DEVELOPMENT OF STORY TEXT-TO-SPEECH SYSTEM BASED ON STORY GENRES

Parakrant Sarkar, K. Sreenivasa Rao, Gurunath Reddy M

Department of Computer Science and Engineering,
Indian Institute of Technology Kharagpur, Kharagpur - 721302, West Bengal, India
parakrantsarkar@gmail.com, ksrao@iitkgp.ac.in,mgurunathreddy@sit.iitkgp.ernet.in

ABSTRACT

Storytelling is a distinct speaking style which embraces various expressive subtleties aimed to draw the attention of the children. Studies have shown that Text-to-speech (TTS) systems have the tendency of not conveying the right emotion expressivity in their speech outputs. The objective of this work is to develop prosody models to capture the story semantics present in the Hindi children stories. In this work, we have considered three prosodic parameters to be modeled: duration, intonation, and intensity of the syllables. We have proposed textual (positional, contextual and phonological) and story-semantic features for modeling the prosody. The proposed prosody models will be integrated with the TTS framework to synthesize speech in storytelling style. We used Classification and Regression Tree (CART), Feed-forward neural network (FFNN), and Support Vector Machine (SVM) for modeling the prosody. We pose the problem of modeling the prosody as a data driven statistical transformation from input text onto the feature space to capture the implicit prosodic and semantic knowledge of the syllables. The evaluation of the models was carried out using objective and subjective measures. A framework is designed to further improve the performance of the prosody models by including the story genre specific prosody modeling. The three genres of the story considered are Folk-tale, Legendary, and Fable. The subjective evaluation of the synthesized story speech shows the proposed framework is capable of improving the storytelling style.

Index Terms— Speech synthesis, storytelling style, emotion expressivity, story genre, text-to-speech system

1. INTRODUCTION

In Text-to-Speech (TTS) system, prosody contributes to the naturalness and expressivity of the synthesized speech [1, 2, 3, 4, 5]. Prosody is a concoction of the supra-segmental features like intonation, duration, intensity, and pause patterns. It provides an inference of additional knowledge in the speech that is not present in the text. An appropriate duration, tone (intonation) with sufficient intensity of speech can help in conveying the speech nuances in a content-rich manner. Although the generation of synthetic speech with a certain degree of expressiveness has been successful for some particular [6] speaking styles (e.g. emotions). In contrast, storytelling is a particular speaking style which constitutes expressive subtleties in the speech, has not experienced much of consideration in the literature. Therefore, a good TTS system should generate a the speech similar to a human storyteller, so it can engage the children.

In this paper, we build a syllable based story TTS system in Hindi. To synthesize a natural and storytelling style speech, we need good prosody models. These prosody models should capture the duration, intonation, and intensity patterns of the story style speech. The present Story TTS system tends to synthesize input text that sounds natural but lacks in demonstrating the storytelling style. It is due to the deficiencies of syntactic and semantic analysis of the raw input text. However, this problem can be mended by efficient extraction of the affective clues that could be used in the synthesis phase. This is our main motivation for carrying out this study and our objective is to determine whether non-linear models can capture the implicit knowledge of the prosodic pattern of the syllables in story style speech.

Automatic prediction of the prosody in a TTS system for Indian Languages is not new, earlier works are shown in [7, 8, 9, 10]. These studies were carried out focusing on the development of the prosody models using features related to a syllable. Here, our primary contribution lies in the text processing module of TTS, in which we concentrate on exploring features that can help in capturing the story semantic knowledge present in the text. In view of this, we developed prosody models trained with our proposed textual (positional, contextual and phonological features) and story-semantic features not only to capture the prosodic pattern but also semantic information. Further, we introduced the genre information of the stories for increasing the performance of the prosody models.

2. SYLLABLE BASED STORY TTS SYSTEM

In this work, a unit selection based TTS system was developed with story speech corpus recorded by a professional female artist in Hindi. In Indian Languages for developing TTS systems, the choice of the basic unit is syllable [11]. Syllable is a basic speech unit consisting of the vowel at the syllable nucleus surrounded by optional consonants on both sides. In [11], authors have developed syllable based TTS system using Festival framework [12] for 13 Indian languages.

