Available online at www.sciencedirect.com
ScienceDirect
Computer Speech and Language 21 (2007) 282–295
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

Modeling durations of syllables using neural networks

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

[a] Department of Electronics and Communication Engineering, Indian Institute of Technology Guwahati, North Guwahati, Guwahati 781 039, India
[b] Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai 600 036, India

Received 29 August 2005; received in revised form 7 June 2006; accepted 9 June 2006
Available online 21 July 2006

Abstract

In this paper, we propose a neural network model for predicting the durations of syllables. A four layer feedforward neural network trained with backpropagation algorithm is used for modeling the duration knowledge of syllables. Broadcast news data in three Indian languages Hindi, Telugu and Tamil is used for this study. The input to the neural network consists of a set of features extracted from the text. These features correspond to phonological, positional and contextual information. The relative importance of the positional and contextual features is examined separately. For improving the accuracy of prediction, further processing is done on the predicted values of the durations. We also propose a two-stage duration model for improving the accuracy of prediction. From the studies we find that 85% of the syllable durations could be predicted from the models within 25% of the actual duration. The performance of the duration models is evaluated using objective measures such as average prediction error (μ), standard deviation (σ) and correlation coefficient (γ).

1. Introduction

During production of speech, human beings seem to impose automatically the required duration, intonation and coarticulation patterns on the sequence of sound units. These patterns reflect the linguistic and production constraints in the prosody (duration and intonation) and coarticulation in speech. The knowledge of these constraints is implicit in the speech signal in the sense that it is difficult to articulate the rules governing this knowledge. But lack of this knowledge can easily be perceived while attempting to communicate the desired message through speech. Capturing this knowledge is essential, for example, for synthesizing natural sounding speech from a text.

One of the most striking features of speech produced by humans is its variability, especially in the prosody (Smith, 2002). Speech synthesizers will sound more natural, only if they reproduce some of this variability. Caroline L. Smith conducted a study to investigate the durational variability in readings of an extended passage of text by an American English speaker. The focus of the study was to observe how the structure of the

* Corresponding author. Tel.: +91 361 2582516; fax: +91 44 2690762.
E-mail addresses: ksrao@iitg.ernet.in (K.S. Rao), yegna@cs.iitm.ernet.in (B. Yegnanarayana).

0885-2308/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2006.06.003

topics (*topic shift, topic continuation and topic elaboration*) in the spoken material can affect the variability in the acoustic durations (Smith, 2002). The following observations were made in this study: (1) Speech rate tends to be slower around a topic shift than at other transitions. (2) Sentence-final lengthening is similar in topic shifts and continuations. However, at a topic continuation, speech rate increases significantly between the end of the present sentence and the beginning of the following sentence. (3) Topic elaborations have significantly less sentence-final lengthening than other transitions. Pauses are of similar durations for elaborations and continuations. Speech rate is faster in elaborations, but does not change at the transition. This study indicates the variability in durations due to the structure present in the spoken material, and the variability in turn reflects in the fluency.

The implicit knowledge of prosody is usually captured using modeling techniques. In this paper, we focus on modeling the duration knowledge in speech. In speech signal the duration of each sound unit is dictated by the linguistic context of the unit and the production constraints. Modeling duration may be rule-based or statistical. Rule-based methods involve analysis of segment durations manually to determine the duration constraints on the sequence of sound units. The derived rules get better in terms of accuracy, as the amount of speech data for analysis is increased. But with large amount of data the process of manually deriving the rules becomes tedious and time consuming. Hence, rule-based methods are limited to small amount of data. One of the early attempts for developing rule-based duration models was in 1970s (Klatt, 1976). The model was based on information present in linguistic and phonetic literature about different factors affecting segmental durations. The rules were derived by analyzing a set of phonetically balanced sentences. Following this model, similar models were developed for other languages like German and French (Kohler, 1988; Bartkova and Sorin, 1987). More recently, a rule-based duration model was developed for a text-to-speech (TTS) system for the Indian language Hindi (Yegnanarayana et al., 1990; Kumar, 2002). The model was developed using the broadcast news speech data in Hindi. In general, the rule-based methods are difficult to study, due to complex interaction among the linguistic features at various levels (Chen et al., 2003). Therefore the rule inference process is restricted to controlled experiments, where only a limited number of contextual factors are involved.

Statistical methods are attractive when large phonetically labeled databases are available. The statistical methods can be based on parametric or nonparametric regression models (Mixdorff et al., 2001). Examples of parametric regression models are Sums-of-Products (SOP) model, generalized linear model and multiplicative model (Mixdorff, 2002). In the SOP model the duration of a segment is represented as a sum of factors that affect the duration, and their interactions (product terms) (Santen, 1994). The generalized linear model is a variant of the SOP model. The advantage of the SOP model is that the model can be developed using small amount of data. But in the SOP model the number of different sums-of-products grows exponentially with the number of factors. Thus it is difficult to find an SOP model that best describes the data. In addition, the SOP model requires significant preprocessing of data to correct the interaction among factors and data imbalance (Goubanova et al., 2000; Sayli, 2002). The model was developed using two sets of data corresponding to two types of speaking styles, read speech and news commentary. The SOP duration models were used in Bell labs multi-lingual text-to-speech system (Sproat, 1998). Van Santen used SOP models for predicting the segment durations, and reported a correlation coefficient of 0.93 between predicted and observed segment durations for English, and 0.90 for German (Mixdorff et al., 2001; Santen, 1994). Hyunsong Chung used SOP models for predicting the durations of vowels and consonants in Korean language. For vowels, the root mean square error (RMSE) and the correlation coefficient values were observed to be 32.13 ms and 0.68, respectively, and for consonants they were 28.86 ms and 0.54, respectively (Chung, 2002a,b). On TIMIT database, RMSE of 9 ms and correlation coefficient of 0.94 were observed using SOP models (Goubanova et al., 2000).

