and total number of segments in a syllable are also important. This is represented by three independent codes specifying three distinct features.

The detailed list of features representing linguistic constraints and the number of input nodes needed for the neural network to represent these features is given in Table 3. These features are coded and normalized before presenting to the neural network.

4.2. Production constraints

The intonation patterns of speech segments are also influenced by the production mechanism of speech sounds in addition to the linguistic constraints. Each sound unit has specific articulatory movements and positions during its production. These production constraints in turn depend on the language. Phonetics deals with the production of speech sounds by humans, often without prior knowledge of the language being spoken. In studying articulation, phoneticians explain how humans produce speech sounds with the interaction of different physiological structures. Therefore, in this study, the features related to different articulatory positions and manners of speech segments (consonants and vowels) are considered as production constraints. These production constraints are represented as articulatory features to predict the intonation patterns of the syllables. The manner of articulation describes the involvement of speech organs such as tongue, lips, jaw in producing a sound. The place of articulation of a consonant is the point of contact where an obstruction occurs in the vocal tract between an active (moving) articulator (typically some part of the tongue) and a passive (stationary) articulator (typically some part of the roof of the mouth). Place of articulation gives the consonant its distinctive sound along with the manner of articulation. For any place of articulation, there may be several manners, and therefore several homorganic consonants.

The schematic representation of different places of articulations is shown in Fig. 5. The quality of the vowel depends on the articulatory features that distinguish different vowel sounds (Association, 1999). 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) (Association, 1999). Height relates to the position of tongue and the degree of lowering of the jaw. Backness relates to the position of the body of the tongue in oral cavity. Rounding refers to the position of the lips. In this study, a 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 roundedness (lip rounding), consonant type, consonant place, consonant voicing, aspiration, nukta

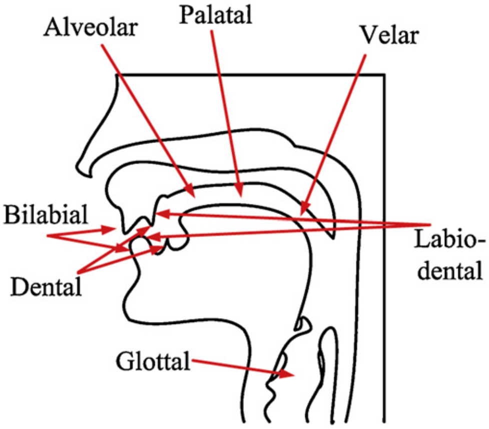

Fig. 5. Schematic view of places of articulation.

(diacritic mark), type of first phone and type of last phone in a syllable. The detailed list of production constraints represented in the form of articulatory features is given in Table 4. Each articulatory feature is uniquely coded.

4.3. Tilt features

The basic unit in the tilt model is called the *intonation event* (Taylor, 2000, 1995). Each intonation event, be an *accent*, *boundary, silence* or *connection* between events, is described by a set of continuous parameters, which are useful for prosodic control in speech synthesis. The intonation events are parameterized by measuring the amplitudes and durations of the rises and falls, denoted by A_{rise}, A_{fall}, D_{rise}, and D_{fall} as shown in Fig. 6. These four parameters are converted into three tilt parameters, namely *tilt*, A_{event} and D_{event}.

The *tilt* is calculated as follows:

$$tilt_{amp} = \frac{|A_{rise}| - |A_{fall}|}{|A_{rise}| + |A_{fall}|} \tag{1}$$

$$tilt_{dur} = \frac{|D_{rise}| - |D_{fall}|}{|D_{rise}| + |D_{fall}|} \tag{2}$$

$$tilt = \frac{1}{2} tilt_{amp} + \frac{1}{2} tilt_{dur} \tag{3}$$

Table 4
List of articulatory features.

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

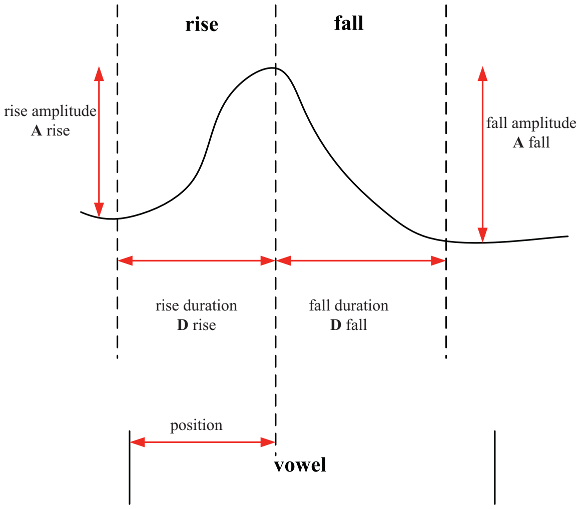

Fig. 6. The tilt model and its parameters.

$$= \frac{|A_{rise}| - |A_{fall}|}{2(|A_{rise}| + |A_{fall}|)} + \frac{|D_{rise}| - |D_{fall}|}{2(|D_{rise}| + |D_{fall}|)} \tag{4}$$

$$A_{event} = |A_{rise}| + |A_{fall}| \tag{5}$$

$$D_{event} = |D_{rise}| + |D_{fall}| \tag{6}$$

In this work, parameter position is also used in addition to above mentioned three tilt parameters. The description of the tilt parameters used in this study are as follows:

(1) Amplitude event (A_{event}): the amplitude of an intonation event relative to starting F_0 (in Hz).
(2) Duration event (D_{event}): the duration of an intonation event (in ms).
(3) *tilt*: a dimensionless parameter in the range of [−1, 1] describing the shape of an intonation event. −1 represents a pure fall, 1 is pure rise, and 0 indicates that the event contains equal portions of rise and fall.
(4) *Position*: the peak location of an intonation event, which is usually defined as the distance between the vowel starting time and the peak location.