Table 1: Statistics of prosodic features of syllables present in the story speech corpus (μ:mean σ: standard deviation)

Prosodic Features	Min	Max	μ	σ
Duration (in ms)	15	650	195	96
Intonation (in Hz)	50	495	223	49.3
Intensity (in dB)	20	90	69	7.8

2.1. Story corpus

The story speech corpus in Hindi consists of both story text and its corresponding story wave files. The children story texts are collected

from story books like *Panchatantra*[1] and *Akbar Birbal*[2]. The speech corpus comprises of 105 stories with a total of 1960 sentences and 25340 words. These stories were recorded by a professional female storyteller in a noise free studio environment. For maintaining the high recording quality of the stories, continuous feedback is given to the narrator for improving the quality of the recordings. The speech signal was sampled at 16 kHz and represented as 16-bit numbers. The total duration of the story corpus is about 3 hours.

To develop the prosody models for the duration, intonation, and intensity, we need to extract these information with respect to each syllable present in the corpus. These prosodic information of the syllables are obtained from the speech utterances that are segmented and labeled into syllable-like units using ergodic hidden Markov models (EHMM). For each utterance, label file is generated, which consist of syllables present in the utterance along with its duration information. From the label file, the duration information of the syllables are obtained. The label files are converted into TextGrid file (Praat label file). The Praat tool is used to extract the intonation (F0) and intensity of syllables. The F0 of the syllable is represented by three values namely, start F0, middle F0, and end F0. The start F0 constitutes the mean of 25% of the beginning frame F0 values of the syllable, the end F0 values constitutes the mean of 25% of the end frame F0 values of the syllable and the mean of the rest frames is considered as middle F0. The average intensity of the syllable is computed in decibel (dB) by using Praat scripts. The structure of the syllable considered in this study are V, CV, CCV, CVCC and CCVC. C is a consonant and V is a vowel. Table 1 gives the minimum, maximum, mean and standard deviations of durations, intonation and intensity values of syllables observed in the corpus.

3. PROPOSED SYSTEM

A storyteller narrates a story in different styles based on the semantics and type of the story. Narration style depends on the genre of the story. In this study, we have considered three genres of the story: Fable, Folktale, and Legend. Fable is a short tale involving animals as essential characters, conveying moral. Folk-tale is a traditional story that is passed on in spoken form from one generation to the next. Legend is a traditional story based on historical or an extremely famous person of a particular geographic region. The proposed framework of the system includes an automatic genre classifier. The purpose of the automatic genre classifier is to decide the genre of the story. The Figure 1 shows the framework for training and testing of the prosody models used for developing the Story TTS system. The training of the genre classifier is carried out based on the POS density and term frequency features derived from the story text. Similar approach for extracting the features is followed, explained in [13]. It is shown in the Figure that DF, IF, and InF are the prosody models for fable genre, DFT, IFT, and InFT are the prosody models for folk-tale genre, and DL, IL, and InL are the prosody models for legend genre. In all the three genres of the story, DF, DFT, and DL; IF, IFT, and IL; InF, InFT, and InL are the models for the duration, intonation, and intensity, respectively. The prosody models are trained offline using our proposed textual (positional, contextual, and phonological) and story-semantic features derived from the story text and later, integrated with TTS system framework to synthesize storytelling style speech. These proposed features capture the affect and semantic information present in the text. The features used for developing the prosody models are given in the Table 3. The story-semantic word

