Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 38 (2012) 2772 – 2783

Procedia Engineering

www.elsevier.com/locate/procedia

ICMOC

Intonation Modeling Using Linguistic, Production and Prosodic Constraints for Syllable based TTS Systems

V. Ramu Reddy, K. Sreenivasa Rao

School of Information Technology, IIT Kharagpur, Kharagpur- 721302, West Bengal

Abstract

This paper proposes linguistic, production and prosodic constraints for modeling the intonation patterns of sequence of syllables. Linguistic constraints are represented by positional, contextual and phonological features, production constraints are represented by articulatory features, and prosodic constraints are represented by durations and intensities of syllables. Neural network models are explored to capture the implicit intonation knowledge using above mentioned features. The prediction performance of the neural network models is evaluated using objective measures such as average prediction error (μ), standard deviation (σ) and linear correlation coefficient ($\gamma X,Y$).The prediction performance of the feed-forward neural network (FFNN) models is compared with other statistical models such as Classification and Regression Tree (CART) and Linear Regression (LR) models. The performance of the intonation models is also analyzed by conducting listening tests to evaluate the quality of synthesized speech after incorporating the models in baseline TTS system.

Keywords: Intonation, TTS, FFNN, CART, Contextual, Phonological, Positional, Articulatory, Prosody, Festival.

Introduction

The goal of text-to-speech (TTS) system is to generate speech from the given input text [1]. People who are visually impaired, having reading disabilities or struggling readers will be benefited greatly with the integration of TTS system to screen reader softwares such as Orca or Non-visual Desktop Access (NVDA) [2]. People from rural areas and illiterates can also browse the Internet for different applications with the help of TTS. Prosody plays an important role in improving the quality of TTS system both in terms of naturalness and intelligibility. Prosody refers to duration, intonation and intensity patterns of speech for the sequence of syllables, words and phrases. In this work, we are focusing on modeling the

1877-7058
doi:10.1016/j.proeng.2012.06.325

intonation patterns of sequence of syllables. Intonation can be defined as the dynamics of F0 contour over time, caused by vocal fold vibration. The perceptual correlate of F0 is pitch. Human hearing system is very sensitive to variations of pitch [3]. In the previous work, we modeled F0 patterns of syllables using neural networks giving linguistic constraints represented by positional, contextual and phonological features as input to the network [4]. In this work, we proposed production and prosodic constraints in addition to linguistic constraints to predict the F0 values of the sequence of syllables. Production constraints are represented by articulatory features where as prosodic constraints are represented by durations and intensities of syllables provided by duration and intensity modules.

Rest of the paper is organized as follows: Following section presents the speech database used in this work. Section 3 discuss the performance of the intonation model using Festival features. Features represented by linguistic, production and prosodic constraints are explained in section 4. Modeling the intonation patterns using neural networks is presented in section 5. Section 6 discuss the performance results of intonation models using proposed features by means of objective and subjective tests. Summary of this paper is laid in the final section.

Speech database

The text utterances of speech database used for this study is collected from Bengali Anandabazar newspaper, various text books and story books which covers wide range of domains. The collected text corpus covers 7762 declarative sentences with 4372 unique syllables and 22382 unique words. The text is recorded with a professional female artist in a noiseless chamber. The duration of total recorded speech is around 10 hrs. The speech signal was sampled at 16 kHz and represented as 16 bit numbers. The speech utterances are segmented and manually labeled into syllable-like units. For every utterance a labeled file is maintained which consists of syllables of the utterance and their timing information. The syllable structures considered here are V, CV, CCV, CVCC and CCVC, where C is a consonant and V is a vowel. For developing intonation models, the fundamental frequencies (F0) of the syllables should be available in the database. The fundamental frequencies of the syllables in the database are computed by using zero frequency filter method [5]. It is observed from the database that F0 values vary from 80 to 290 Hz. The average and standard deviation of pitch (F0) of female speaker in the database was found to be 223 and 47 Hz, respectively.

Intonation modeling using Festival default features

The baseline Bengali TTS system was developed using Festival framework [6]. The prediction performance of the CART model with the features used by Festival is given in Table1.

Table1. Performance of the CART based intonation model using Festival features.

