Two-Stage Duration Model for Indian Languages Using Neural Networks

K. Sreenivasa Rao, S.R. Mahadeva Prasanna, and B. Yegnanarayana

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

Abstract. In this paper we propose a two-stage duration model using neural networks for predicting the duration of syllables in Indian languages. The proposed model consists of three feedforward neural networks for predicting the duration of syllable in specific intervals and a syllable classifier, which has to predict the probability that a given syllable falls into an interval. Autoassociative neural network models and support vector machines are explored for syllable classification. Syllable duration prediction and analysis is performed on broadcast news data in Hindi, Telugu and Tamil. The input to the neural network consists of a set of phonological, positional and contextual features extracted from the text. From the studies it is found that about 80% of the syllable durations are predicted within a deviation of 25%. The performance of the duration model is evaluated using objective measures such as mean absolute error (μ), standard deviation (σ) and correlation coefficient (γ).

1 Introduction

Modeling syllable durations by analyzing large databases manually is a tedious process. An efficient way to model 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 durations of the 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 [1998v]. 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. Neural networks are known for their ability to generalize and capture the functional relationship between the input-output pattern pairs [1999]. Neural networks have the ability to predict, after an appropriate learning phase, even patterns they have never seen before. For predicting the syllable duration, Feedforward Neural Network (FFNN) models are proposed [1990]. The existing neural network based duration models consists of single neural network for predicting the durations of

N.R. Pal et al. (Eds.): ICONIP 2004, LNCS 3316, pp. 1179–1185, 2004.

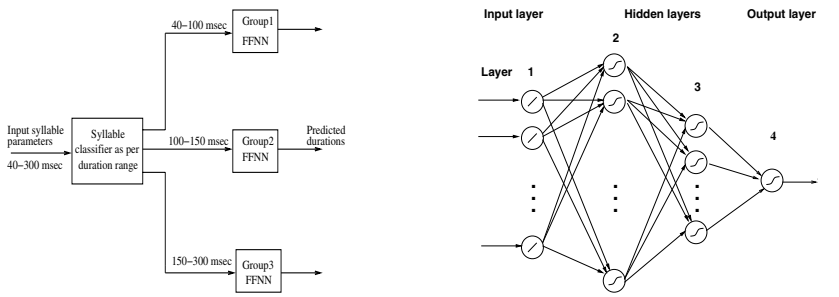
all sound units. With this the sound units around the mean of the distribution will be predicted better, and for other (long and short) sound units prediction will be poor [1998v][1990].

This paper proposes a two-stage model for predicting the syllable duration. The first stage consists of syllable classifier, which classify the syllable into one of the groups based on duration range. The second stage constitutes three neural network models, which are meant for predicting the duration of syllable in the specific intervals. The paper presents the duration analysis of broadcast news data for three Indian languages (Hindi, Telugu and Tamil) using syllables as basic units.

The paper is organized as follows: Section 2 describes the proposed two-stage model and the performance of duration models intended for specific intervals. The first stage in the proposed duration model is a syllable classifier, which is discussed in Section 3. Evaluation of the proposed duration model is presented in section 4. Final section discusses about the issues to be addressed further.

2 Two-Stage Duration Model

The block diagram of the proposed two-stage duration model is shown in Fig. 1(a). The first stage consists of syllable classifier which groups the syllables based on their duration. The second stage is for modeling the syllable duration which consists of specific models for the given duration interval. In the database, most of the syllable durations are varying from 40-300 ms. We have chosen three successive duration intervals (40-100, 100-150 and 150-300 ms) such that they will cover the entire syllable duration range.



(a) (b)

Fig. 1. (a) Two-stage duration model (b) Four layer Feedforward neural network.

2.1 Neural Network Structure

For modeling syllable durations, we employed a four layer feedforward neural network whose general structure is shown in Fig. 1(b). The first layer is the input layer which consists of linear elements. The second and third layers are hidden layers, and they can be interpreted as capturing some local and global features in the input space [1999]. The fourth layer is the output layer having one unit representing the syllable duration. For better generalization, several

network structures are experimentally verified. The optimum structure arrived is $22L\,44N\,11N\,1N$, where L denotes a linear unit, N denotes a nonlinear unit and 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 normalized to the range [-1 to +1] 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.