[1]https://en.wikipedia.org/wiki/Panchatantra

[2]https://en.wikipedia.org/wiki/Akbar_Birbal

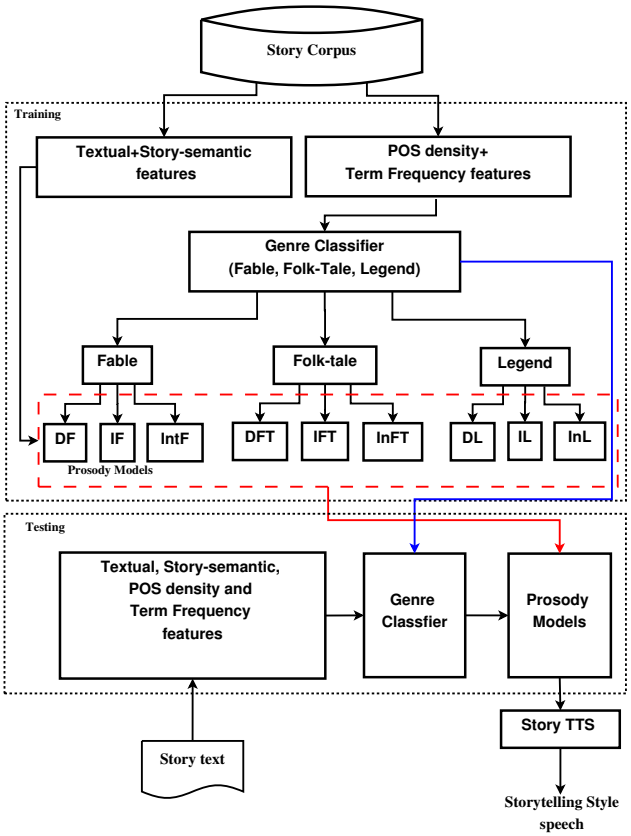

Fig. 1: Flowchart for training and testing of prosody (duration, intonation and pitch) models and automatic genre classifier based on genre classification of story (D:duration, I:intonation, Int:intensity, F:fable, FT: folk-tale, L:legend).

level features such as discourse of the sentence and emotion of the word are manually labeled by four text analysis experts in Hindi. The task of each annotator is to assign a label to the sentence with one of the modes of discourse (i.e. descriptive, dialogue and narrative) [14] and also, with one of the story-specific emotions [15] (i.e sad, happy, fear, anger and neutral). The inter-annotator agreement is given by Fleiss Kappa (κ). The κ values above 0.65 or so is considered to be substantial. The κ values are 0.73 and 0.67 for the two annotation tasks: discourse and emotion respectively. Furthermore, while analyzing the structure of stories, the importance of POS tags with respect to genres of story are observed. A story consists of more named entities and POS tags such as nouns, adjectives, quantifiers, and verbs can be used as a useful features for discriminating the story genres. It motivated us to take POS tags a suitable feature for training the prosody models.

CART, FFNN, and SVM are used for modeling the duration, intonation and intensity of the syllables for three story genres. The 47-dimensional feature vector is used for representing the proposed set of features. In this study, we explored different network structures for FFNN to obtain the optimal performance. It was done by incrementally varying the hidden layer neurons. The structure of the FFNN is represented by A L B N C N D L, where L denotes linear unit and N denotes non-linear unit. A, B, C and D are the integer values that indicate the number of units used in different layers. The

activation function used in the non-linear unit (N) is tanh(s) function, where 's' is activation value of that unit. The structure of the FFNN model used for duration, pitch and intensity are 47L 95N 21N 1L, 47L 92N 19N 3L, and 47L 89N 17N 1L, respectively. The amount of data used for training, validation, and test data are 70%, 15% and 15%, respectively. The training data is used to estimate the weights for FFNN models. The validation data is used to determine the architecture (not weights) of the classifiers, for instance, finding the size of the leaf node in CART, determining the number of hidden units in FFNN and for determining appropriate cost (C) and γ in SVM to minimize over-fitting. It is also used to verify the performance error of the model and to stop training once the non-training validation error estimate stops decreasing. The test data is used once on the best-designed model to obtain an estimate of predicted error for unseen non-training data.

We used the stories that are not used to train the prosody models In testing,. The children story text is given as input to the system. Initially, the genre classifier determines the genre of the story and then the prosody models of that genre are activated for predicting the prosody. For instance, the genre classifier assigns a label to the story as a fable. Therefore to predict the appropriate prosody the DF, IF, and InF models are used to determine the duration, intonation, and intensity of the syllables, respectively. The Table 2 shows the statistics of duration, intonation, and intensity of the syllables for three story genres. It is observed from the Table that the average duration of the syllables is more compared to fable and folktale. It is also noted that the variation in the pitch and intensity is more in the fable genre as compared to folktale and legend.