The parameter *tilt* gives the actual shape of the intonation event. Therefore, tilt parameters have the ability to capture the linguistic relevance information and provide more naturalness to the synthesized speech. Hence, in this study in addition to positional, contextual, phonological and articulatory features we also use the tilt parameters. In this work, each syllable is treated as an intonation event and the tilt parameters are extracted from each syllable.

5. Intonation modeling using single-stage feedforward neural networks

In this work, a four layer feedforward neural network (FFNN) (Tamura and Tateishi, 1997; Rao, 2005) with an input layer, two hidden layers and an output layer is used for modeling the intonation patterns of syllables. The structure of the FFNN for predicting the F_0 values of syllables using linguistic and production constraints as features is shown in Fig. 7. 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. 7) of the neural network consists of more units compared to the input layer (first layer in Fig. 7), so that the network can capture some local variations of features in the input space. The second hidden layer (third layer in Fig. 7) of the neural network has fewer units compared to the input layer, so that the network can capture global variations of features in the input space (Haykin, 1999; Yegnanarayana, 1999). The last layer (fourth layer in Fig. 7) is the output layer having three linear units representing the average F_0 values of syllable. 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

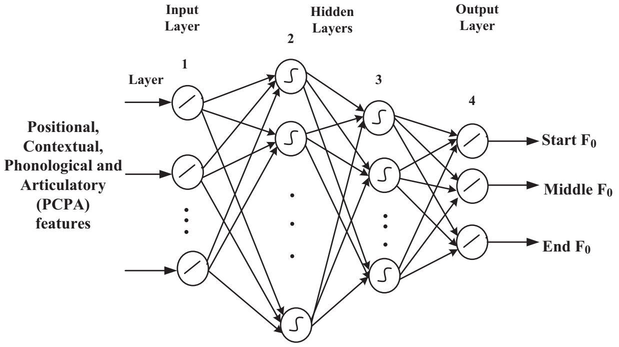
Input Layer
Hidden Layers
Output Layer
Layer
1
2
3
4
Positional, Contextual, Phonological and Articulatory (PCPA) features
Start F0
Middle F0
End F0

Fig. 7. Architecture of four layer feedforward neural network for predicting the F_0 values of syllables.

35 dimensional input vectors representing positional, contextual, phonological and articulatory features are presented as input, and the average F_0 values corresponding to start, middle and end positions of syllables are presented as desired outputs to the FFNN model. Here, three F_0 values for each syllable are chosen to capture the broad shape of the intonation contour of syllables.

The generalization by the network is influenced by three major factors: (1) the architecture of the network, (2) the amount of data used in training the network, and (3) the complexity of the problem. We have no control over the second and third factors. Different network structures were explored in this study to obtain the optimal performance, by incrementally varying the number of hidden layer neurons between 5 and 100. 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 indicating the number of units used in different layers. The (*empirically arrived*) final optimal structure 35*L* 72*N* 19*N* 3*L* is obtained with minimum generalization error. The activation function used in the non-linear unit (*N*) is *tanh(s)* function, where '*s*' is activation value of that unit. The input and output features are normalized between [−1, 1], before presenting to the neural network.

The training process of FFNN is carried out using Levenberg–Marquardt backpropagation algorithm to adjust the weights of the neural network, by backpropagating the mean-squared error to the neural units and optimizing the free parameters (synaptic weights) to minimize the error. Gradient descent with momentum weight function is used as an adaptation learning function. The backpropagation network learns by examples. So, we use input-output examples to show the network what type of behavior is expected, and the backpropagation algorithm allows the network to adapt. The backpropagation learning process works in small iterative steps as follows:

(i) One of the example cases is presented to the network.
(ii) The network produces an output based on the current state of its synaptic weights (initially, the output will be random).
(iii) The network output is then compared to the desired output and a mean-squared error signal is calculated.
(iv) The error value is then propagated backwards through the network, and weights are updated to decrease the error in each layer.
(v) The whole process is repeated for each of the examples.

For each syllable a 35 dimensional feature vector is formed, representing the positional, contextual, phonological and articulatory information. In this work, data consisting of 177,820 syllables is used for modeling the intonation. The data is divided into two parts namely design data and test data. The design data is used to determine the network topology. The design data in turn is divided into two parts namely training data and validation data. Training data is used to estimate the weights (includes biases) of the neural network and validation data is used to minimize the over-fitting of network, to verify the performance error and to stop training once the non-training validation error estimate stops decreasing. The test data is used only once on the best design, to obtain an unbiased estimate of the predicted error for unseen non-training data. The percentage 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 which is a subset of the one used for parameter estimation. As generalization is the goal of the neural network, we have used cross validation.

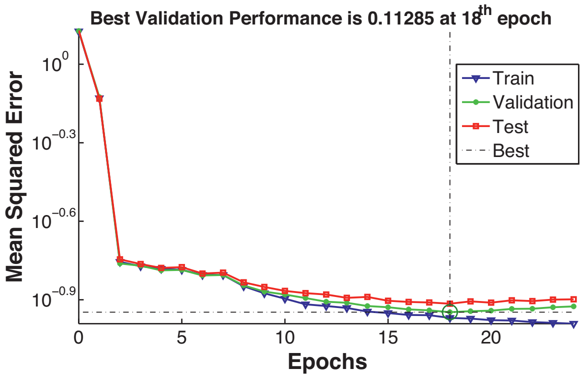

Fig. 8. Train, validation and test errors of the FFNN model developed for modeling the F_0 values of syllables.

The early stopping method is used to avoid over-fitting of the neural network. The mean-squared errors for training, validation and testing of the FFNN for modeling the intonation patterns of the sequence of syllables are shown in Fig. 8. The mean-squared error decreases with an increasing number of 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-squared 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. Finally, the predicted F_0 values of the syllables from the output of the network are obtained in normalized form. The denormalization is then carried out to convert back the normalized F_0 values to absolute F_0 values. The normalization and denormalization are carried out using Eqs. (7) and (8).

$$F_0^{Norm} = 2\left\{\frac{F_0 - F_0^{min}}{F_0^{max} - F_0^{min}}\right\} - 1 \tag{7}$$

$$F_0 = \frac{(1 + F_0^{Norm})F_0^{max} + (1 - F_0^{Norm})F_0^{min}}{2} \tag{8}$$

where F_0^{Norm} is normalized F_0 value, F_0^{max} and F_0^{min} are maximum and minimum values of absolute F_0.

