Intonation modeling for Indian languages

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 study neural network models to capture intonation patterns of speech in Indian languages. We examine the performance of neural networks and support vector machines (SVM) for this purpose. Modeling the intonation pattern is the task of predicting the sequence of fundamental frequency (f_0) values for the sequence of syllables in the given text. Analysis is performed on broadcast news data in the languages Hindi, Telugu and Tamil, in order to predict the f_0 of syllables in these languages using neural network and SVM models. The input to both the models consists of a set of phonological, positional and contextual features extracted from the text.

1. Introduction

The generation of the correct intonation pattern for an input text is important for text-to-speech synthesis. The temporal changes of the fundamental frequency (f_0) value depends on the information contained in the text at various levels. They are segmental coarticulation at the phonemic level, emphasis of the words and phrases, syntax, semantic contents of a given sentence, prominence and presence of new information in an utterance [1]. In most of the existing approaches, intonation patterns are derived by using a set of phonological rules [2–5]. Phonological rules are inferred by observing a large set of utterances with the help of linguists and phoneticians. The relationship between the linguistic features of input text and the f_0 contour pattern of utterances is explored. Although this is done by induction, it is generally difficult to analyze the effect of mutual interaction of linguistic features at different levels. Hence, the inferred phonological rules for intonation modeling is always incomplete.

In an alternative approach statistical approaches can be used to model f_0 contours [6]. In this paper we examine models using neural networks and support vector machines (SVM) for predicting the f_0 contours from the sequence of syllables of the input text. Neural networks are known for their ability to generalize and capture the functional relationship between the input-output pattern pairs and have the ability to predict, after an appropriate learning phase, even patterns not presented before [7] [8]. For predicting the f_0 contour, feedforward neural network models are proposed [9]. SVMs also provide a good generalization performance on classification and function approximation problems. SVM performs function estimation from training samples using nonlinear mapping of the data on to high-dimensional feature space [10].

The paper is organized as follows: Section 2 describes the models used for predicting the f_0 contour for the sequence of syllables and the features for representing the linguistic context and production constraints of the input text. Section 3 gives the details of the database used for the proposed intonation analysis. Prediction performance of the proposed models is presented in Section 4. A summary of the paper is given in the final section along with a discussion on some issues to be addressed further.

2. Features and models for predicting the f_0 of a syllable

2.1. Features

The features considered for modeling the fundamental frequency of a syllable 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 starting position of the phrase. 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: Fundamental frequency of the syllable may be influenced by its adjacent syllables. Hence for modeling the f_0 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 represented with three independent codes specifying three distinct features.

Duration and pitch: Length of the present syllable and preceding syllable pitch value may influence the f_0 of the present syllable. Therefore these two syllable parameters are used as constituents in feature vector for predicting the f_0 of the syllable.

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

Table 1: List of the factors affecting the f_0 of the syllable, features representing the factors and the number of nodes needed for neural network or SVM 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
Duration	Duration of the present syllable	1
Pitch	f_0 of the previous syllable	1

2.2. Neural network model

For modeling the f_0 of a syllable, 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, and they can be interpreted as capturing some local and global features in the input space [7] [8]. The fourth layer is the output layer having one unit representing the f_0 of the syllable. For better generalization, several network structures are experimentally verified. The optimum structure arrived is 23*L* 46*N* 11*N* 1*N*, 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 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 f_0 value.

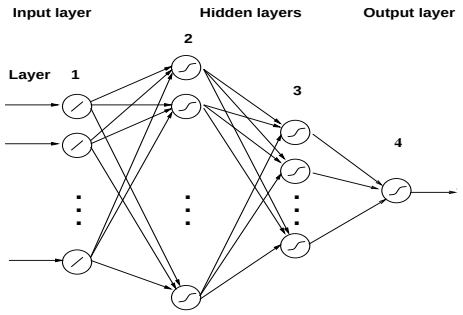

Figure 1: Four layer feedforward neural Network

2.3. SVM regression model

Support vector machines predict the f_0 of a syllable using regression (function approximation). For a given training data $\{(\mathbf{x_1}, \mathbf{y_1}), (\mathbf{x_2}, \mathbf{y_2}), ...(\mathbf{x_l}, \mathbf{y_l})\}$, where $\mathbf{x_i}$ denotes the input pattern and $\mathbf{y_i}$ is the desired response (target) for the corresponding input pattern. In SVM regression, the goal is to find a function $f(x)$ that has at most ϵ deviation from the actually obtained targets $\mathbf{y_i}$ for all the training data [8] [10]. This is achieved by mapping the data points on to high-dimensional feature space using a nonlinear mapping function and then constructing a hyperplane that lies close to as many of the data points as possible. Therefore the objective is to choose a hyperplane with small norm, while simultaneously minimizing the sum of the distances from the data points to the hyperplane. The general SVM architecture for regression is shown in Fig. 2. After the training phase, the approximate function can be estimated by a subset of training data, such data points are known as support vectors. The number of support vectors depends on the precision required for approximating the original data. Here Gaussian kernel Φ is used for nonlinear transformation of data from input space to high-dimensional feature space. For predicting the f_0 of a syllable, dot products are computed with the images of the support vectors under the mapping