F0 position in syllable	% Predicted syllables within deviation					Objective measures		
	2%	5%	10%	15%	25%	μ (Hz)	σ (Hz)	γ
Start	6.31	14.96	29.87	43.29	62.1	54.21	47.17	0.65
Middle	8.65	17.67	34.13	47.10	67.32	44.47	38.13	0.67
End	6.02	10.93	25.59	41.16	58.35	56.82	48.19	0.64

Here the prediction accuracy of the model is computed using objective measures such as average prediction error (μ), standard deviation (σ) and linear correlation coefficient (γX,Y). The number of predicted syllables within different deviations are also shown in Table1. The details of objective measures is discussed in Section 6.1. From the Table1 it is observed that the average prediction error seems to be large. This is mainly is due to poor prediction of F0 values by CART model and this in turn depends upon the features used to model the intonation patterns of syllables. The features used by Festival to model the intonation are more stress related features and ignores the positional, phonological and articulations of sound units information. These features are more suitable for stress-timed languages such as English, German, Danish, Swedish, Norwegian, Faroese, Dutch and Portuguese. But however, Bengali is syllable-timed language having equal stress on each syllable. Therefore, more deviation is observed between predicted and actual F0 values. Therefore, there is need for the features which are more suitable for syllable based TTS systems. Hence, in this work for prediction of intonation, we proposed syllable specific features for modeling the intonation patterns of sequence of syllables.

Proposed features

It is known that there exists some inherent relationship between linguistic and production constraints of speech to the intonation patterns in speech [4][7]. The linguistic constraints of syllables can be expressed using positional, contextual and phonological (PCP) features and production constraints of syllables can be expressed using articulatory features (A). In this study we use 35 dimensional feature vector representing the linguistic constraints and production constraints of each syllable. Out of 35 features, 24 features represents the linguistic constraints in the form of positional, contextual and phonological information and remaining 11 features represents articulatory information of production constraints of each syllable. The positional features are further classified based on syllable position in a word and sentence, and word position in a sentence.

4.1. Linguistic Constraints

The list of features representing linguistic constraints and the number of input nodes needed for the neural network to represent these features are given in Table2. These features are coded and normalized before giving to neural network.

Table2. List of factors affecting the intonation of syllables, features representing the factors and the number of nodes needed for neural network to represent the features.

Factors	Features	# Nodes
Syllable position in the sentence	Position of syllable from beginning of the sentence	
	Position of syllable from end of the sentence	3
	Number of syllables in the sentence	

Syllable position in the word	Position of syllable from beginning of the word	
	Position of syllable from end of the word	3
	Number of syllables in the word	
Word position in the sentence	Position of word from beginning of the sentence	
	Position of word from end of the sentence	3
	Number of words in the sentence	
Syllable identity	Segments of the syllable (consonants and vowels)	4
Context of the syllable	Previous syllable	4
	following syllable	4

4.2. Production Constraints

Intonation patterns of speech segments are also influenced by the production mechanism of speech sounds. Each sound unit has specific articulatory movements and positions while producing the sound. These production constraints in turn depends on the language. In this study the features related to different articulatory positions and manners of speech segments (consonants and vowels) are considered as production constraints. These production constraints are represented as articulatory features to predict the F0 values of syllables. The features used to represent the articulatory information are vowel length, vowel height, vowel frontness, vowel roundness (lip rounding), consonant type, consonant place, consonant voicing, aspiration, nukta (diacritic mark), type of first phone and type of last phone in a syllable. The detailed list of production constraints represented in the form of articulatory features is given in Table3.

Table2. List of articulatory features.

Features	Description
vlen	Length of the vowel in a syllable (short, long, dipthong and schwa).
vheight	Height of the vowel in a syllable (high, mid and low).
vfront	Frontness of the vowel in syllable (front, mid and back).
vrnd	Lip roundness (no rounding and rounding).
ctype	Type of consonant (stop, fricative, affricative, nasal, and liquid).
cplace	Place or position of the production of the consonant (labial, alveolar, palatal, labio-dental, dental and velar).
cvox	Whether consonant is voiced or unvoiced (voiced and unvoiced).
asp	Whether consonant is aspirated or unaspirated (aspirated and unaspirated).
nuk	Whether consonant with nukta or not nukta (with nukta and without nukta).
fph	Type of first phone in a syllable (vowel, voiced consonant, unvoiced consonant, nasal, semi-vowel, nukta and fricative).
lph	Type of last phone in a syllable (vowel, voiced consonant, unvoiced consonant, nasal, semivowel,

Features	Description
	nukta and fricative).