5.1. Evaluation of the intonation model

The prediction performance of the FFNN model is evaluated by means of objective and subjective tests. In this study, F_0 values of the syllables are also modeled using Linear Regression (LR) and CART with the same PCPA features used for FFNN. The performance of the FFNN model is compared with LR and CART models developed by PCPA features, and the intonation model of Festival, in predicting the average F_0 values of the syllables. The details of objective and subjective tests used for evaluating the models are discussed in the following subsections.

5.1.1. Objective evaluation

The intonation model is evaluated with the syllables in the test set. The three average F_0 values located at start, middle and end positions 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 prediction performance of the FFNN model with PCPA features is given in

Table 5
Performance of LR, CART and FFNN models for predicting the F_0 values of the syllables.

Model (features)	F_0 position in the syllable	% predicted syllables within the deviation					Objective measures		
		2%	5%	10%	15%	25%	μ (Hz)	σ (Hz)	$\gamma_{X,Y}$
	Start	6.31	14.96	29.87	43.29	62.10	54.21	47.17	0.65
CART (Festival)	Middle	8.65	17.67	34.13	47.10	67.32	44.47	38.13	0.67
	End	6.02	10.93	25.59	41.16	58.35	56.82	48.19	0.64
	Start	12.56	27.15	51.97	70.50	82.19	41.01	33.93	0.77
CART (PCPA)	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
	Start	10.11	21.32	45.39	61.50	80.91	43.72	39.10	0.74
LR (PCPA)	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
	Start	14.35	31.26	55.25	73.16	87.34	32.19	27.13	0.82
FFNN (PCPA)	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 5. Column 2 of Table 5 indicates F_0 position of syllable and columns 3–7 indicates the percentage of syllables predicted within different deviations from their actual F_0 values. The deviation (D_i) is calculated 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.

The prediction accuracy 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 F_0 values is shown in columns 8–10 in Table 5.

The computation of objective measures is given below:

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

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 duration 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

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

where 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 performance of FFNN is compared with that of LR and CART models. The prediction performance of the models along with the prediction performance of intonation model using CART with Festival default features is also given in Table 5. From Table 5, it is observed that the prediction performance of LR, CART and FFNN models using PCPA features is better than CART model using Festival features. This indicates that the proposed syllable specific features are more appropriate for accurate prediction of F_0 values. It is also observed that among LR, CART and FFNN models, the prediction performance of LR model is 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. Among all models, FFNN performs better in predicting the intonation patterns of sequence of syllables. From this we can hypothesize that neural network models capture the inherent complex relationships between PCPA features and F_0 values of syllables reasonably well compared to other models.

Table 6
Mean opinion scores for the quality of speech by using different intonation models.

Model with input features	Mean opinion score (MOS)	
	Intelligibility	Naturalness
(1) CART model with Festival features	3.24	2.60
(2) LR model with PCPA features	3.29	2.64
(3) CART model with PCPA features	3.37	2.70
(4) FFNN model with PCPA features	**3.41**	**2.79**

5.1.2. *Subjective evaluation*

Naturalness and intelligibility are two important features to measure the quality of synthesized speech. Naturalness can be defined as, how close the synthesized speech is to human speech, whereas intelligibility is defined as how well the message is understood from the speech. Here, the intonation models developed using LR, CART and FFNN with PCPA features are incorporated into baseline TTS system. In this work, 20 subjects in the age group of 23–35 years 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. Ten sentences were selected randomly and the synthesized speech signals were played through headphones to evaluate the quality. Subjects have to assess the quality on a 5-point scale for each of the synthesized sentences (Reddy and Rao, 2011). The perceptual rankings of 5-point scale are as follows: 1 – unsatisfactory, 2 – poor, 3 – fair, 4 – good and 5 – excellent. Subjects have assessed the quality by listening to the synthesized sentences generated by using predicted F_0 values from LR, CART and FFNN models using PCPA features, and CART model using Festival features. Mean opinion score (MOS) is calculated by averaging the perceptual rankings given by the subjects for all the synthesized sentences. In this work, MOS values are calculated for the assessment of naturalness and intelligibility. Table 6 shows MOS values for the models mentioned above. From Table 6, it is observed that the MOS values of FFNN model for both naturalness and intelligibility are better compared to other 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 for the observed differences in the pairs of MOSs between proposed FFNN model and other models (shown in Table 6) are given in Table 7. From Table 7, 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 better than that using the other models at the perceptual level. The level of confidence between CART models using Festival default features and PCPA features is also given in the last row of Table 7. From this, we conclude that the accuracy of prediction of F_0 values of the syllables has improved by using syllable specific features compared to Festival default features.

6. Proposed intonation model using two-stage feedforward neural networks

From the existing works, it was observed that tilt parameters can capture the rise-fall pattern of the intonation contour (Taylor, 2000). Intonation patterns for the sequence of syllables may be better predicted by including tilt parameters, in addition to linguistic and production constraints. But, the tilt parameters may not be derived directly from the input text. Therefore, in this study, first, the tilt parameters are derived from the linguistic and production constraints using

Table 7
Level of confidence values for different intonation models.

Models	Level of confidence (%)	
	Intelligibility	Naturalness
(4)–(1)	>99.0	>99.5
(4)–(2)	>97.5	>99.0
(4)–(3)	>90.0	>97.5
(1)–(3)	>97.5	>97.5

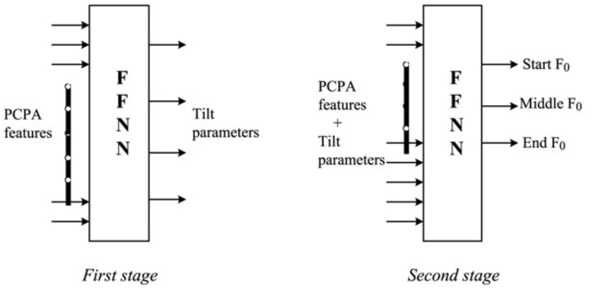

Fig. 9. Block diagrams of first and second stages of two-stage intonation model.