2.2 Speech Database

The database consists of 15 Hindi, 20 Telugu and 25 Tamil 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 around 3.25, 4.5 and 4 hours, respectively. The speech utterances were 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 timing information in the form of sample numbers. The total database consists of 46222 syllables in Hindi, 81630 syllables in Telugu and 69811 syllables in Tamil.

2.3 Features for Developing Neural Network Model

The features considered for modeling syllable duration are based on positional, contextual and phonological information. 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	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
Gender identity	Gender	1

2.4 Performance of the Models in Specific Duration Range

Initially syllables of each of the three languages (Hindi, Telugu and Tamil) are manually classified into 3 groups (40-100, 100-150 and 150-300 ms) based on the duration. For each language three FFNN models are used for predicting the syllable durations in the specific duration intervals. For each syllable the phonological, positional and contextual features are extracted and a 22 dimension input vector is formed. The extracted input vectors are given as input and the corresponding syllable durations are given as output to the FFNN models, and the networks are trained for 500 epochs. The duration models are evaluated with the corresponding syllables in the test set. The deviation of predicted duration from the actual duration is estimated. The number of syllables with various deviations from actual syllable durations are presented in Table 2. In order to objectively evaluate the prediction accuracy, between predicted values and actual duration values, standard deviation of the difference (σ) and linear correlation coefficient (γ) were computed. The standard deviation of the difference between predicted and actual durations is found to be about 13.2 ms and the correlation between predicted and actual durations is found to be 0.91 across the languages in specific duration intervals.

Table 2. Number of syllables having predicted duration within the specified deviation from actual syllable duration for different duration intervals from each of the three languages Hindi, Telugu and Tamil.

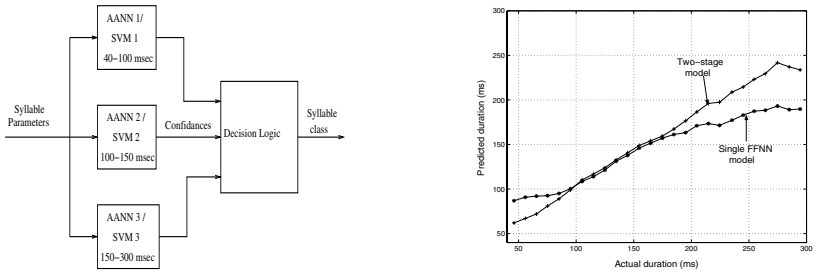
Language	Duration range	Training # syls	Testing # syls	# Syllables within deviation			
				< 10%	10-25%	25-50%	> 50%
Hindi	40-100	10000	3057	1611	1155	244	47
	100-150	13000	4112	2462	1641	9	-
	150-300	12000	4053	1989	1875	189	-
Telugu	40-100	19670	5000	1802	2304	692	202
	100-150	24011	6000	3324	2556	120	-
	150-300	20949	6000	2718	2656	622	4
Tamil	40-100	15000	4260	1570	2292	313	85
	100-150	20000	7156	4177	2834	145	-
	150-300	18000	5395	2834	2242	319	-

3 Syllable Classification

In the proposed two-stage duration model, first stage consists of a syllable classifier, which divides the syllables into three groups based on their duration. In this paper Autoassociative Neural Network (AANN) models and Support Vector Machine (SVM) models are explored for syllable classification. The block diagram of syllable classification model is shown in Fig. 2(a).

3.1 AANN Models

Autoassociative neural network models are feedforward neural networks performing an identity mapping of the input space, and are used to capture the



(a) (b)

Fig. 2. (a) Syllable classification model (b) Prediction performance of two-stage and single FFNN models.

distribution of the input data [1999]. The optimum structures arrived for the study in Hindi, Telugu and Tamil are $22L\ 30N\ 14N\ 30N\ 22L$, $22L\ 30N\ 16N\ 30N\ 22L$ and $22L\ 30N\ 10N\ 30N\ 22L$, respectively. For each language three AANN models are prepared for the duration intervals 40-100, 100-150 and 150-300 ms. For classification task, the syllable parameters are given to each of the model. The output of each model is compared with the input to compute the square error. The error (e) is transformed into a confidence (c) value by using the equation $c = exp(-e)$. The confidence values are given to a decision logic, where the highest confidence value among the models is used for classification. The classification performance of the AANN models are shown in Table 3.