Φ. This corresponds to evaluating kernel functions $\mathbf{K}$ at locations $\mathbf{K}(\mathbf{x_i}, \mathbf{x})$. Finally the dot products are added up using weights $\mathbf{v_i}$. This plus the constant term $\mathbf{b}$ yields the final prediction output.

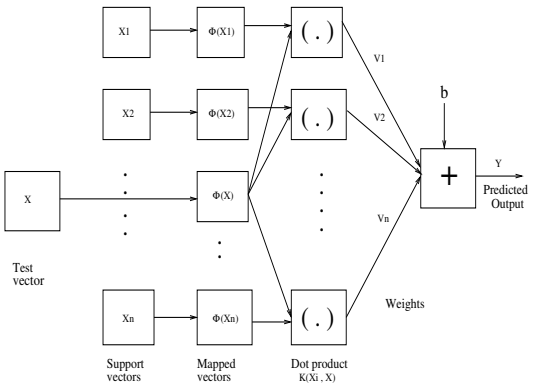

Figure 2: SVM model for regression

3. Speech database

The database consists of three Hindi, seven Telugu and three Tamil news bulletins. Total durations of speech in Hindi, Telugu and Tamil are around 45 minutes, 105 minutes and 65 minutes, 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 fundamental frequencies of the syllables are computed using Entropic Speech Processing Software (ESPS) package. The average pitch (f_0) for male speakers and female speakers in the database is found to be 129 and 231 Hz, respectively. The total database consists of 9530 syllables in Hindi, 27187 syllables in Telugu and 14185 syllables in Tamil.

4. Evaluation of the proposed intonation models

4.1. Evaluation of FFNN model

A separate model is prepared for each of the speakers in three languages. For each syllable the phonological, positional and contextual features are extracted and a 23 dimension input vector is formed. The fundamental frequency of the each syllable is obtained from the database. The extracted input vectors are given as input, and the corresponding f_0 value of the syllable is given as output to the FFNN model. The network is trained for 500 epochs. For each syllable in the test set, predict the f_0 using FFNN by giving the input vector of syllable as input to the neural network. The deviation of predicted f_0 from the actual f_0 is estimated. The results in terms of number of syllables for various deviations from actual f_0 value of the syllable are given in Table 2. In Table 2, the first column indicates the language, the second column shows the number of syllables specific to the speaker used for testing, and the other columns indicate the number of syllables having predicted f_0 within the specified deviation with respect to actual f_0 of the syllable. In order to evaluate the prediction accuracy (objective measures) between predicted and actual values, mean absolute error (MAE) and correlation coefficient (γ) are computed using the following formulations.

$$MAE = \frac{\sum_i |(x_i - y_i)|}{N}, \quad \gamma = \frac{\sum_i |(x_i - \bar{x})||(y_i - \bar{y})|}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}$$

Where x_i and y_i are the actual and predicted f_0 values, $\bar{x}$ and $\bar{y}$ are the mean values of actual and predicted f_0 values. The prediction accuracy results for different speakers in the three languages are given in Table 4. Fig. 3 shows the distribution of actual f_0 values against the predicted values for syllables of Hindi male speaker data. Although the distribution seems to be tightly clustered, the orientation of the distribution is slightly deviated from the expected. This is due to the prediction error of the neural network model.

Table 2: The number of syllables having predicted f_0 within the specified deviation from actual f_0 value for different speakers from three languages Hindi, Telugu and Tamil using neural network model.

Language	Speaker gender # Syllables	# Syllables with deviation			
		$< 10\%$	10-15%	15-25%	$> 25\%$
Hindi	Female(786)	445	149	145	46
	Female(660)	443	101	89	26
	Male(1084)	806	176	93	8
Telugu	Male(4427)	2455	835	762	374
	Male(984)	628	185	136	34
	Female(1276)	916	237	103	19
Tamil	Male(949)	494	166	193	95
	Female(1495)	973	262	211	48
	Female(741)	569	108	57	6

4.2. Evaluation of SVM regression model

Training and testing data used for evaluating the performance of the SVM regression model is the same used for neural network models in Section 4.1. For the given training data SVM model is optimized by varying the width of the error pipe ϵ and standard deviation σ of the Gaussian kernel Φ. The performance of the SVM regression model in predicting the f_0 of the syllables for different speakers in the three languages are presented in Table 3 and Table 4. The performance measures indicate that SVM models also captures the intonation patterns similar to neural network models. The overall prediction error of the models is about 12 Hz (9.3% w.r.t mean f_0) for male speakers and 19 Hz (8.2% w.r.t mean f_0) for female speakers. About 87% of the syllables are predicted within 15% deviation.