FFNN. Later, the derived tilt parameters are used for predicting the intonation contours. Hence, in this work a two-stage FFNN model is proposed for predicting the desired F_0 values using tilt parameters.

6.1. Principle of two-stage intonation model

The positional, contextual, phonological and articulatory features can be extracted for each syllable from the text transcription of each utterance (text corpus). The tilt parameters and the three F_0 values of each syllable can be extracted from the acoustic data (speech corpus). Therefore, two-stage intonation model consists of two FFNNs. In the first stage, one of the FFNNs is used to model the tilt parameters from PCPA features (Fig. 9). In the second stage, the other FFNN is used to model the F_0 values of the syllables from the combination of PCPA and the tilt parameters (Fig. 9).

In the testing phase, from the text transcription, PCPA features are extracted for each syllable of the test set, and they are fed to the first stage of the network. The tilt parameters of the syllables are predicted at the output of the first stage based on the knowledge gained in the training phase. The predicted tilt parameters are concatenated with the PCPA features, and they are given as input to the second stage of the network to predict the F_0 values at the output of second stage.

The final optimal structures used in two-stage intonation model are 35L 69N 15N 4L and 39L 80N 19N 3L after exploring different structures. The overall architecture of the two-stage intonation model using FFNNs for predicting the F_0 values is given in Fig. 10.

6.2. Evaluation of two-stage intonation model

Performance of the two-stage intonation model is evaluated by using objective and subjective measures. The prediction performance of the two-stage intonation model using FFNNs is compared with that of the two-stage LR and CART models. Prediction performance of the intonation models developed using PCPA and PCPA+tilt features is also compared.

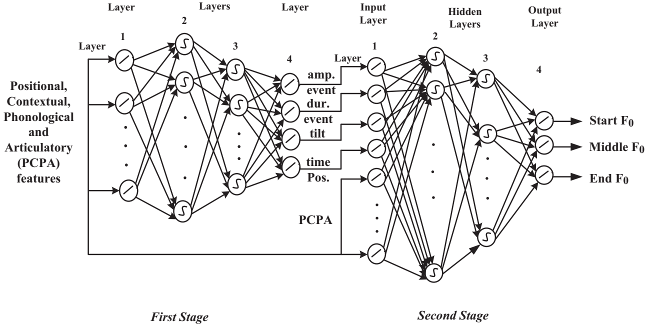

Fig. 10. Architecture of two-stage intonation model using FFNNs.

Table 8
Prediction performance of two-stage intonation models developed using LR, CART and FFNNs with PCPA and tilt parameters, and FFNN model with only PCPA features.

Model	F_0 position in the syllable	% predicted syllables within the deviation					Objective measures		
		2%	5%	10%	15%	25%	μ (Hz)	σ (Hz)	$\gamma_{X,Y}$
Single-stage 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
Two-stage CART	Start	28.93	55.24	77.91	80.29	90.38	25.74	23.70	0.89
	Middle	28.99	58.14	79.27	83.99	92.29	23.91	20.49	0.90
	End	23.27	50.98	72.21	79.93	88.60	26.92	24.19	0.88
Two-stage LR	Start	19.39	33.71	58.23	73.07	85.39	32.19	27.56	0.83
	Middle	23.38	39.90	65.39	74.66	90.18	29.36	25.54	0.86
	End	17.12	36.23	59.96	71.33	83.08	33.97	28.01	0.82
Two-stage FFNN	Start	30.48	59.83	80.99	88.88	95.46	19.46	18.92	0.94
	Middle	31.43	61.79	81.32	89.86	96.73	19.13	17.19	0.95
	End	25.13	52.68	74.46	84.23	92.56	25.65	21.11	0.91

6.2.1. *Objective evaluation*

Performance of the intonation models for predicting the F_0 values of the syllables is given in Table 8. From the results shown in Table 8, it is observed that the two-stage intonation model using FFNNs performs better than the two-stage intonation model using LR and CART models. It is also observed that there is a significant improvement in the prediction accuracy of F_0 values by two-stage intonation models (with PCPA and tilt parameters) compared to single-stage intonation model developed using FFNN (with only PCPA features). This improvement in the prediction performance is mainly due to the contribution of tilt parameters. From this we can hypothesize that the tilt parameters have the ability to capture linguistic relevance information (shape of the intonation contour) from the acoustic signal.

The prediction performance of the two-stage intonation models developed using LR, CARTs and FFNNs is also examined using scatter plots shown in Fig. 11. The scatter plots are generated by jointly plotting the actual and the predicted F_0 values of syllables. In ideal case, the predicted values should coincide with the actual values. This case is represented by dotted line in Fig. 11. The thick solid line in Fig. 11 represents average predicted F_0 vs. the average F_0 values of the syllables. The angle between the solid line and dotted line (diagonal) is inversely proportional to accuracy in prediction.

From the scatter plots, it is observed that the angle between solid and dotted lines is very less in case of FFNN model compared to other models. From Fig. 11, it is observed that predicted F_0 values are more deviated from the actual values at lower and higher F_0 values. It is also observed that the predicted F_0 values of LR and CART models are more deviated from actual F_0 values compared to FFNN model. The prediction performance of the FFNN model seems to be better in the range of 140–250 Hz for the start F_0, 175–280 Hz for the middle F_0, and 140–270 Hz for the end F_0.

For demonstrating the accuracy of prediction of the F_0 values of the sequence of syllables at the utterance level, the actual and the predicted syllable F_0 values of an utterance *"bArAmotI eArsTriper pAsh die kud∼i pon∼chish kimi gele gaond"*, are plotted in Fig. 12. From Fig. 12 it is observed that the predicted F_0 patterns of start, middle and end positions of syllables by the proposed two-stage intonation model using FFNNs is very close to the original contours, compared to two-stage intonation model using LR and CARTs.