Nonlinear regression models are either Classification and Regression Tree (CART) models or neural network models (Mixdorff, 2002; Riley, 1992). 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 (Riley, 1992). The feature vector of a segment represents the positional, contextual and phonological information. The segment duration is predicted by passing the feature vector through the tree so as to minimize the variance at each terminal node. The tree construction algorithm 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. Riley used CART model to predict the phoneme durations using their context (Riley, 1992). The training data comprised of 1500

short sentences spoken by a single speaker. The performance of the model was shown to be better compared to previous rule-based methods. Recently, CART models were used for predicting the phoneme durations and F_0 (pitch) values in the Festival framework for developing TTS systems in different languages (Black et al., 2000). On TIMIT database the performance of CART model was observed to be RMSE of 20 ms and correlation coefficient of 0.82 (Goubanova et al., 2000). For predicting the durations of vowels and consonants in Korean language, the CART model gave the RMSE of 27.51 ms for vowels and 24.2 ms for consonants, and correlation coefficient of 0.78 for vowels and 0.71 for consonants (Chung, 2002a,b). Sridhar Krishna *et al.* used CART models for predicting the segment durations in two Indian languages namely, Hindi and Telugu (Krishna and Murthy, 2004). The RMSE and correlation coefficient values were observed to be 22.86 ms, and 0.80 for Telugu, and 27.14 ms and 0.75 for Hindi, respectively.

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 durations of syllables in continuous speech (Campbell, 1990, 1992, 1993; Campbell and Isard, 1991; Barbosa and Bailly, 1994; Barbosa and Bailly, 1992; Cordoba et al., 1999; Hifny and Rashwan, 2002; Sonntag et al., 1997; Teixeira and Freitas, 2003). Campbell used a feedforward neural network trained with feature vectors, each representing six features of a syllable (Campbell, 1992). The six features are: Number of phonemes in a syllable, the nature of syllable, position in the tone group, type of foot, stress, and word class. Syllable durations are predicted with these feature vectors as input to the neural network. The durations of the phonemes are estimated from the predicted durations of the syllables using the *elasticity principle* (Campbell and Isard, 1991). Neural network models were developed using two different databases. (1) Spoken English Corpus (SEC) from broadcast news was used for examples of fluent speech with natural prosody, and (2) Spoken Corpus Readings in British English (SCRIBE) database of phonetically rich sentence readings was used for balanced segmental information. Barbosa and Bailly used a neural network model to capture the perception of rhythm in speech (Barbosa and Bailly, 1994). The model predicts the duration of a unit, known as Inter-Perceptual Center Group (IPCG). The IPCG is delimited by the onset of a nuclear vowel and the onset of the following vowel. The model was trained with the following information: Boundary marker at phrase level, sentence mode, accent marker, number of consonants in IPCG, number of consonants in coda and nature of the vowel. Campbell reported a correlation coefficient of 0.89 between observed and predicted syllable durations (Campbell, 1993). Teixeira *et al.* used neural network for predicting the segment durations, and they reported the RMSE of 19.5 ms and correlation coefficient of 0.839 (Teixeira and Freitas, 2003).

The objective in the present study is to determine whether the nonlinear neural network models can capture the implicit knowledge of the syllable duration in a language. 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 may infer that the network indeed captures the implicit relations in the input–output pairs. We propose to examine the ability of neural network models to capture the duration knowledge for speech in different Indian languages. We consider three Indian languages (Hindi, Telugu and Tamil) using syllable as the basic sound unit. The reason for choosing the 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 prediction performance of the neural network model depends on the nature of the training data used. Distributions of the durations of syllables (Fig. 1) indicate that majority of the durations are concentrated around mean of the distribution. The distribution of the syllable durations for the three languages is shown in Fig. 1. This kind of training data forces the model to be biased towards mean of the distribution. To avoid this problem, some postprocessing and preprocessing methods are proposed. Postprocessing methods modify the predicted values further using some durational constraints. Preprocessing methods involve use of multiple models, one for each limited range of duration. This requires a two-stage duration model, where the first stage is used to segregate the input into groups according to the number of models, and the second stage is used for prediction.