4.3. Prosodic Constraints

In speech signal, the duration, intonation and intensity patterns of the sequence of sound units are interrelated at some higher level through emphasis (stress) and prominence of the words and phrases. But representation of the feature vector to capture these dependencies is difficult. In this study, the durations of the syllables are considered as duration constraints and the intensities of the syllables are considered as intensity constraints. The duration and intensity constraints can be obtained from the duration and intensity models.

Modeling the F0 patterns with proposed features using feed-forward neural networks

In this work, a four layer feed-forward neural network (FFNN) [8][9][10]with input layer, two hidden layers and output layer is used to model the intonation of the syllables. The structure of the four layer FFNN is shown in Fig.1. Different structures were explored to obtain the optimal four layer FFNN, by incrementally varying the hidden layers neurons in between 5 to 50. The optimal structure 35L 72N 19N 3L is obtained with minimum generalization error after exploring several structures, where L denotes linear unit and N denotes non-linear unit. The activation function used for non-linear unit (N) is hyperbolic tangent i.e., *tanh(s)* function. Here, the input and output features are normalized between [-1, 1] before giving to the neural network. The mapping function is between the 35-dimensional input vector and the 3-dimensional output vector. Here, three F0 values for each syllable were chosen to capture the broad shape of the intonation contour of syllables. Two hidden layers which comprise 72 and 19 hidden units in the first and second hidden layers can capture the local and global variations across the features. The FFNN operates from left to right and training was carried using Lavenberg-Marquardt back-propagation algorithm. In this work, the total data consists of 177820 syllables for modeling the intonation, is divided into two parts namely design data and test data. The design data is used to determine the network topology, training algorithm and training algorithm parameters. The design data composed of two parts namely training data and validation data. Training data is used to estimate the weights (includes biases) of candidate units and validation data is used to estimate the non-training performance error of candidate units, and can also be used to stop training once the non-training validation error estimate stops decreasing. The test data is used once and only once on the best design, to obtain an unbiased estimate for the predicted error of unseen non-training data. The percentages of data divided for training, validation and testing the network is 70%, 15% and 15% respectively.

Evaluation of intonation model using proposed features

The prediction performance of the proposed FFNN model is evaluated by means of objective and subjective tests. In addition to FFNN model we also explored other non-linear statistical model like Classification and Regression Tree (CART) and linear statistical model like Linear Regression (LR) models in predicting the F0 values. The performance of the FFNN model is compared with LR and

CART models developed by positional, contextual, phonological and articulatory (PCPA) features. The details of objective and subjective tests of the models are discussed in the following subsections.

6.1.Objective Evaluation

The intonation model is evaluated with the syllables in the test set. The three average F0 values located at start, middle and end positions of each syllable in the test set are predicted using FFNN by presenting the feature vector of each syllable as input to the network. The prediction performance of the FFNN model with PCPA features is given in Table 4. The percentage of syllables predicted within different deviations

from their actual F0 values are shown in Table4. The prediction accuracy is evaluated by means of objective measures such as average prediction error, standard deviation and linear correlation coefficient between actual and predicted F0 values is shown in Table4. The performance of FFNN is compared with LR and CART models. The prediction performance of the LR and CART models is also given in Table4. From the Table4, it is observed that the prediction performance of LR, CART and FFNN models using PCPA features is better than CART model using Festival features (Table1). This indicate that the proposed syllable specific PCPA features are more appropriate for accurate prediction of the F0 values. It is also observed that among LR, CART and FFNN models, the prediction performance of LR model is low compared to other models. The lower performance of the linear models can be attributed to their inability to capture the non-linear (complex) relations present in the data. Among all models FFNN model perform better in predicting the intonation patterns of sequence of syllables. From this we can hypothesize, that the neural network model capture the inherent complex relationships between PCPA features and F0 values of syllables reasonably well compared to other models.