6.2.2. *Subjective evaluation*

The subjective tests are conducted by incorporating two-stage intonation models developed by using LR, CART and FFNNs into baseline TTS system. The overall architecture representing the incorporation of two-stage models into TTS system is shown in Fig. 13. The MOS values for naturalness and intelligibility of the two-stage intonation models are given in Table 9. From Table 9, it is observed that the MOS values of two-stage intonation model developed using FFNNs is better compared to other models for both naturalness and intelligibility. The level of confidence for the observed differences in the pairs of MOSs between the proposed two-stage intonation model using FFNNs and

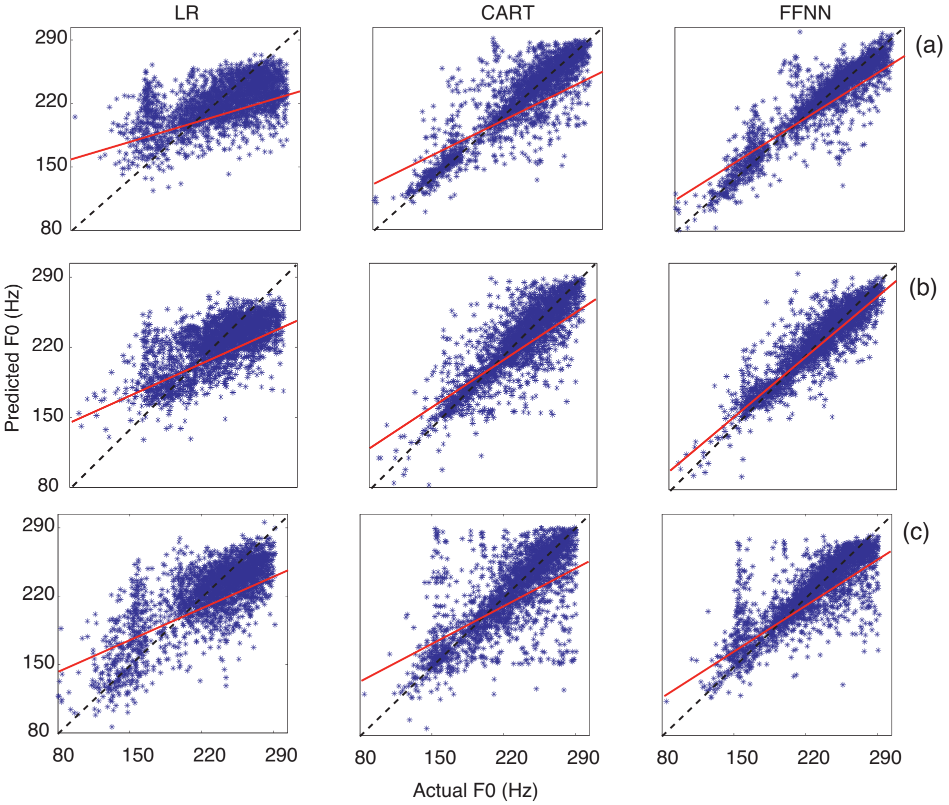

Fig. 11. Prediction performance of (a) start F_0, (b) middle F_0, and (c) end F_0 values from two-stage intonation models developed using LR, CARTs and FFNNs with scatter plots.

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

Model with input features	Mean opinion score (MOS)	
	Intelligibility	Naturalness
(1) FFNN using only PCPA features	3.41	2.79
(2) LR using PCPA and tilt parameters	3.53	2.84
(3) CART using PCPA and tilt parameters	3.60	2.91
(4) FFNN using PCPA and tilt parameters	**3.69**	**2.97**

other models (shown in Table 9) are given in Table 10. From Table 10, it is observed that the level of confidence is high (> 95.0) 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 proposed two-stage intonation model using FFNNs is significantly better than the two-stage intonation models using CARTs and LR at perceptual level. The level of confidence between the two-stage

Table 10
Level of confidence values for different intonation models.

Models	Level of confidence (%)	
	Intelligibility	Naturalness
(4)–(1)	>99.5	>99.0
(4)–(2)	>99.0	>97.5
(4)–(3)	>97.5	>95.0

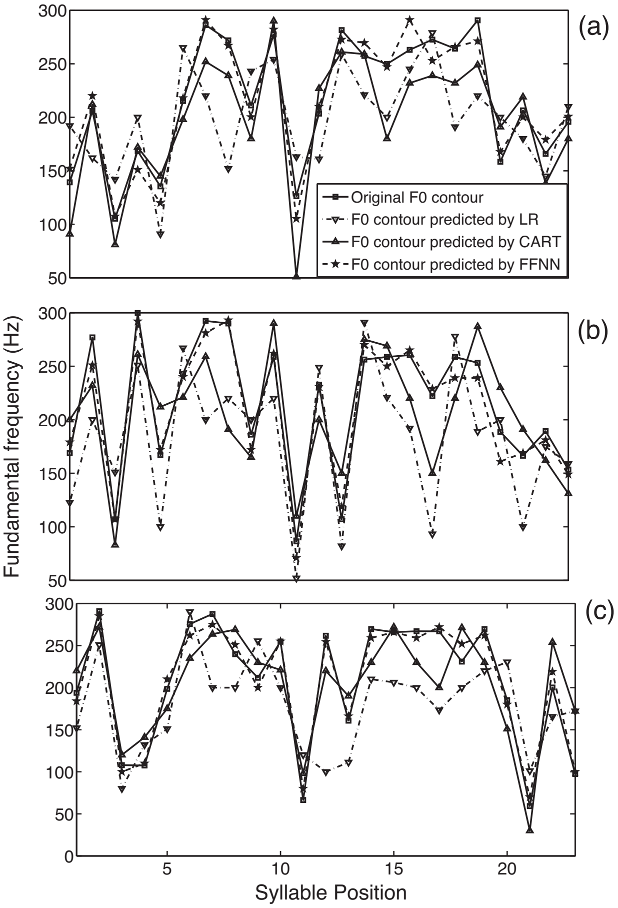

Fig. 12. Comparison of predicted (a) start F_0, (b) middle F_0, and (c) end F_0 patterns from two-stage LR, CART and FFNN models with original F_0 patterns for the utterance *"bArAmotI eArsTriper pAsh die kud∼i pon∼chish kimi gele gaond"*.