The paper is organized as follows: Section 2 discusses the factors that affect the syllable duration. The database used for the proposed duration model is described in Section 3. Section 4 discusses the features used as

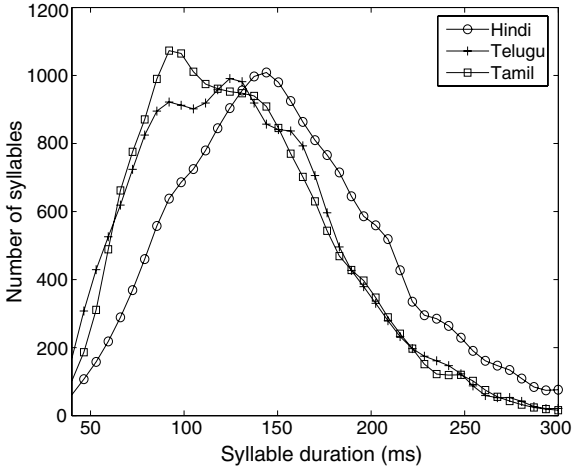

Fig. 1. Distributions of durations of syllables in the languages Hindi, Telugu and Tamil.

input to the neural network for capturing the knowledge of the syllable duration. Section 5 gives the details of the neural network model, and discusses its performance in predicting the durations of syllables. A two-stage duration model is proposed in Section 6 to reduce the error in the predicted durations of the syllables. 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.

2. Factors affecting the syllable duration

The factors affecting the durations of the basic sound units in continuous speech can be broadly categorized into phonological, positional and contextual (Huang et al., 2001). The vowel is considered as the nucleus of a syllable, and consonants may be present on either side of the vowel. The duration of a syllable is influenced by the position of the vowel, the category of the vowel and the type of the consonants associated with the vowel. The positions that affect the durations of the syllables are: Word initial position, word final position, phrase boundary and sentence ending position. The contextual factors include the preceding and following syllables. The manner and place of articulation of the syllables in the preceding and following positions also affect the duration of the present syllable. In addition, the gender of the speaker, psychological state of the speaker (happy, anger, fear, etc.), age, relative novelty in the words and words with relatively large number of syllables, also affect the duration.

3. Speech database

The database for this study consists of 19 Hindi, 20 Telugu and 33 Tamil broadcast news bulletins. In each language these news bulletins are 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 coded 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 syllable durations vary from 30 to 450 ms. The details of the data with respect to duration for the three languages are given in Table 1 (Khan et al., 2003).

4. Features for developing duration models

In this study, we use 25 features (together called as feature vector) for representing the linguistic context and production constraints of each syllable.

Table 1
Details of the broadcast news data for the languages Hindi, Telugu and Tamil

Language	# Speakers		# Utterances	# Words	# Syllables	Syllable duration (ms)	
	Male	Female				Mean	SD
Hindi	06	13	4191	26,090	50,237	157.10	57.37
Telugu	11	09	6484	25,463	84,349	133.65	54.69
Tamil	10	23	7359	30,688	100,707	132.10	48.84

These features represent positional, contextual and phonological information of each syllable. Features representing the positional information are further classified based on the position of the word in a 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 duration of a syllable may depend on the position of the word in an utterance. Therefore the word position is used for developing the duration 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 very less (less than 1%). Each segment of a syllable is encoded separately, so that each syllable identity is represented by a four-dimensional feature vector.

Context of the syllable: Syllable duration may be influenced by its adjacent syllables. Hence, for modeling the duration of a syllable, the contextual information is represented by the previous and following syllables. Each of these syllables is represented by a four-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 three-dimensional feature vector specifying the consonant–vowel structure present in the syllable.

Gender identity: The database contains speech from both male and female speakers. This gender information is represented by a single feature.

The list of features and the number of input nodes in a neural network needed to represent the features are given in Table 2.

5. Duration modeling with feedforward neural networks

A four layer feedforward neural network (FFNN) is used for modeling the durations of syllables. The general structure of the FFNN is shown in Fig. 2. 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 one-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

Table 2
List of the factors affecting the syllable duration, 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	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 a 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
Gender identity	Gender of the speaker	1

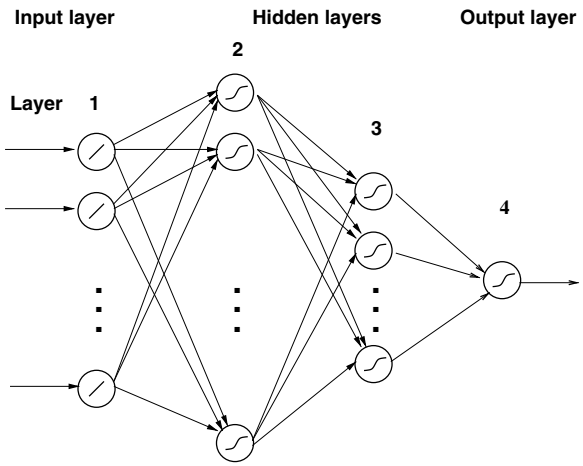

Fig. 2. Four layer feedforward neural network.

features in 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 duration 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 were 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. The nonlinear units use $tanh(s)$ as the activation function, where s is the activation value of that unit. For studying the effect of the positional and contextual factors on syllable duration, the network structures $14L\ 28N\ 7N\ 1N$ and $13L\ 26N\ 7N\ 1N$ are used, respectively. The proportions of the number of units in

each layer are similar as in the earlier network. The inputs to these networks represent the positional and contextual factors. All the input and output features are normalized to the range [−1, +1] before presenting them to the neural network. The backpropagation learning algorithm is used for adjusting the weights of the network to minimize the mean squared error for each syllable duration (Yegnanarayana, 1999).