Table4. Performance of the LR, CART and FFNN models for predicting the F0 values of syllables for Bengali language.

Model	F0 position in syllable	% Predicted syllables within deviation					Objective measures		
		2%	5%	10%	15%	25%	μ (Hz)	σ (Hz)	γ
CART	Start	12.56	27.15	51.97	70.50	82.19	41.01	33.93	0.77
	Middle	16.93	35.74	60.09	77.73	88.92	32.53	28.74	0.80
	End	9.13	23.99	47.75	68.41	79.97	42.15	37.73	0.76
LR	Start	10.11	21.32	45.39	61.50	80.91	43.72	39.10	0.74
	Middle	13.93	29.17	52.65	78.93	87.56	33.96	28.19	0.78
	End	9.01	19.97	45.32	59.99	77.60	44.20	34.87	0.73
FFNN	Start	14.35	31.26	55.25	73.16	87.34	32.19	27.13	.82
	Middle	19.17	40.58	66.35	80.81	90.12	28.31	24.74	0.83
	End	10.57	23.18	51.72	69.89	84.99	35.52	30.91	0.79

The prediction performance of LR, CART and FNN models is also examined by using scatter plots shown in Fig. 1. The scatter plot is generated by jointly plotting the actual and the predicted F0 values. In ideal case, the predicted values should coincide with the actual values, with that all the points should fall on the diagonal, represented by dotted line in Fig. 1. The thick solid line in Fig. 1 represents the average predicted F0 vs. the average F0 of the syllables. The angle between the solid line and dotted line (diagonal) is inversely proportional to accuracy in prediction. From the scatter plots, it is observed that

the angle between solid and dotted lines is less in case of FFNN, compared to LR and CART models. It is observed from the Fig.1, that the predicted F0 values are more deviated from the actual values at lower and higher F0 values. From the Fig.1, it is also observed that the predicted F0 values of LR and CART models are more deviated from actual F0 values compared to FFNN model. The prediction accuracy of FFNN model seems to be better in the range of 140-250 Hz for start F0 , 175-280 Hz for middle F0 , and 140-270 Hz for end F0.

For studying the influence of prosodic constraints along with linguistic and production constraints in predicting the intonation, three FFNN models are developed. Two FFNN models are developed with respect to each duration and intensity constraints, and one FFNN model with respect to combination of duration and intensity constraints. The performance of these models is given in Table5.

The results indicate that the prediction performance has improved by imposing the duration and intensity constraints. But, the prediction accuracy is outperformed when imposing all the constraints. From the Table5, it is observed that the percentage of syllables predicted within 2-25% is increased by including the features related to different constraints. A similar phenomenon is observed in objective measures also.

Table5. Prediction performance of FFNN model using different constraints (L: Linguistic, P: Production, D: Duration, and I: Intensity).

Model	F0 position in syllable	% Predicted syllables within deviation					Objective measures		
		2%	5%	10%	15%	25%	μ (Hz)	σ (Hz)	γ
L+P+D	Start	17.23	33.65	57.53	76.11	88.99	30.91	26.97	0.83
	Middle	21.76	42.91	68.87	82.93	92.34	26.13	24.10	0.84
	End	12.96	27.10	55.24	71.18	85.51	33.39	29.53	0.81
L+P+I	Start	16.93	32.98	55.99	75.45	87.69	31.53	27.01	0.83
	Middle	20.19	41.33	67.18	81.19	90.93	27.70	24.23	0.83
	End	11.13	25.85	54.71	70.78	85.10	43.94	29.98	0.80
L+P+D+I	Start	19.01	34.56	59.97	77.81	90.07	27.17	24.53	0.84
	Middle	24.34	43.11	70.08	84.24	93.41	25.91	23.75	0.85
	End	16.97	29.72	57.93	73.17	86.93	30.33	27.18	0.82