Table 2: Statistics of prosodic features of syllables present in three genre of the story (μ:mean σ: standard deviation)

Prosodic Features	**Fable**		**folk-tale**		**Legend**	
	μ	σ	μ	σ	μ	σ
Duration (in ms)	208	105.45	212	130.4	216	143
Intonation (in Hz)	240	55.7	235	46	250.5	52
Intensity (in dB)	69	8.4	67	6.6	71	7.2

4. RESULTS AND DISCUSSION

In this section, we discuss about the performance of the proposed prosody models. The performance of the prosody models is carried out by using (i) objective and (ii) subjective evaluation. These are explained briefly in the following subsections.

4.1. Objective Measures

The prosody models comprise of duration, intonation (F0) and intensity are evaluated with the syllables in the test set. The duration, F0 and intensity of each syllable in the test are predicted using CART, FFNN, and SVM. The prediction performance of the CART, FFNN, and SVM models is evaluated by means of objective measures. It includes average prediction error (μ), standard deviation (σ) and correlation coefficient ($\gamma_{x,y}$) between the actual and predicted prosodic values of the syllable. The equations used for computing is given below:

$$\mu = \frac{\sum_i |x_i - y_i|}{N} \tag{1}$$

$$\sigma = \sqrt{\frac{\sum_i d_i^2}{N}}, d_i = e_i - \mu, e_i = x_i - y_i \tag{2}$$

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

$$\gamma_{X,Y} = \frac{V_{X,Y}}{\sigma_X.\sigma_Y}, \text{ where } V_{X,Y} = \frac{\sum_i |x_i - \bar{x}|.|y_i - \bar{y}|}{N} \tag{3}$$

where σ_X, σ_Y are the standard deviations for the actual and predicted prosodic values respectively, and $V_{X,Y}$ is the correlation between the actual and predicted prosodic values. The percentage of the syllables predicted within different deviation from their actual

Table 3: Details of the features used for developing the prosody models (* new proposed feature)

Syllable position in the sentence [3]
Position of syllable from beginning of the sentence Position of syllable from end of the sentence Number of syllables in the sentence
Syllable position in the word [3]
Position of syllable from beginning of the word Position of syllable from end of the word Number of syllables in the word
***Syllable position in the story** [3]
*Position of syllable from beginning of the story *Position of syllable from end of the story *Number of syllables in the story
Word position in the sentence [3]
Position of word from beginning of the sentence Position of word from end of the sentence Number of words in the sentence
***Word position in the story** [3]
*Position of word from beginning of the story *Position of word from end of the story *Number of words in the story
***Sentence position in the story** [3]
*Position of sentence from beginning of the story *Position of sentence from end of the story *Number of sentences in the story
Syllable identity [4]
Identities of individual sounds in the syllable (consonants and vowels)
Context of the syllable [16]
Previous syllable Following syllable Previous previous syllable Following following syllable
Syllable nucleus [3]
Number of segments before the nucleus Number of segments after the nucleus Number of segments in a syllable
Story-semantic features at word level [6]
*Emotion of the sentence *Discourse of the sentence. Parts-of-speech (POS) of the word *Word is stressed or not while speaking *Adjective and adverb count in the sentence. *Number of adjectives and adverbs in the story.

duration, intonation, and intensity values for the proposed prosody model. The deviation is calculated as

$$D_i = \frac{|x_i - y_i|}{x_i} * 100 \quad (4)$$

The objective measures obtained are given in the Table 6 for the proposed prosody models. The performance of the proposed system is compared with the baseline system. The baseline system consist of the prosody models for duration, intonation, and intensity of the syllables which are trained by using the PCP features (positional, contextual, and phonological) as proposed in [7]. Later, these prosody models are integrated with the story TTS system that acts as baseline. The Table 6 shows the performance of the prosody models used in the development of Story TTS i.e. proposed system and baseline system. The numbers outside and within the brackets indicates the performance of the prosody models present in proposed system and baseline, both derived from the predicted and actual prosody values of the syllable. The percentage of the syllables predicted within the different deviation from their actual duration, intonation (F0) and intensity values is also shown in Table 6. The performance of the SVM is the best and CART is the least. The lower performance of the CART can be associated with the incompetence to capture the complex relations present in the data whereas SVM does it better. From the prediction performance of the both the story TTS systems, our proposed system is better.