A separate model is developed for each of the three languages. For Hindi 35,000 syllables are used for training the network, and 11,222 syllables are used for testing. For Telugu 64,000 syllables are used for training, and 17,630 are used for testing. For Tamil 75,000 syllables are used for training, and 21,493 are used for testing. For each syllable a 25-dimensional input vector is formed, representing the positional, contextual and phonological features. The duration of each syllable is obtained from the timing information available in the database. Syllable durations seem to follow a logarithmic distribution, and hence the logarithm of the duration is used as the target value (Campbell, 1992). Since the number of training samples are significantly larger than the number of weights, the possibility of over-fitting of the training data is rare. Also, the training was stopped after about 500 epochs in each case, as there was no significant reduction in the error for further increase in the number of epochs. The learning ability of the network from the training data can be observed from training error. The training errors for neural network models for the three languages are shown in Fig. 3. The decreasing trend in the training error indicates that the network is capturing the implicit relation between the input and output.

The duration model is evaluated using syllables in the test set. For each syllable in the test set, the duration is predicted using the FFNN by giving the feature vector of each syllable as input to the neural network. The deviation of the predicted duration from the actual duration is obtained. The prediction performance of the models for the three languages is shown in Fig. 4. Each plot represents the average predicted duration *vs.* the average duration of a syllable. These duration values are derived as follows: The range of the available syllable durations is uniformly divided into a finite number of bins. In this work, the range of duration for each bin is 10 ms. The average duration corresponds to the mean of the syllable durations that fall into a particular bin. The average predicted duration corresponds to the mean of the predicted durations for the syllables whose actual durations fall into a particular bin. Fig. 4 gives the prediction performance only in the average sense. For illustrating the actual prediction performance, a set of 1000 syllables were chosen randomly from the test set of Tamil language. Fig. 5 shows the points corresponding to actual *vs.* predicted syllable durations. The plot shows that the prediction is better in the range of 80–170 ms. A similar performance can be observed in Fig. 4.

For studying the effect of positional and contextual factors on syllable duration, the features associated with the syllable position and syllable context were used separately. The features representing the positional factors are: (a) Syllable position in the phrase (three-dimensional feature), (b) syllable position in the word

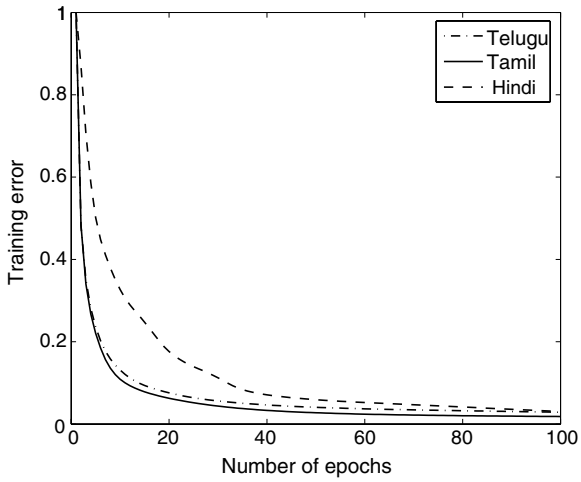

Fig. 3. Training errors for the neural network models for three Indian languages (Hindi, Telugu and Tamil).

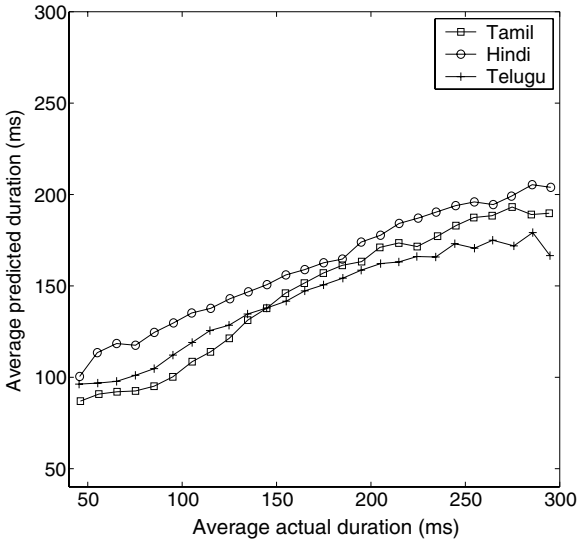

Fig. 4. Prediction performance of the neural network models for the languages Hindi, Telugu and Tamil.

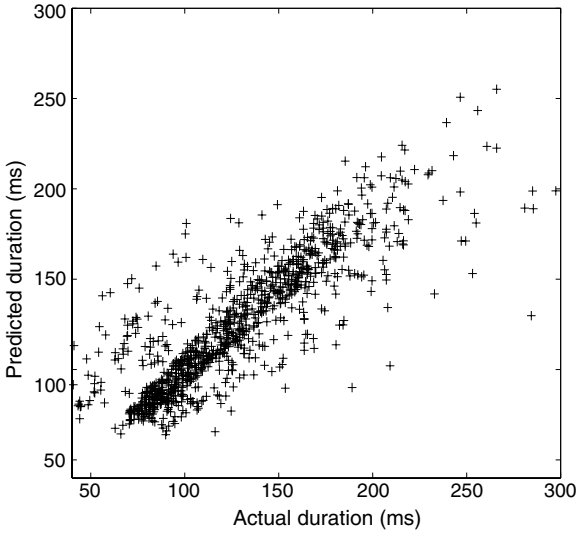

Fig. 5. Actual *vs.* predicted duration values of the syllables in Tamil language.

