MODELING SYLLABLE DURATION IN INDIAN LANGUAGES USING NEURAL NETWORKS

K. Sreenivasa Rao and B. Yegnanarayana

Speech and Vision Laboratory
Department of Computer Science and Engineering
Indian Institute of Technology Madras, Chennai - 600 036, INDIA.
E-mail: {ksr,yegna}@cs.iitm.ernet.in

ABSTRACT

In this paper we propose a neural network model for predicting the syllable duration in Indian languages. A four layer feedforward neural network trained with a backpropagation algorithm is used for modeling the syllable duration. Analysis is performed on broadcast news data in the languages Hindi, Telugu and Tamil, in order to predict the duration of syllables in these languages using a neural network model. The input to the neural network consists of a set of phonological, positional and contextual features extracted from the text. About 88% of the syllable durations are predicted within 25% of the actual duration. The relative importance of the positional and contextual features are examined separately.

1. INTRODUCTION

Modeling the syllable durations by analyzing large databases manually is a tedious process. An efficient way to model the syllable durations is by using features of neural networks. Duration models help to improve the quality of Text-to-Speech (TTS) systems. In most of the TTS systems the durations of syllables are estimated using a set of rules derived manually from a limited database.

Mapping a string of phonemes or syllables and the linguistic structures (positional, contextual and phonological information) to the continuous prosodic parameters is a complex nonlinear task [1]. This mapping has traditionally been done by a set of sequentially ordered rules derived based on introspective capabilities and expertise of the individual research workers. Moreover, a set of rules cannot describe the nonlinear relations beyond certain point. The rules are usually as general as possible, and exceptions to them tend to complicate the rule-set.

Neural networks are known for their ability to generalize and capture the functional relationship between the input-output pattern pairs [2] [3]. Neural networks have the ability to predict, after an appropriate learning phase, even patterns not presented before. For predicting the syllable duration, a feedforward neural network model is proposed [4].

The duration models can be grouped into rule-based approaches and statistical approaches. Rule-based approach is based on the results of experimental studies on segment durations. The most important rule-based duration model is the one developed by Klatt [5]. Another rule-based duration model was developed for a TTS system for the Indian language Hindi at IIT Madras [6].

Statistical models became popular with the availability of large phonetically labeled databases. The statistical approaches can be divided into parametric and nonparametric regression models [7]. In parametric regression model the structure of processing the input parameters is determined a priori. Examples of a parametric regression model are sums-of-products model, generalized linear models and additive and multiplicative models. Nonparametric regression models are developed by unsupervised training, and the model structure is determined automatically. Examples of nonparametric regression models are neural network based approaches and Classification and Regression Trees (CART) based approaches [8]. Campbell used the neural network models for estimating syllable duration and computed the segment durations from syllable duration using the concept of Z-score [4]. Barbosa and Bailly also presented a neural network model for French [9]. Neural network models for predicting syllable durations were also presented for other languages.

This paper presents the duration analysis of broadcast news data for three Indian languages (Hindi, Telugu and Tamil) using syllables as basic units. The syllable is a natural and convenient unit for speech in Indian languages. In Indian scripts syllables are generally used as characters. The syllable-like units are thus more relevant from both speech production and perception point of view. The syllable also capture some coarticulation effects. A character in an Indian language scripts is close to a syllable, and is typically of the following form: V, CV, CCV, CCVC and CVCC, where C is a consonant and V is a vowel. All the Indian languages

have a common phonetic base, and the phoneset consists of about 35 consonants and 18 vowels.

The paper is organized as follows: Section 2 discusses the basic factors that affect the syllable duration. The database for the proposed duration analysis is described in section 3. Section 4 discusses the features used as input to the neural network for capturing the information about syllable duration. Section 5 gives the details of the neural network model. Evaluation of the model is presented in section 6. A summary of the paper is given in the final section along with a discussion on some issues to be addressed further.

2. FACTORS AFFECTING THE SYLLABLE DURATION

Acoustic analysis and synthesis experiments have shown that duration and intonation patterns are the two most important prosodic features responsible for the quality of synthesized speech [10]. A good prosodic model should capture the durational and intonational properties of natural-speech. In continuous speech large number of factors affect the durations of the basic units. They are broadly classified into phonological, positional and contextual factors. The vowel is considered as a nucleus of a syllable, and consonants may present on either side of the vowel. The syllable duration may be influenced by the vowel position, category of the vowel present in the syllable and the type of the consonants associated with the vowel. Positional factors affect the durations of the basic units according to the position of the unit in the text. Different positions that affect the duration of the basic unit are: Word final position, phrase boundary, sentence ending position and word initial position. The other factors that affect the durations of the basic units depend on the contexts in which the units occur. Contextual factors include the influence of the preceding and following units on the present unit. Different manners and places of articulation of the units in the preceding and following positions also affect the duration of present unit to different extents. Apart from the factors mentioned above, gender of the speaker, psychological state of the speaker (happy, anger, fear etc.), age, relative novelty in the words and words with relatively higher number of syllables also affect the duration. These effects are difficult to describe. Also most of these effects occur relatively less frequently.