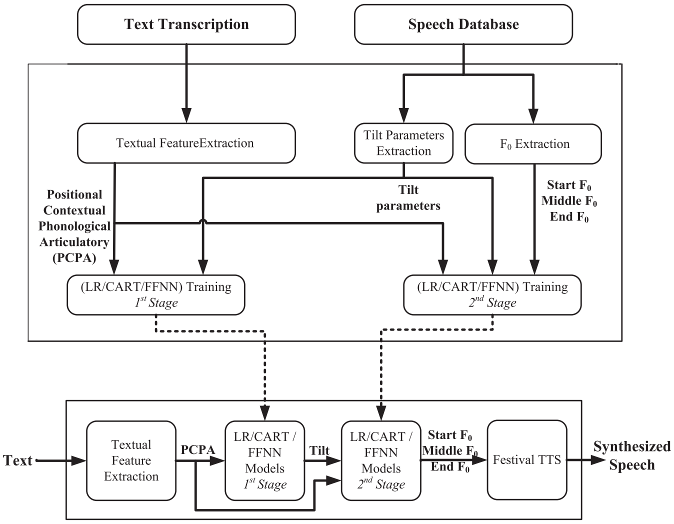
Text Transcription
Speech Database
Textual FeatureExtraction
Tilt Parameters Extraction
F0 Extraction
Positional Contextual Phonological Articulatory (PCPA)
Tilt parameters
Start F0 Middle F0 End F0
(LR/CART/FFNN) Training 1st Stage
(LR/CART/FFNN) Training 2nd Stage
Text
Textual Feature Extraction
PCPA
LR/CART / FFNN Models 1st Stage
Tilt
LR/CART / FFNN Models 2nd Stage
Start F0 Middle F0 End F0
Festival TTS
Synthesized Speech

Fig. 13. Architecture of the two-stage intonation models in TTS system.

Table 11
Optimal network architectures obtained for individual features for predicting the durations, intonation and intensities of syllables.

	First stage	Second stage
Positional	13L 25N 8N 4L	17L 36N 6N 3L
Contextual	12L 26N 5N 4L	16L 30N 7N 3L
Phonological	7L 14N 3N 4L	11L 23N 4N 3L
Articulatory	15L 33N 7N 4L	19L 37N 8N 3L

intonation model using the PCPA features and the tilt parameters, and FFNN model using only PCPA features is given in the first row of Table 10. The results indicate that the contribution of tilt parameters to PCPA features is significant in improving the quality of TTS system at the perceptual level.

From the objective and subjective evaluation results, it is observed that models using combination of tilt parameters with PCPA features outperform the models developed using only PCPA features. Among different two-stage models, the model based on FFNN has shown better prediction performance.

7. Influence of individual features for predicting the intonation

From the above results, it is observed that the prediction of F_0 values by the two-stage intonation model using FFNNs is comparatively better than the other models. Hence, for studying the effect of positional, contextual, phonological and articulatory features in predicting the syllable F_0 values, separate models were developed using two-stage approach. The network structures used for studying the effect of the positional, contextual, phonological and articulatory features for two-stages are given in Table 11. The prediction performance of the two-stage intonation model using individual features is given in Table 12. From the results, it is observed that the F_0 values of syllables depend on positional, contextual, phonological and articulatory features. However, the positional features, seem to perform slightly better compared to the other features for predicting the syllable F_0 values. But, the combination of all features, performs better than the individual features for predicting the F_0 values.

Table 12
Prediction performance of two-stage FFNN model with individual features for predicting F_0 values of syllables.

Input features	F_0 position in syllable	% predicted syllables within deviation					Objective measures		
		2%	5%	10%	15%	25%	μ (Hz)	σ (Hz)	$\gamma_{X,Y}$
Positional	Start	30.16	59.58	80.22	88.88	94.77	19.58	19.01	0.93
	Middle	31.43	61.79	80.83	89.30	96.65	19.42	16.99	0.94
	End	24.80	51.12	73.60	84.18	92.39	26.01	22.23	0.91
Contextual	Start	29.87	58.72	79.69	88.80	94.75	19.96	18.99	0.93
	Middle	30.94	59.86	80.18	88.93	96.32	19.46	17.25	0.94
	End	24.56	50.84	72.86	83.49	92.28	26.28	23.30	0.90
Phonological	Start	29.83	58.43	78.69	87.45	94.64	20.17	18.53	0.92
	Middle	30.60	59.21	79.81	88.59	96.19	19.54	17.97	0.93
	End	24.19	50.14	72.53	83.29	92.25	26.62	22.28	0.90
Articulatory	Start	28.19	55.27	79.11	88.83	94.76	19.67	18.27	0.94
	Middle	30.27	57.90	80.28	89.50	95.91	19.43	17.30	0.94
	End	25.01	52.17	73.90	85.59	91.93	25.99	23.13	0.90
All	Start	30.48	59.83	80.99	88.88	95.46	19.46	18.92	0.94
	Middle	31.43	61.79	81.32	89.86	96.73	19.13	17.19	0.95
	End	25.13	52.68	74.46	84.23	92.56	25.65	21.11	0.91

Table 13
Prediction performance of FFNN model using different constraints (L: linguistic, P: production, D: duration, and I: intensity).