(three-dimensional feature), (c) word position in the phrase (three-dimensional feature), (d) syllable identity (four-dimensional feature) and (e) identity of gender. Features representing the contextual factors are the identities of the present syllable, its previous and following syllables and the identity of gender. The percentages of syllables predicted within different deviations from their actual durations are given in Table 3. 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 durations, respectively. The first column indicates the number of syllables specific to the particular language used in testing, the second column shows the features used as input to the neural network, and the other columns indicate the percentage of syllables having predicted durations

Table 3
Percentages of syllables predicted within different deviations for different input features for the languages Hindi, Telugu and Tamil

Language (# Syllables)	Features	% Predicted syllables within deviation		
		10%	25%	50%
Hindi (11,222)	All	29	68	84
	Positional	26	63	81
	Contextual	27	65	82
Telugu (17,630)	All	29	66	86
	Positional	25	60	82
	Contextual	26	62	83
Tamil (21,493)	All	34	75	96
	Positional	29	70	91
	Contextual	31	72	93

within the specified deviation with respect to their actual durations. Compared to the positional and contextual factors separately, the features using all the factors seem to yield a good prediction of the duration.

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 the actual and predicted duration values. These results are given in Table 4. The definitions of average prediction error (μ), standard deviation (σ), and linear correlation coefficient ($\gamma_{X,Y}$) are given below

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

$$\sigma = \sqrt{\frac{\Sigma_i d_i^2}{N}},\ d_i = e_i - \mu,\ \text{and}\ e_i = x_i - y_i,$$

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

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

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

The accuracy of prediction (Fig. 4) of the duration models is not uniform for the entire duration range. Better prediction is observed around the mean of the distribution of the original training data. Syllables with longer durations than the mean tend to be underestimated, while those with shorter durations than the mean seem to be overestimated. For improving the accuracy of prediction, the predicted values are further modified by imposing piecewise linear transformation (Bellegarda et al., 2001; Silverman and Bellegarda, 1999). The prediction performance is observed to have improved slightly by using the transformation.

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 durations of the adjacent syllables are considered as duration constraints, and the intonation patterns (average pitch values) of the present syllable and its adjacent syllables are considered as intonation constraints, for estimating the duration of the

Table 4
Performance of neural network models using objective measures (μ, σ and γ)

Language	Average prediction error in ms (μ)	SD in ms (σ)	Correlation coefficient (γ)
Hindi	32	26	0.75
Telugu	29	23	0.78
Tamil	26	22	0.82

Table 5
Performance of the neural network models using different constraints

Features	Language (# Syllables)	% Predicted syllables within deviation			Objective measures		
		10%	25%	50%	μ (ms)	σ (ms)	γ
Linguistic	Hindi (1084)	25	57	85	34	27	0.71
	Telugu (1107)	24	56	83	35	28	0.74
	Tamil (949)	26	60	88	30	25	0.73
Linguistic and duration	Hindi (1084)	25	59	86	34	27	0.72
	Telugu (1107)	24	58	85	34	27	0.75
	Tamil (949)	27	62	88	30	25	0.74
Linguistic and intonation	Hindi (1084)	27	60	88	33	27	0.73
	Telugu (1107)	25	58	86	34	27	0.76
	Tamil (949)	28	64	90	29	25	0.75
Linguistic, duration and intonation	Hindi (1084)	29	63	90	32	26	0.73
	Telugu (1107)	27	60	88	33	26	0.77
	Tamil (949)	30	65	92	29	24	0.77

syllable. For studying the influence of duration and intonation constraints in predicting the durations 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 5. The first column indicates the features used for developing the neural network models, the second column shows the number of syllables specific to the particular language used in testing, the columns 3–5 indicate the percentage of syllables having predicted durations within the specified deviation with respect to their actual durations, and the columns 6–8 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 25% or 50% 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 5).

6. Duration modeling using two-stage approach

Since a single feedforward neural network model is used for predicting the durations of syllables for the entire range of 40–300 ms, the accuracy of prediction (Fig. 4) is biased towards the mean of the distribution of the training data. This leads to poor prediction for long and short duration syllables which lie at the tail portions of the distributions. This problem can be alleviated to some extent by using multiple models, one for each limited range of duration. Here the number of models (number of intervals) is not crucial. The number of models and the range of each interval can be arrived at experimentally. But this requires preprocessing of the data to categorize syllables into different groups based on duration.

For implementing this concept, a two-stage duration model is proposed. The first stage consists of a syllable classifier which groups the syllables based on their duration. The second stage is a function approximator for modeling the syllable duration, which consists of specific models for the given duration interval. The block diagram of the proposed two-stage duration model is shown in Fig. 6. The performance of the proposed model depends on the performance of the syllable classifier (1st stage), since the error at the 1st stage will route the syllable features to the unintended model for prediction. The duration intervals are chosen based on the distributions of the durations data for different languages (see Fig. 1). From the figure, it is evident that the distributions of durations of Telugu and Tamil are similar within central region approximately in the range 100–150 ms. The prediction performance plots in Fig. 4 for Telugu and Tamil are also similar. Hence, the duration intervals for Telugu and Tamil are chosen as 40–100 ms, 100–150 ms and 150–300 ms. For Hindi, the central region of the distribution of durations is higher than for Telugu and Tamil, and it is approximately in the range 120–170 ms