3. SPEECH DATABASE

The database consists of four Hindi, five Telugu and five Tamil news bulletins. In each language these news bulletins are read by a male speaker. Total durations of speech in Hindi, Telugu and Tamil are around 45 minutes, 75 minutes and 65 minutes, respectively. The speech signal was sampled at 16kHz sampling frequency and encoded as 16 bit integers. The speech utterances are manually transcribed into text using common Transliteration code (ITRANS) for Indian languages. The speech utterances are segmented and labeled manually into syllable-like units. Each bulletin is organized in the form of syllables, words, orthographic text representations of the utterances and the utterances in wav format. Each of the syllable and word files contains the text transcriptions and timing information in the form of sample numbers. The total database consists of 9415 syllables in Hindi, 19719 syllables in Telugu and 16315 syllables in Tamil.

4. FEATURES FOR DEVELOPING NEURAL NETWORK MODEL

The features considered for modeling syllable duration are based on positional, contextual and phonological information. Features representing positional information are further classified based on syllable position in a word and phrase.

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

Position in 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. They indicate the location of the syllable in a word from the starting and terminating positions. Another feature indicates the number of syllables in a word.

Syllable identity: Syllable constitutes combination of segments of consonants and vowels. In our analysis syllables with more than four segments are ignored. Each segment of the syllable is encoded separately, so that each syllable is represented by four features indicating its identity.

Context of a syllable: Syllable duration may be influenced by its adjacent syllables. Hence for modeling the duration of a syllable its context information is represented by the previous syllable and following syllable. Each of these syllables is represented by a four dimensional feature vector, representing the identity of the syllable.

Syllable nucleus: Another important feature consists of vowel position in a syllable, and the number of segments before and after the vowel in a syllable. This feature is rep-

resented with three independent codes specifying three distinct features.

The list of features and the number of nodes in a neural network needed to represent the features are given in Table 1.

Table 1. 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	1. Position of syllable from beginning of the phrase 2. Position of syllable from end of the phrase 3. Number of syllables in a phrase	3
Syllable position in the word	1. Position of syllable from beginning of the word 2. Position of syllable from end of the word 3. Number of syllables in a word	3
Syllable identity	Segments of syllable	4
Context of syllable	1. Previous syllable 2. Following syllable	4 4
Syllable nucleus	1. Position of the nucleus 2. Number of segments before nucleus 3. Number of segments after nucleus	3

5. NEURAL NETWORK STRUCTURE

Feedforward neural networks (FFNN) are used to capture the relationship between input and output vectors [3] [2]. For modeling syllable durations, we employed a four layer feed forward neural network whose general structure is shown in Fig. 1. The first layer is the input layer which consists of linear elements. The second and third layers are hidden layers. The second layer (first hidden layer) of the network has more units than the input layer, and it can be interpreted as capturing some local features in 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 [3] [2]. The fourth layer is the output layer having one unit representing the syllable duration. Activation functions at the input layer are linear, and at the hidden layers they are nonlinear. Generalization is influenced by three factors: (1) The size of the training set, (2) the architecture of the neural network, and (3) the complexity of the problem. We have no control over the first and last factors. Hence for better generalization, several network structures are experimentally verified. The optimum structure arrived for the study in our system is 21*L* 40*N* 10*N* 1*N*. For analyzing the influence of positional and contextual factors on syllable duration, the network structures used are 10*L* 20*N* 5*N* 1*N* and 12*L* 24*N* 6*N* 1*N*, respectively. In the network structure *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. All the input and output parameters were transformed to fit in [-1 to +1] range before applying to the neural network. The standard backpropagation learning algorithm is used for adjusting the weights of the network to minimize the mean squared error for each syllable duration [3].