Input constraints	Output features	% predicted syllables within deviation					Objective measures		
		2%	5%	10%	15%	25%	μ (Hz)	σ (Hz)	$\gamma_{X,Y}$
L + P + D	Start F_0	32.11	62.72	81.31	90.01	96.99	18.37	16.13	0.95
	Middle F_0	34.69	64.18	82.97	91.73	98.13	17.99	13.48	0.96
	End F_0	26.98	57.31	77.15	86.59	92.97	23.41	19.69	0.92
L + P + I	Start F_0	31.53	61.10	81.14	89.19	95.56	19.03	17.79	0.94
	Middle F_0	33.07	63.91	81.53	90.27	97.43	18.75	14.87	0.95
	End F_0	25.65	55.35	75.29	86.11	92.78	24.90	20.16	0.91
L + P + D + I	Start F_0	33.64	64.19	83.82	90.15	97.92	17.51	14.90	0.96
	Middle F_0	35.87	67.23	83.91	92.25	98.79	16.13	13.82	0.96
	End F_0	28.99	61.20	81.63	87.56	93.20	22.93	18.17	0.92

8. Influence of duration and intensity constraints

In speech signal, duration, intonation and intensity patterns of a sequence of sound units are interrelated at some higher level through emphasis (stress) and prominence of the words and phrases. But, representation of the feature vector to capture these dependencies is difficult. In this study, the durations of the syllables are considered as duration constraints and the intensities of the syllables are considered as intensity constraints. The duration and intensity constraints can be obtained from the duration and intensity models. For studying the influence of these constraints in predicting the intonation, three FFNN models are developed: (i) model with duration constraints, (ii) model with intensity constraints, and (iii) model with duration and intensity constraints together. The performance of these models is given in Table 13. The results indicate that the prediction performance has improved by imposing the constraints. From Table 13, it is observed that the percentage of syllables predicted within 2–25% has increased by including duration and intensity constraints. A similar phenomenon is observed in objective measures also. The prediction performance of the proposed models is observed to be significantly better compared to our previous intonation model (Reddy and Rao, 2011). The superior performance of the proposed models is mainly due to inclusion of articulatory and tilt features.

9. Summary and conclusions

A two-stage intonation model using neural networks is proposed for predicting the F_0 values of a sequence of syllables. Linguistic and production constraints represented by positional, contextual, phonological and articulatory features are proposed for modeling the intonation patterns. Tilt parameters are also explored in addition to linguistic and production constraints. The intonation model is evaluated by objective measures such as average prediction error, standard deviation and correlation coefficient. The prediction accuracy of the proposed two-stage intonation model using FFNNs is better than the two-stage CART and LR models. It is observed that the prediction accuracy has significantly improved with the inclusion of tilt parameters. Verification of the prediction accuracy of the intonation models was also carried out by conducting perceptual tests on synthesized speech by incorporating the derived F_0 values from the models. The mean opinion scores of the perceptual tests have indicated that the quality of synthesized speech with the predicted F_0 values by the two-stage FFNN model was better than the other models. Both objective and subjective results have indicated the superior performance of the proposed two-stage FFNN model. The prediction accuracy has further improved by imposing the duration and intensity constraints. Performance may be further improved by accurate labeling and diversity of data.

Acknowledgements

This work was carried out as a part of the project sponsored by Department of Information Technology (DIT), Government of India. We would like to acknowledge DIT for offering this project. 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. Finally, we would like to thank Mr. Guruprasad Seshadri, Mr. Sourjya Sarkar and Mr. Rohan Banerjee for helping us in proofreading the article.

References

Adriaens, L.M.H., 1991. Ein Modell Deutscher Intonation. PhD Thesis, Technical University Eindhoven, Eindhoven.

Association, I.P., 1999. Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet. Cambridge University Press.

Beckman, M., Pierrehumbert, J., 1986. Intonational structure in Japanese and English. In: Phonology Yearbook, vol. 3, Cambridge University Press, pp. 255–309.

Black, A.W., Lanzo, K., 2009. Building synthetic voices in the Festival speech synthesis system. http://www.festvox.org, December 2009.

Black, A.W., Taylor, P., Caley, R., 2009. The Festival speech synthesis system. Manual and source code available at www.cstr.ed.ac.uk/projects/festival.html

Bodo, A., Buza, O., Toderean, G., Zsolt, A.B., 2009. Experiments with the prediction and generation of Romanian intonation. In: Proc. of the 5th Conf. on Speech Technology and Human–Computer Dialogue, June, pp. 1–9.

Botinis, A., Granstrom, B., Mobius, B., 2001. Developments and paradigms in intonation research. Speech Communication 33, 263–296.

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

Campillo, F., Banga, E.R., 2011. Multiple fo contour parallel Viterbi search for unit selection speech synthesis. IEEE Electronics Letters 47 (August), 937–938.

Damadi, M.S., Azami, B., Eslami, M., 2010. Prosody generation in TTS system for Azeri. In: IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM), July, pp. 1335–1338.

de Pijper, J.R., 1983. Modeling British English Intonation. Foris, Dordrecht.

Fujisaki, H., Ohno, S., 1995. Analysis and modeling of fundamental frequency contours of English utterances. In: Proceedings Eurospeech 95, (Madrid), pp. 985–988.

Fujisaki, H., Hirose, K., Halle, P., Lei, H., 1971. A generative model for the prosody of connected speech in Japanese. Annual Report of Engineering Research Institute 30, 75–80.

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: Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 2039–2042.

Fujisaki, H., Ohno, S., Nakamura, K., Guirao, M., Gurlekian, J., 1994. Analysis and synthesis of accent and intonation in standard Spanish. In: Proc. Int. Conf. Spoken Language Processing, Yokohama, pp. 355–358.

Fujisaki, H., Ohno, S., Yagi, T., 1997. Analysis and modeling of fundamental frequency contours of Greek utterances. In: Proceedings Eurospeech 97, Rhodes, Greece, pp. 465–468.

Fujisaki, H., 1983. Dynamic characteristics of voice fundamental frequency in speech and singing. In: MacNeilage, P.F. (Ed.), The Production of Speech. Springer-Verlag, New York, USA, 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, USA, pp. 347–355.