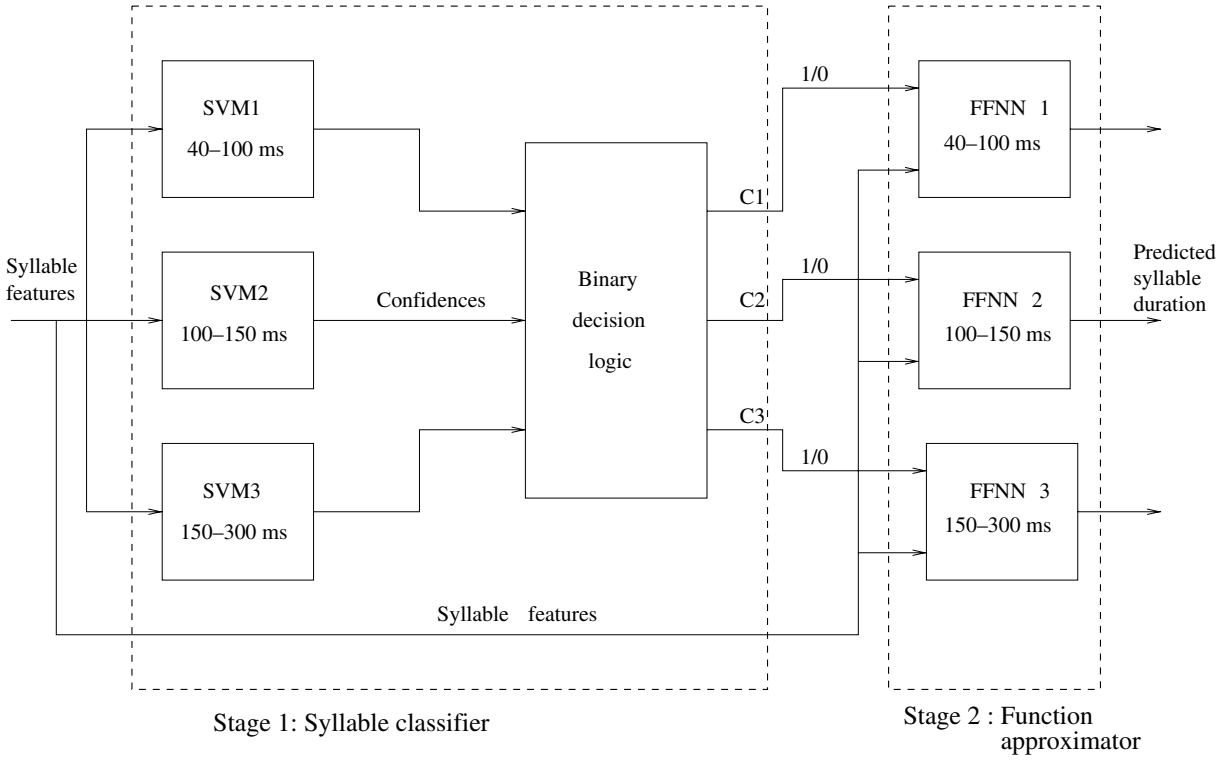

Fig. 6. Two-stage duration model.

(see Fig. 1). The prediction performance for Hindi is also different from that for Telugu and Tamil, and hence the duration intervals for Hindi are chosen as 40–120 ms, 120–170 ms and 170–300 ms.

Support Vector Machine (SVM) models are used for classification of syllables into three intervals (Haykin, 1999; Burges, 1998). The SVM uses discriminative learning for classification tasks. In the training phase, the SVM model finds a global minimum, and it tries to maintain the training error close to zero. The SVM can provide good generalization on pattern classification problems, by implementing the concept of structural risk minimization. The block diagram of the syllable classifier using SVM models is shown in stage 1 of Fig. 6. The decision logic provides three outputs (corresponding to the number of duration intervals), which are connected to the corresponding models in the following stage (function approximator). For each syllable the decision logic activates only the duration model corresponding to the maximum confidence. The selected duration model now predicts the duration of the syllable in that limited interval. The performance of classification using SVM models is shown in Table 6.

For modeling the syllable duration, the syllable features are presented to all the models of the syllable classifier. The decision logic in the syllable classifier routes the syllable features to one of the duration models present in the second stage (function approximator) for predicting the duration of the syllable in a specific limited duration interval. The performance of the proposed two-stage duration model is given in Table 7. The numbers within brackets indicate the performance of the single FFNN model. The first column indicates the syllables from different languages used for testing the models, the columns 2–4 indicate the percentage of

Table 6
Classification performance using SVM models for the languages Hindi, Telugu and Tamil

Language	% Syllables correctly classified
Hindi	81.92
Telugu	80.17
Tamil	83.26

Table 7
Performance of the two-stage model (numbers outside the brackets)

Language	% Predicted syllables within deviation			Objective measures		
	10%	25%	50%	μ (ms)	σ (ms)	γ
Hindi	37(29)	80(68)	96(84)	25(32)	20(26)	0.82(0.75)
Telugu	39(29)	83(66)	96(86)	23(29)	23(23)	0.82(0.78)
Tamil	44(34)	86(75)	97(96)	20(26)	20(22)	0.85(0.82)

The numbers within brackets indicate the performance of the single FFNN model.

syllables having predicted durations within the specified deviation with respect to their actual durations, and the columns 5–7 indicate the objective measures. The number of syllables used for testing the model in each language is same as those used in Section 5 (Table 3).