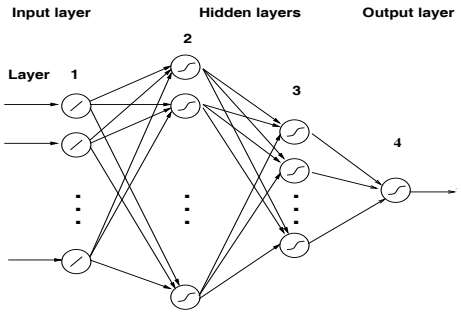

Fig. 1. Four layer feedforward neural Network

6. EVALUATION OF THE NEURAL NETWORK MODEL

A separate model is prepared for each of the three languages. For Hindi 7000 syllables are used for training the network and 2415 syllables are used for testing. For Telugu 15000 syllables are used for training and 4719 are used for testing. For Tamil 12000 syllables are used for training and 4315 are used for testing. For each syllable the phonological, positional and contextual features are extracted and a 21 dimension input vector is formed. Duration of the each syllable is obtained from the timing information available in the database. Usually syllable durations are ranging from 50-500 msec. Before applying to neural network model, both input and output parameters are normalized in the range of -1 to +1. The extracted input vectors are given as input and the corresponding syllable durations are given as output to the FFNN model, and the network is trained for 100 epochs. The duration model is evaluated with the syllables in the test set. For each syllable in the test set, predict the duration using FFNN by giving the input vector of syllable as input to the neural network. The deviation of predicted duration from the actual duration is estimated. In order to evaluate the prediction accuracy, between predicted values and actual duration values, standard deviation of the difference is computed. The standard deviation of the differ-

ence between predicted and actual durations is found to be 34.7, 29.3 and 26.2 msec for Hindi, Telugu and Tamil data respectively. For analyzing the influence of positional and contextual factors on syllable duration, features associated with syllable position and syllable context are extracted separately. Syllable durations are predicted and the deviations from the original duration are estimated using positional parameters and contextual parameters separately with the corresponding neural network models mentioned in Section 5. The number of syllables with various deviations from actual syllable durations are presented in Table 2. In Table 2, the first column indicates the number of syllables specific to the particular language used for testing, the second column shows the parameters used as input to the neural network, and the other columns indicate the number of syllables having predicted duration with in the specified deviation with respect to actual syllable duration. Compared to individual factors, features using all the factors seems to yield a good duration model.

Table 2. The table shows the number of syllables having predicted duration within the specified deviation from actual syllable duration for different input features from each of the three languages Hindi, Telugu and Tamil.

Language # Syllables	Features	# Predicted syllables with deviation			
		$< 10\%$	10-25%	25-50%	$> 50\%$
Hindi (2415)	All parameters	1321	816	176	102
	Positional parameters	1187	860	206	162
	Contextual parameters	1206	858	210	141
Telugu (4719)	All parameters	2576	1723	285	135
	Positional parameters	2351	1752	381	235
	Contextual parameters	2216	1884	387	232
Tamil (4315)	All parameters	2841	1182	238	54
	Positional parameters	2505	1379	326	105
	Contextual parameters	2606	1249	335	125

7. SUMMARY AND CONCLUSIONS

A four layer feedforward neural network trained with standard backpropagation algorithm was used for predicting syllable durations. Phonological, positional and contextual parameters were extracted from the syllables of each of the three languages. Suitable neural network structures were derived by experimental verification. The model is objectively evaluated by computing the standard deviation for the difference between predicted and actual syllable durations. Phonological, positional and contextual features were analyzed separately, and also in combination. Performance of the model is improved by considering the features from all the three factors together. The performance can be further improved by including the accent and prominence of the syllable in the feature vector. The accuracy of labeling, diversity of data in the database, and fine tuning of neural network parameters, all of these may also play a role in improving the performance.

8. REFERENCES

[1] M. Vainio and T. Altosaar, "Modeling the microprosody of pitch and loudness for speech synthesis with neural networks," in *Proc. Int. Conf. Spoken Language Processing*, (Sidney, Australia), Sept. 1998.

[2] S. Haykin, *Neural Networks: A comprehensive foundation*. New Delhi, India: Pearson Education Aisa, Inc., 1999.

[3] B. Yegnanarayana, *Artificial Neural Networks*. New Delhi, India: Printice-Hall, 1999.

[4] W. N. Campbell, "Analog i/o nets for syllable timing," *Speech Communication*, vol. 9, pp. 57–61, Feb. 1990.

[5] D. H. Klatt, "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence," *Journal of Acoustic Society of America*, vol. 59, pp. 1209–1221, 1976.

[6] B. Yegnanarayana, H. A. Murthy, R. Sundar, V. R. Ramachandran, A. S. M. Kumar, N. Alwar, and S. Rajendran, "Develoment of text-to-speech system for indian languages," in *Proc. Int. Conf. Knowledge Based Computer Systems*, (Pune, India), pp. 467–476, Dec. 1990.

[7] H. Mixdorff and O. Jokisch, "Building an integraated prosodic model of german," in *Proc. European Conf. Speech communication and Technology*, vol. 2, (Aalborg, Denmark), pp. 947–950, Sept. 2001.

[8] M. Riley, "Tree-based modeling of segmental durations," *Talking Machines: Theories, Models and Designs*, pp. 265–273, 1992.

[9] P. A. Barbosa and G. Bailly, "generation of pauses within the z-score model," *Progress in Speech Synthesis*, pp. 365–381, 1997.

[10] X. Huang, A. Acero, and H. W. Hon, *Spoken Language Processing*. Prentice-Hall, Inc., 2001.