3.2 SVM Models

Support vector machines provide an alternate approach to the pattern classification problems. SVMs are initially designed for two-class pattern classification. Multiclass (n-class) pattern classification problems can be solved using a combination of binary support vector machines. Here we need to classify the syllables into three groups based on duration. An SVM is constructed for each class by discriminating that class against the remaining two classes. The classification system consists of three SVMs. The set of training examples $\{\{(\mathbf{x}_i, k)\}_{i=1}^{N_k}\}_{k=1}^{n}$ consists of N_k number of examples belonging to k^{th} class, where the class label $k \in \{1, 2, \ldots, n\}$. The SVM for the class k is constructed using a set of training examples and their desired outputs, $\{\{(\mathbf{x}_i, y_i)\}_{i=1}^{N_k}\}_{k=1}^{n}$. The desired output y_i for a training example $\mathbf{x}_i$ is defined as follows:

$$y_i = \begin{cases} +1 & : \quad If \quad \mathbf{x}_i \in k \\ -1 & : \quad otherwise \end{cases}$$

The examples with $y_i = +1$ are called positive examples, and those with $y_i = -1$ are called negative examples. An optimal hyperplane is constructed to separate positive examples from negative examples. The separating hyperplane (margin) is chosen in such a way as to maximize its distance from the closest training examples of different classes [1999][1998b]. For a given test pattern $\mathbf{x}$,

Table 3. Classification performance of AANN and SVM models.

Language	% of syllables correctly classified	
	AANN models	SVM models
Hindi	74.68	81.92
Telugu	79.22	80.17
Tamil	76.17	83.26

the evidence is obtained from each of the SVMs, and the maximum evidence is hypothesized as the class of the test pattern. The performance of the classification model using SVMs is shown in Table 3.

4 Evaluation of the Two-Stage Duration Model

For modeling the syllable duration using the proposed two-stage model, syllable parameters are given to all the classification models. Here SVM models are used for syllable classification. The decision logic followed by classification models route the syllable parameters to one of the three FFNN models for predicting the syllable duration. The prediction performance of the two-stage model is presented in Table 4. For comparison purpose, syllable durations are estimated using single FFNN model and its performance is presented in Table 4. Prediction performance of single FFNN and two-stage models for Tamil data is shown in Fig. 2(b). Performance curves in the figure show that short and long duration syllables are better predicted in the case of proposed two-stage duration model. Table 4 and Fig. 2(b) shows that the proposed two-stage model predicts the durations of syllables better compared to single FFNN model.

Table 4. Number of syllables having predicted duration within the specified deviation from actual syllable duration and objective measures for the languages Hindi, Telugu and Tamil using two-stage and single FFNN duration models.

Duration models	Language # Syllables	# Syllables within deviation				Objective measures		
		< 10%	10-25%	25-50%	> 50%	Avg. Err	Std. dev.	Corr.
Two-stage model	Hindi(11222)	4002	4676	2242	302	26.04	20.42	0.81
	Telugu(17000)	6277	6955	2923	842	23.44	23.28	0.82
	Tamil(16811)	7283	6687	2251	590	20.70	21.34	0.85
Single FFNN Model	Hindi(11222)	3312	4012	2875	1023	32.39	25.55	0.74
	Telugu(17000)	4810	5911	4230	2049	28.64	23.92	0.77
	Tamil(16811)	5580	6695	3709	827	25.69	22.56	0.82

5 Conclusions

A two-stage neural network model for predicting the duration of the syllable was proposed in this paper. The performance of the proposed model is shown to be

superior compared to the single FFNN model. The performance of the two-stage model may be improved by appropriate syllable classification model and the selection criterion of duration intervals. 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 the linguistic and phonetic importance may further improve the performance. The accuracy of labeling, diversity of data in the database, and fine tuning of neural network parameters, all of these may also play a role in improving the performance.

References

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

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

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

[1998b] Burges C. J. C.: A tutorial on support vector machines for pattern recognition, Data Mining and Knowledge Discovery, vol. 2, no. 2, pp. 121-167, 1998.