The results show that the prediction accuracy has improved with the two-stage model compared to the single model for the entire duration range. For comparing the performance of the two-stage model with the single FFNN model, the Tamil broadcast news data was chosen. The prediction performance of the single FFNN model and the two-stage model are shown in Fig. 7. The performance curves in the figure show that the syllables having duration around the mean of the distribution are estimated better in both the models. On the other hand the short and long duration syllables are poorly predicted in the case of the single FFNN model. The prediction accuracy of these extreme (long and short) syllables has improved in the two-stage model, because of the use of specific models for each of the duration interval, even though the performance of the classification stage is only about 80% (shown in Table 6).

The prediction performance of the proposed models are compared with the results using the (Classification and Regression Trees) CART models. The performance of the CART models is given in Table 8. The first column indicates the syllables from different languages used for testing the models, the columns 2–4 indicate the percentage of syllables having predicted durations within the specified deviation with respect to their actual durations, and the columns 5–7 indicate the objective measures. The number of syllables used for testing the model in each language is same as those used in Section 5 (Table 3). The performance of the FFNN models (shown within brackets in Table 8) is comparable to CART models, whereas the two-stage models (Table 7) seem to perform better than the CART models.

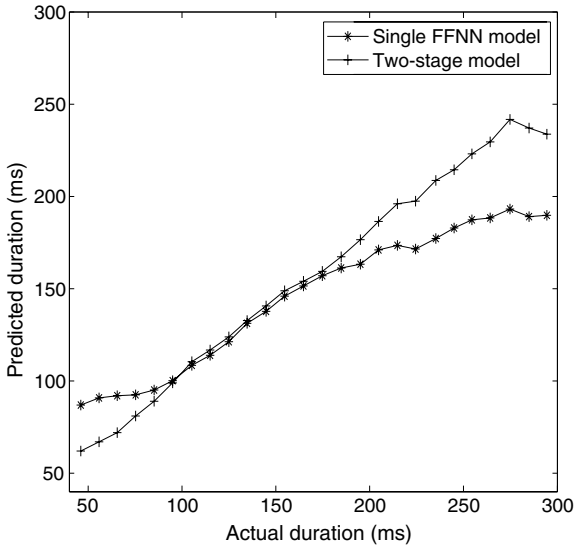

Fig. 7. Prediction performance of single FFNN model and two-stage model.

Table 8
Performance of the CART models (numbers outside the brackets)

Language	% Predicted syllables within deviation			Objective measures		
	10%	25%	50%	μ (ms)	σ (ms)	γ
Hindi	31(29)	67(68)	92(84)	32(32)	26(26)	0.76(0.75)
Telugu	30(29)	64(66)	88(86)	29(29)	24(23)	0.78(0.78)
Tamil	33(34)	71(75)	93(96)	25(26)	21(22)	0.81(0.82)

The numbers within brackets indicate the performance of FFNN models.

7. Summary and conclusions

Feedforward neural network (FFNN) models were proposed for predicting the durations of syllables. The linguistic context and production constraints associated with the syllable are represented with positional, contextual and phonological features. Suitable neural network structures were arrived at empirically. The models were evaluated by computing the average prediction error (μ), standard deviation (σ) and correlation coefficient (γ) between the predicted and actual durations of the syllables. The accuracy in prediction was also analyzed in terms of percentages of syllables predicted within different deviations with respect to their actual durations. The positional and contextual features were analyzed separately, and also in combination. The performance of the model has improved by using the features from all the factors together. The effect of duration and intonation constraints was examined by modeling the durations of the syllables with these constraints as input to the FFNN models. A two-stage duration model was proposed to alleviate the problem of poor prediction due to a single FFNN model. The performance of the two-stage model was improved by appropriate syllable classification model and the selection criterion of duration intervals. The performance of the neural network models was compared with the results obtained by CART models. The performance 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 also improve the performance. The accuracy of labeling, diversity of data in the database, and fine tuning of the neural network parameters, all of these factors may also play a role in improving the prediction of the syllable duration. The proposed duration models can be used in applications such as speech recognition, speaker recognition, language identification and text-to-speech synthesis.

References

Barbosa, P.A., Bailly, G. 1992. Generating segmental duration by p-centers. In: Proceedings of the Fourth Workshop on Rhythm Perception and Production, Bourges, France, June, pp. 163–168.

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

Bartkova, K., Sorin, C., 1987. A model of segmental duration for speech synthesis in French. Speech Communication (6), 245–260.

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 (Jan), 52–66.

Black, A.W., Taylor, P., Caley, R., 2000. The festival speech synthesis system: System documentation. The Centre for Speech Technology Research (CSTR), University of Edinburgh, 1.4.0 edition. Available from: http://www.cstr.ed.ac.uk/projects/festival/manual/festival_toc.html.

Burges, C.J.C., 1998. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2 (2), 121–167.

Campbell, W.N., 1990. Analog i/o nets for syllable timing. Speech Communication 9 (February), 57–61.

Campbell, W.N., 1992. Syllable based segment duration. In: Bailly, G., Benoit, C., Sawallis, T.R. (Eds.), Talking Machines: Theories, Models and Designs. Elsevier, Amsterdam, pp. 211–224.

Campbell, W.N., 1993. Predicting segmental durations for accommodation within a syllable-level timing framework. In: Proceedings of the European Conference Speech Communication and Technology, vol. 2, Berlin, Germany, September, pp. 1081–1084.