Table 4: Description of E:experienced and I:inexperienced users selected for subjective evaluation

Users	Occupation	Expertise	Age	#subjects
E	Research Scholars	Expert	23-30	5
	Under Graduates	Intermediate	18-22	5
I	High_School_Students	Novice	13-16	5
	Primary_School_Students	Novice	8-12	5

4.2. Subjective Evaluation

We conducted subjective evaluation in which native Hindi speakers were used as subjects to evaluate the synthesis quality of the speech. The subjective evaluation involves two tests (i) mean opinion score (MOS) and (ii) preference Test. The subjects were asked to listen to the utterances synthesized from the two developed Story TTS systems (baseline and proposed). The subjects evaluated the quality of synthesized speech in terms of naturalness and story style on a five-point scale. The scores are described as 1: very poor quality in conveying story effect; 2: poor quality in conveying story effect;3: good quality in conveying story effect; 4: very good quality in conveying story effect,5: as good as natural story narrated by a storyteller. In preference tests, the subjects are asked to select the best synthesized speech fragment which exhibits the story style synthesized from the two developedsStory TTS systems.

The listening test was carried out on five stories that were not used in the training of the prosody models for the two story TTS systems. We divided the stories into fragments. A fragment consists of one or more sentences, out of which one sentence belongs to one of the five story-specific emotions [15]. The reason for choosing fragment over single sentence is that the story-semantic information is properly captured. There are 15 fragments selected from the five story texts for the listening test. The corresponding speech has been synthesized from the two TTS systems. A total of 35 synthesized speech fragments are used. In the laboratory environment for each of the subjects participating in the test the fragments are played randomly through headphones. In preference tests, subjects were asked to give their preference whether the synthesized fragment conveys the storytelling style. As our work aims at synthesizing speech in story style, we hoped that the children would be a better option for carrying out the listening test. Therefore, we included children for the listening test. The Table 4 shows the description of the 20 subjects involved in the test with varying age groups.

Table 5: Results of MOS scores for various user profiles for baseline and proposed Story TTS systems. The numbers inside the brackets shows the preference value of the system in percentage

User Profiles	MOS (Preference %)	
	Baseline	Proposed
Experienced	2.75 (30)	3.25(70)
Inexperienced	2.90(20)	3.55(80)
Overall	2.82(25)	3.40(75)

The Table 5 shows the results of the listening test according to relevant user profiles. The 2^{nd} and 3^{rd} column of the Table shows the MOS scores for the baseline and proposed story TTS systems. The overall score is the average of all values over the user profiles. The overall MOS of our proposed system is 3.40 which is better than the baseline system. It is noted from the results of the preference test that the subjects preferred speech fragments synthesized using the proposed framework for our proposed system in 75% of the cases and remaining 25% preferred the baseline, indicating that the story style is better perceived in our proposed TTS system. An important observation is drawn from the results of the preference test is that the children preferred our proposed framework in 80% of the cases. It indicates that the target audience (i.e. children) are more satisfied with the quality of the speech synthesized.

5. CONCLUSIONS

We have presented a framework of story genre specific prosody models to improve the synthesis quality of the speech from a Hindi Story TTS system. We have developed two systems: baseline Story TTS system and proposed Story TTS system. The baseline system uses PCP based features for training the prosody models. The proposed story TTS system uses story genre specific prosody models. The story genre specific prosody models are developed using the textual (positional, contextual and phonological) and story-semantic features. The prosody models are developed using CART, FFNN and SVM, for modeling the duration, intonation, and intensity values of the syllables. The prosody models are evaluated by carrying out the objective and subjective measures. The results of the evaluation suggests that our proposed models outperforms the baseline and conveys the story style speech.

In the future work, the prosody modeling can be extended to other Indian Languages such as Bengali, Telugu, Tamil, Marathi, and Malayalam. Further, the prosody modeling can be carried out for different parts of the story such as introduction, main part, and moral. In this work, story-semantic features such as POS tags, discourse, and emotion are examples of supervised features, either obtained automatically from a tagger or from manual annotation by the text experts. Therefore, we can further investigate the unsupervised features to model the prosody for a syllable at each story genre.