6.2. Subjective Evaluation

Naturalness and intelligibility are two important key features to measure the quality of the synthesized speech. The perceptual evaluation is conducted by incorporating LR, CART and FFNN based intonation models into baseline TTS system. The overall architecture representing the incorporation of intonation models in TTS system is given in Fig.2. In this work, 20 subjects within the age group of 23-35 were considered for perceptual evaluation of synthesized speech. After giving appropriate training to the subjects, evaluation of TTS system is carried out in a laboratory environment. Randomly 10 sentences were selected and played the synthesized speech signals through headphones to evaluate the quality. Subjects have to assess the quality on a 5-point scale for each of the synthesized sentences. The subjective

listening tests are carried out for the synthesized sentences generated by LR, CART and FFNN models using PCPA features and CART model using Festival default features. The mean opinion scores (MOS) are calculated for naturalness and intelligibility of the synthesized speech. Table6 shows the MOS values for the synthesized speech correspond to different intonation models. From the Table6, it is observed that the MOS values for naturalness and intelligibility of proposed FFNN model seems to be better compared to other models. The scores indicate that the intelligibility of the synthesized speech is fairly acceptable, whereas the naturalness seems to be poor. Naturalness is mainly attributed to individual perception. Naturalness can be improved to some extent by incorporating the appropriate duration and intensity information along with intonation. The significance of the differences in the pairs of the mean opinion scores for intelligibility and naturalness is tested using hypothesis testing. The level of confidence for the observed differences in the pairs of MOSs between proposed FFNN model and other models are given in observed differences in the pairs of MOSs between proposed FFNN model and other models are given in cases. This indicates that the differences in the pairs of MOS in each case is significant. From this study, we conclude that the quality of speech using proposed FFNN model is better than other models at perceptual level and also the accuracy in prediction of F0 values of syllables is significantly better by

using proposed features compared to Festival default features.

Table6. Mean opinion scores and confidence levels for the quality of synthesized speech of Bengali TTS after incorporating the intonation models.

Model with input features	Mean opinion score (Confidence level)	
	Naturalness	Intelligibility
CART model with Festival features	3.24 (>99.5)	2.60 (>99.5)
LR model with PCPA	3.29 (>99.5)	2.64 (>99.0)
CART model with PCPA	3.37 (>99.0)	2.70 (>97.5)
FFNN model with PCPA	**3.41**	**2.79**

Summary and Conclusions

The intonation pattern of the sound unit depends on its linguistic, production and prosodic constraints. Linguistic constraints are represented by positional, contextual and phonological features, production constraints are represented by articulatory features and prosodic constraints are represented by durations and intensities of syllables. Feed-forward neural networks were proposed for predicting the F0 values of the syllables in Bengali using above mentioned features. The performance of the neural network models is compared with LR and CART models. The evaluation of quality of TTS system is carried out by incorporating the intonation models into baseline Bengali TTS system.

Acknowledgements

We would like to acknowledge Department of Information Technology (DIT), Government of India for offering this project. We also acknowledge the research scholars of IIT Kharagpur for their voluntary participation in listening tests to evaluate the Bengali TTS system developed in this work.

References

[1] Benesty J, Sondhi MM, Huang Y (Eds.). *Springer Handbook on Speech Processing.* Springer Publishers; 2008.

[2] Narendra NP, Rao KS, Ghosh K, Reddy VR, Maity S. Development of Bengali screen reader using Festival speech synthesizer. In: *2011 Annual IEEE India Conference (INDICON)*. Hyderabad, India; 2011, p. 1–4.

[3] Moore BCJ. *An introduction to the psychology of hearing*. 3rd ed. San Diego, CA: Academic Press; 1989.

[4] Reddy VR, Rao KS. Intonation Modeling using FFNN for Syllable based Bengali Text To Speech Synthesis. In: *Proc. Int. Conf. Computer and Communication Technology*, MNNIT, Allahabad; 2011, p. 334–339.

[5] Murty K, Yegnanarayana B. Epoch extraction from speech signals. *IEEE Trans. Audio, Speech, and Language Processing* 2008; 16: 1602–1613.

[6] Narendra NP, Rao KS, Ghosh K, Reddy VR, Maity S. Development of syllable-based text to speech synthesis system in Bengali. *Int. J. of Speech Technology, Springer* 2011; 14(3) : 167–181.

[7] K. S. Rao, B. Yegnanarayana, Intonation modeling for Indian languages, *Computer Speech and Language* 2009; 23: 240–256.