Campbell, W.N., Isard, S.D., 1991. Segment durations in a syllable frame. Journal of Phonetics: Special issue on speech synthesis 19, 37–47.

Chen, S.H., Lai, W.H., Wang, Y.R., 2003. A new duration modeling approach for Mandarin speech. IEEE Transactions on Speech and Audio Processing 11 (July), 308–320.

Chopde, A. Itrans Indian language transliteration package version 5.2 source. Available from: http://www.aczone.con/itrans/.

Chung, H., 2002a. Duration models and the perceptual evaluation of spoken Korean. In: Proceedings of Speech Prosody, Aix-en-Provence, France, pp. 219–222.

Chung, H., 2002b. Perceptual evaluation of duration models in spoken Korean. The Korean Journal of Speech Sciences 9, 207–215.
Cordoba, R., Vallejo, J.A., Montero, J.M., Gutierrezarriola, J., Lopez, M.A., Pardo, J.M. 1999. Automatic modeling of duration in a Spanish text-to-speech system using neural networks. In: Proceedings of the European Conference on Speech Communication and Technology, September, Budapest, Hungary.
Goubanova, O, Taylor, P. 2000. Using bayesian belief networks for modeling duration in text-to-speech systems. In: Proceedings of the International Conference on Spoken Language Processing, vol. 2, Beijing, China, October 2000, pp. 427–431.
Haykin, S., 1999. Neural Networks: A Comprehensive Foundation. Pearson Education Asia, Inc., New Delhi, India.
Hifny, Y, Rashwan, M. 2002. Duration modeling of Arabic text-to-speech synthesis. In: Proceedings of the International Conference on Spoken Language Processing, Denver, CO, USA, September, pp. 1773–1776.
Huang, X., Acero, A., Hon, H.W., 2001. Spoken Language Processing. Prentice-Hall, Inc., New York, NJ, USA.
Khan, A.N., Gangashetty, S.V., Yegnanarayana, B., 2003. Syllabic properties of three Indian languages: Implications for speech recognition and language identification. In: International Conference on Natural Language Processing, Mysore, India, December, pp. 125–134.
Klatt, D.H., 1976. Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. Journal of Acoustic Society of America 59, 1209–1221.
Kohler, K.J., 1988. Zeistrukturierung in der Sprachsynthese. ITG-Tagung Digitalc Sprachverarbeitung (6), 165–170.
Krishna, N.S., Murthy, H.A., 2004. Duration modeling of Indian languages Hindi and Telugu. In: 5th ISCA Speech Synthesis Workshop, Pittsburgh, USA, May, pp. 197–202.
Kumar, K.K., 2002. Duration and intonation knowledge for text-to-speech conversion system for Telugu and Hindi, Master's thesis, Department of Computer Science and Engineering, Indian Institute of Technology Madras, India, May.
Mixdorff, H., 2002. An integrated approach to modeling German prosody. PhD thesis, Technical University, Dresden, Germany, July.
Mixdorff, H., Jokisch, O. 2001. Building an integrated prosodic model of German. In: Proceedings of the European Conference on Speech Communication and Technology, vol. 2, Aalborg, Denmark, September, pp. 947–950.
Riley, M., 1992. Tree-based modeling of segmental durations. Talking Machines: Theories, Models and Designs, 265–273.
Santen, J.P.H.V., 1994. Assignment of segment duration in text-to-speech synthesis. Computer Speech and Language 8 (April), 95–128.
Sayli, O, 2002. Duration analysis and modeling for Turkish text-to-speech synthesis, Master's thesis, Department of Electrical and Electronics Engineering, Bogaziei University, 2002.
Silverman, K.E.A., Bellegarda, J.R. 1999. Using a sigmoid transformation for improved modeling of phoneme duration. In: Proceedings of the IEEE International Conference on Acoustic Speech, Signal Processing, Phoenix, AZ, USA, March 1999, pp. 385–388.
Smith, C.L., 2002. Modeling durational variability in reading aloud a connected text. In: Proceedings of the International Conference on Spoken Language Processing, Denver, CO, USA, September, pp. 1769–1772.
Sonntag, G.P., Portele, T., Heuft, B. 1997. Prosody generation with a neural network: Weighing the importance of input parameters. In: Proceedings of the IEEE International Conference on Acoustic, Speech, Signal Processing, Munich, Germany, April, pp. 931–934.
Sontag, E.D., 1992. Feedback stabilization using two hidden layer nets. IEEE Transactions on Neural Networks 3 (November), 981–990.
Sproat, R. (Ed.), 1998. Multilingual Text-to-Speech Synthesis: The Bell Labs Approach. Kluwer Academic Publishers, Dordrecht, The Netherlands.
Teixeira, J.P., Freitas, D. 2003. Segmental durations predicted with a neural network. In: Proceedings of the European Conference on Speech Communication and Technology, Geneva, Switzerland, September, pp. 169–172.
Yegnanarayana, B., 1999. Artificial Neural Networks. Prentice-Hall, New Delhi, India.
Yegnanarayana, B., Murthy, H.A., Sundar, R., Ramachandran, V.R., Kumar, A.S.M., Alwar, N., Rajendran, S., 1990. Development of text-to-speech system for Indian languages. In: Proceedings of the International Conference on Knowledge Based Computer Systems, Pune, India, December, pp. 467–476.