Acknowledgements

The authors would like to thank the Department of Information Technology, Government of India, for funding the project, *Development of Text-to-Speech synthesis for Indian Languages Phase II*, Ref. no. 11(7)/2011HCC(TDIL). The authors would also, like to thank all the participants volunteered for the listening tests.

Table 6: Performance of the prosody (duration, intensity, and intonation) models developed using PCP features is given within the bracket and outside the bracket the performance of the prosody models developed using PCP and story-semantic features.

	Model	Objective Measures			% predicted syllables with deviation			
		μ	σ	γ	5%	10%	15%	25%
Duration Model (μ and σ are in ms)								
	CART	42.2 (54.5)	43.8 (59.6)	0.68 (0.71)	30.9 (24)	41.7 (34.7)	62.2 (54.3)	69.7 (61.5)
	FFNN	39.6 (51.3)	42.9 (53.5)	0.71 (0.68)	34.5 (27)	44 (38.3)	64.9 (57.3)	72.3 (65.3)
	SVM	38.3 (48)	39.8 (50.1)	0.72(0.70)	34.4 (25)	43.4 (40.1)	64.9 (55.3)	73.9 (69.8)
Intensity Model (μ and σ are in dB)								
	CART	5.4 (5.9)	5 (4.5)	0.66 (0.66)	38.3 (36.3)	66.03 (65.8)	83.9 (84.5)	94.5 (95.6)
	FFNN	5.1 (5.2)	4.7 (5)	0.68 (0.68)	38 (33.4)	66.1 (63.6)	85.2 (83)	96 (93)
	SVM	4.9(4.7)	4.5 (5.2)	0.68 (0.7)	39.2(36.3)	67.7 (65.8)	85.8 (85.7)	96.3 (94.2)
Intonation Model (μ and σ are in Hz)								
Initial F0	**CART**	31.05 (36.4)	28.5 (30.7)	0.64 (0.61)	21.7 (20)	40.23 (38.12)	56 (53.3)	78 (77.2)
	FFNN	31.2 (34.8)	34.5 (33.5)	0.64 (0.63)	25.3 (24.65)	41.8 (40.3)	58.3 (56.8)	78.3 (78.4)
	SVM	30.2 (33.2)	30.7 (30.8)	0.65 (0.67)	25.4 (25.2)	42.7 (44)	61.3 (58.4)	80.2 (79.1)
Middle F0	**CART**	35.8 (38.6)	34.7 (35.4)	0.62 (0.63)	22.4 (15.6)	37.3 (33)	56.1 (50.2)	74.9 (74.5)
	FFNN	33.5 (35.7)	34.5 (34.8)	0.65 (0.64)	24.4 (17.5)	37.6 (35.6)	53.2 (52.4)	76.8 (76)
	SVM	32.1 (34.5)	38.4 (32.4)	0.65 (0.66)	23.6 (19.6)	39.7(36.9)	55.8 (53.6)	79.6 (75.2)
End F0	**CART**	44.1 (48.4)	45.7 (47.9)	0.59 (0.58)	21.5 (15.2)	37.9 (32.4)	47 (47.3)	72 (70.3)
	FFNN	41.9 (45.8)	44 (44.9)	0.63 (0.62)	23.2 (15.8)	41.7 (34.6)	50.2 (48.4)	75.7 (72.4)
	SVM	40.5 (44.6)	43.5 (44.2)	0.64 (0.62)	25.3 (18.3)	43.7 (34.8)	52.5 (49.3)	77.1 (71.4)

6. REFERENCES

[1] Janet E. Cahn, "The generation of affect in synthesized speech," *Journal of the American Voice I/O Society*, vol. 8, pp. 1–19, 1990.

[2] Oudeyer Pierre-Yves, "The production and recognition of emotions in speech: features and algorithms," *International Journal of Human-Computer Studies*, vol. 59, no. 12, pp. 157 – 183, 2003, Applications of Affective Computing in Human-Computer Interaction.