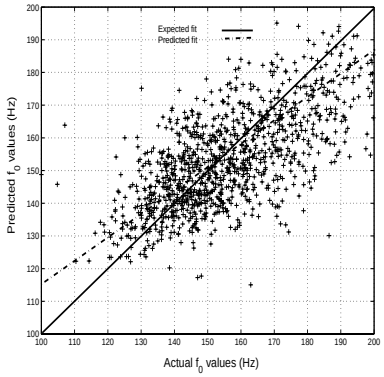

Figure 3: Distribution of actual f_0 values against the predicted values for the syllables of Hindi.

Table 3: The number of syllables having predicted f_0 within the specified deviation from actual f_0 value for different speakers from three languages Hindi, Telugu and Tamil using SVM model.

Language	Speaker gender # Syllables	# Syllables with deviation			
		$< 10\%$	10-15%	15-25%	$> 25\%$
Hindi	Female(786)	475	139	135	36
	Female(660)	415	82	111	51
	Male(1084)	838	142	94	9
Telugu	Male(4427)	2508	776	776	359
	Male(984)	665	170	116	32
	Female(1276)	970	201	96	8
Tamil	Male(949)	500	168	167	113
	Female(1495)	970	261	207	56
	Female(741)	561	122	50	7

5. Summary and Conclusions

In this work a feedforward neural network and SVM regression models were used for predicting the f_0 contour for the sequence of syllables. Phonological, positional and contextual parameters were extracted from the syllables of each of the three Indian languages. The models are objectively evaluated by computing the mean absolute error (MAE) and correlation coefficient (γ) between predicted and actual f_0 values of the syllables. 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 and SVM parameters, all of these may also play a role in improving the performance.

Table 4: The objective measures (MAE and γ) for FFNN/SVM models in predicting the f_0 for syllables spoken by different speakers from three languages Hindi, Telugu and Tamil.

Language	Speaker gender # Syllables	FFNN model		SVM model	
		MAE(Hz)	γ	MAE(Hz)	γ
Hindi	Female(786)	21.63	0.74	21.06	0.72
	Female(660)	20.41	0.79	19.60	0.81
	Male(1084)	11.26	0.79	10.84	0.79
Telugu	Male(4427)	13.78	0.73	13.96	0.81
	Male(984)	10.22	0.79	9.94	0.78
	Female(1276)	16.55	0.78	15.95	0.79
Tamil	Male(949)	14.76	0.81	14.93	0.80
	Female(1495)	21.54	0.83	21.69	0.82
	Female(741)	18.01	0.85	18.18	0.83

6. References

[1] D. H. Klatt, "Review of text-to-speech conversion for English," *Journal of Acoustic Society of America*, vol. 82(3), pp. 737–793, Sep. 1987.

[2] J. P. Olive, "Fundamental frequency rules for the synthesis of simple declarative english sentences," *Journal of Acoustic Society of America*, no. 57, pp. 476–482, 1975.

[3] H. Fujisaki, K. Hirose, and N. Takahashi, "Acoustic characteristics and the underlying rules of the intonation of the common Japanese used by radio and TV anouncers," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Processing*, pp. 2039–2042, 1986.

[4] P. A. Taylor, "Analysis and synthesis of intonation using the Tilt model," *Journal of Acoustic Society of America*, no. 107, pp. 1697–1714, 2000.

[5] A. S. Madhukumar, S. Rajendran, C. C. Sekhar, and B. Yegnanarayana, "Synthesizing intonation for speech in Hindi," in *Proc. of the 2nd European Conf. on speech communication and technology*, vol. 3, (Geneva, Italy), pp. 1153–1156, 1991.

[6] S. H. Chen, S. M. Lee, and S. Chang, "A Chinese fundamental frequency synthesizer based on a statistical model," in *Proc. Int. Conf. Spoken Language Processing*, (Kobe, Japan), pp. 829–832, 1990.

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

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

[9] S. H. Hwang and S. H. Chen, "Neural-network-based F0 text-to-speech synthesizer for Mandarin," *IEE Proc. Image Signal Processing*, vol. 141, pp. 384–390, Dec. 1994.

[10] A. Smola and B. Scholkopf, *A Tutorial on Support Vector Regression*. Technical report Neuro COLT NC-TR-98-030, 1998.