Gårding, E., 1983. A generative model of intonation. In: Cutler, A., Ladd, D.R. (Eds.), Prosody: Models and Measurements. Springer-Verlag, Berlin, Germany, pp. 11–25.

Gerazov, B., Ivanovski, Z., 2012. Analysis of extracted pitch contours across speakers for intonation modelling in TTS synthesis. In: 2012 5th International Symposium on Communications Control and Signal Processing (ISCCSP), May, pp. 1–4.

Ghosh, K., Reddy, V.R., Narendra, N.P., Maity, S., Koolagudi, S.G., Rao, K.S., 2010. Grapheme to phoneme conversion in Bengali for Festival based TTS framework. In: 8th International Conference on Natural Language Processing (ICON), IIT Kharagpur, Kharagpur, India, December.

Gronnum, N., 1992. The Groundworks of Danish Intonation: An Introduction. Museum Tusculanum Press, Copenhagen.

Gronnum, N., 1995. Superposition and subordination in intonation – a non-linear approach. In: Proceedings of the 13th International Congress – Phon. Sc. Stockholm, Stockholm, pp. 124–131.

Gu, W., Hirose, K., Fujisaki, H., 2006. Modeling the effects of emphasis and question on fundamental frequency contours of Cantonese utterances. IEEE Transactions on Audio, Speech, and Language Processing 14 (July), 1155–1170.

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

Hwang, S.H., Chen, S.H., 1994. Neural network-based F0 text-to-speech synthesizer for Mandarin. IEE Proceedings of the Image Signal Processing 141, 384–390.

Jilka, M., Mohler, G., Dogil, G., 1999. Rules for generation of TOBI-based American English intonation. Speech Communication 28, 83–108.

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

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

Mixdorff, H., Fujisaki, H., 1994. Analysis of voice fundamental frequency contours of German utterances using a quantitative model. In: Proceedings of the International Conference Spoken Language Processing, vol. 4, Yokohama, pp. 2231–2234.

Moore, B.C.J. (Ed.), 1989. An Introduction to the Psychology of Hearing, 3rd ed. Academic Press, San Diego, CA.

Murty, K., Yegnanarayana, B., 2008. Epoch extraction from speech signals. IEEE Transactions on Audio, Speech, and Language Processing 16, 1602–1613.

Nacinovic, L., Martincic-Ipsic, S., Ipsic, I., 2010. Intonation modeling for Croatian speech synthesis. In: Proceedings of the 33rd International Convention MIPRO, May, pp. 766–770.

Narendra, N.P., Rao, K.S., Ghosh, K., Reddy, V.R., Maity, S., 2011. Development of syllable-based text to speech synthesis system in Bengali. International Journal of Speech Technology, Springer 14 (3), 167–181.

Odéjobí, O.A., Beaumont, A.J., Wong, S.H.S., 2006 Oct. Intonation contour realisation for Standard Yorùbá text-to-speech synthesis: a fuzzy computational approach. Computer Speech and Language 20, 563–588.

Ode, C., 1989. Russian Intonation: A Perceptual Description. Rodopi, Amsterdam.

O'Shaughnessy, D., 1984. Design of a real-time French text-to-speech system. Speech Communication 3 (3), 233–243.

O'Shaughnessy, D., 1987. Speech Communication: Human and Machine. Addison-Wesley Publishing Company.

Peng, Z., Lihong, W., Sheng, L., 2008. On fundamental frequency contour synthesis and control method for Chinese Speech Synthesis. In: 27th Chinese Control Conference, July, pp. 739–742.

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

Rao, K.S., Yegnanarayana, B., 2007. Modeling durations of syllables using neural networks. Computer Speech and Language 21 (April), 282–295.

Rao, K.S., Yegnanarayana, B., 2009. Intonation modeling for Indian languages. Computer Speech and Language 23 (April), 240–256.

Rao, K.S., Acquisition and incorporation prosody knowledge for speech systems in Indian languages. PhD Thesis, Dept. of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai, India, May 2005.

Reddy, V.R., Rao, K.S., 2011. Intonation Modeling using FFNN for Syllable based Bengali text to speech synthesis. In: Proc. Int. Conf. Computer and Communication Technology, MNNIT, Allahabad, pp. 334–339.

Scordilis, M.S., Gowdy, J.N., 1989. Neural network based generation of fundamental frequency contours. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, vol. 1, Glasgow, Scotland, May, pp. 219–222.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J., 1992. ToBI: a standard for labelling English prosody. In: Proc. Int. Conf. Spoken Language Processing, Banff, Alberta, pp. 867–870.

Sonntag, G.P., Portele, T., Heuft, B., 1997. Prosody generation with a neural network: weighing the importance of input parameters. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Munich, Germany, April, pp. 931–934.

Sproat, R. (Ed.), 1998. Multilingual Text-to-Speech Synthesis: The Bell Labs Approach. Kluwer Academic Publishers, Norwell, MA, USA.

Tamura, S., Tateishi, M., 1997. Capabilities of a Four-Layered Feedforward Neural Network: Four Layers Versus Three. IEEE Transactions on Neural Networks 8 (March), 251–255.

Taylor, P.A., 1995. The rise/fall/connection model of intonation. Speech Communication 15 (15), 169–186.

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

Terken, J., 1993. Synthesizing natural sounding intonation for Dutch: rules and perceptual evaluation. Computer Speech and Language 7, 27–48.

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

Traber, C., 1992. F0 generation with a database of natural f0 patterns and with a neural network. In: Benoit, C., Wawallis, T.R., Bailey, G. (Eds.), Talking Machines: Theories, Models, and Designs. Elsevier Science, Amsterdam, The Netherlands, pp. 287–304.

Uslu, B., Ilk, H., 2009. Fujisaki intonation model in Turkish text-to-speech synthesis. In: IEEE 17th Signal Processing and Communications Applications Conference, April, pp. 844–847.

Vainio, M., Altosaar, T., 1998. Modeling the microprosody of pitch and loudness for speech synthesis with neural networks. In: Proc. Int. Conf. Spoken Language Processing, Sydney, Australia.

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

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