[3] Marc Schröder, "Expressive Speech Synthesis: Past, Present, and Possible Futures," in *Jianhua Tao; Tieniu Tan: Affective Information Processing*, pp. 111–126. Springer, London, 2009.

[4] Donn Morrison, Ruili Wang, and Liyanage C. De Silva, "Ensemble Methods for Spoken Emotion Recognition in Call-centres," *Speech Commun.*, vol. 49, no. 2, pp. 98–112, 2007.

[5] Mostafa Al Masum Shaikh, Antonio Ferreira Rebordao, and Keikichi Hirose, "Improving TTS synthesis for emotional expressivity by a prosodic parameterization of affect based on linguistic analysis," in *International Conference on Speech Prosody, 5th, Proceedings*, 2010, p. 4.

[6] Marc Schröder, "Approaches to emotional expressivity in synthetic speech," in *The Emotion in the Human Voice (Izdebski, K., ed.)*, vol. 3, pp. 307–321. Plural, San Diego, 2008.

[7] V. Ramu Reddy and K. Sreenivasa Rao, "Prosody Modeling for Syllable Based Text-to-speech Synthesis Using Feedforward Neural Networks," *Neurocomput.*, vol. 171, no. C, pp. 1323–1334, 2016.

[8] K. Sreenivasa Rao and B. Yegnanarayana, "Modeling syllable duration in Indian languages using neural networks," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2004, Montreal, Quebec, Canada, May 17-21, 2004*, 2004, pp. 313–316.

[9] K. Sreenivasa Rao and B. Yegnanarayana, "Modeling Durations of Syllables Using Neural Networks," *Comput. Speech Lang.*, vol. 21, no. 2, pp. 282–295, 2007.

[10] K. Sreenivasa Rao and B. Yegnanarayana, "Intonation Modeling for Indian Languages," *Comput. Speech Lang.*, vol. 23, no. 2, pp. 240–256, 2009.

[11] H. A. Patil, T. B. Patel, N. J. Shah, H. B. Sailor, R. Krishnan, G. R. Kasthuri, T. Nagarajan, L. Christina, N. Kumar, V. Raghavendra, S. P. Kishore, S. R. M. Prasanna, N. Adiga, S. R. Singh, K. Anand, P. Kumar, B. C. Singh, S. L. Binil Kumar, T. G. Bhadran, T. Sajini, A. Saha, T. Basu, K. S. Rao, N. P. Narendra, A. K. Sao, R. Kumar, P. Talukdar, P. Acharyaa, S. Chandra, S. Lata, and H. A. Murthy, "A syllable-based framework for unit selection synthesis in 13 Indian languages," in *Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), 2013 International Conference*, Nov 2013, pp. 1–8.

[12] A. W. Black and P. Taylor, "The Festival Speech Synthesis System: System Documentation," Tech. Rep. HCRC/TR-83, Human Communciation Research Centre, University of Edinburgh, Scotland, UK, 1997.

[13] Harikrishna D M and K. S. Rao, "Children story classification based on structure of the story," in *Advances in Computing, Communications and Informatics (ICACCI), 2015 International Conference on*, Aug 2015, pp. 1485–1490.

[14] P. Sarkar and K. S. Rao, "Data-driven pause prediction for synthesis of storytelling style speech based on discourse modes," in *Electronics, Computing and Communication Technologies (CONECCT), 2015 IEEE International Conference on*, July 2015, pp. 1–5.

[15] Parakrant Sarkar, Arijul Haque, Arup Kumar Dutta, Gurunath Reddy M., Harikrishna D. M., Prasenjit Dhara, Rashmi Verma, N. P. Narendra, Sunil Kr. S. B., Jainath Yadav, and K. Sreenivasa Rao, "Designing prosody rule-set for converting neutral TTS speech to storytelling style speech for Indian languages: Bengali, Hindi and Telugu," in *IC3*. 2014, pp. 473–477, IEEE Computer Society.

[16] C.D. Jones, A.B. Smith, and E.F. Roberts, ," in *Proceedings Title*. IEEE, 2003, vol. II, pp. 803–806.