[8] Tamura S, Tateishi M. Capabilities of a Four-Layered Feedforward Neural Network: Four Layers Versus Three. I*EEE Trans. on Neural Networks* 1997; 8 (2): 251–255.

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

[10] Yegnanarayana B. *Artificial Neural Networks*, Prentice-Hall, New Delhi, India, 1999.

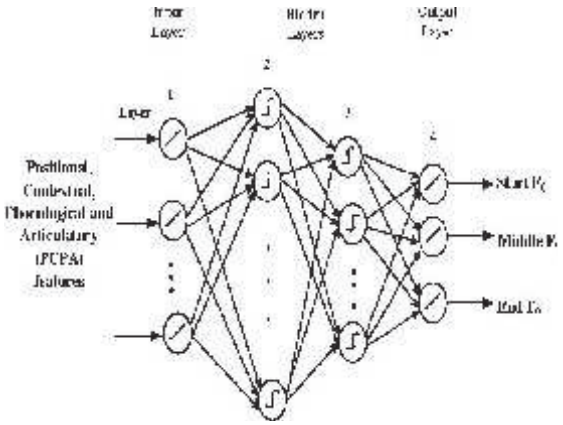

Fig. 1. Architecture of four layer feed-forward neural network for predicting the F0 values of syllables.

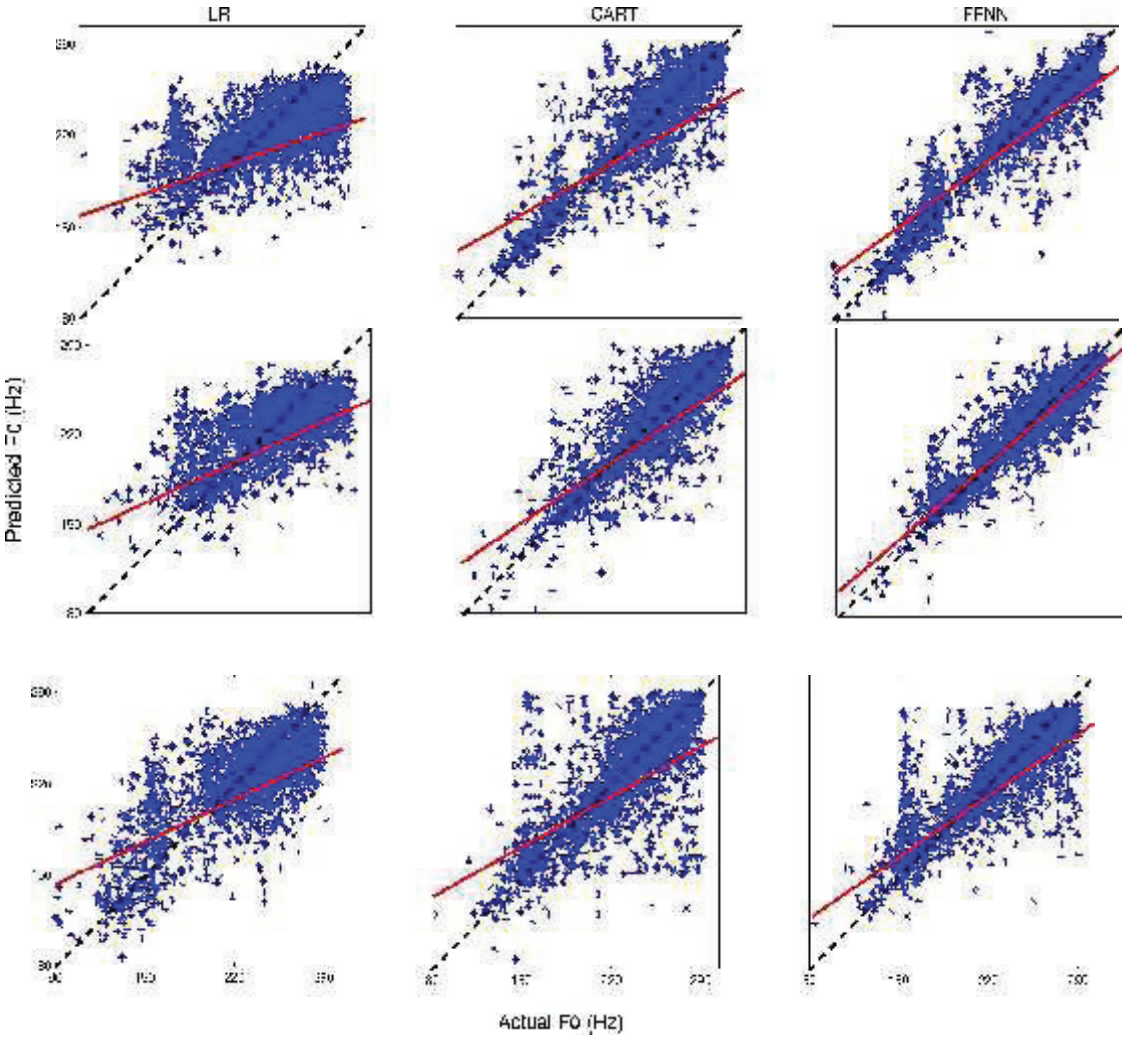

Fig. 2. Prediction performance of intonation models developed using LR, CART and FFNN with scatter plots.

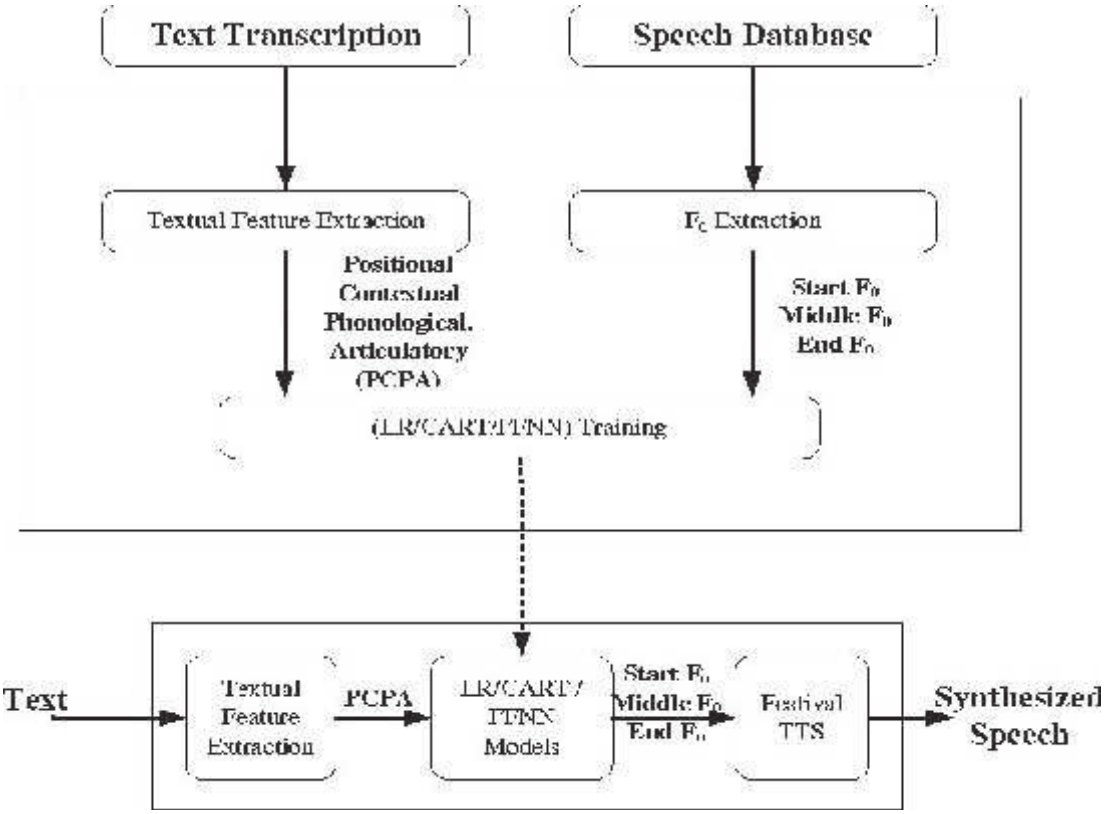

Fig. 3. Incorporation of intonation models in TTS system.