Data-Driven Pause Prediction for Synthesis of Storytelling Style Speech based on Discourse Modes

Parakrant Sarkar and K. Sreenivasa Rao
School of Information Technology
Indian Institute of Technology Kharagpur
Kharagpur - 721302, West Bengal, India.
Email: parakrantsarkar@gmail.com, ksrao@iitkgp.ac.in

***Abstract*—In storytelling style, a storyteller generally uses prosodic variations with subtle speech nuances for the better apprehension of the listeners. It is achieved by emphasizing prominent words, using various emotions, mimicking voices and providing appropriate pauses. This work is a part of building the Story Text-to-Speech (TTS) [1] synthesis systems in Indian Languages, which aims at synthesizing the storytelling style speech from the neutral TTS. The neutral speech is converted to storytelling style by modifying the specific prosodic parameters (i.e. duration, pitch, tempo, intensity and pauses). The main contribution of this paper is to model the pause patterns present in storytelling style speech based on the modes of discourse: narrative, descriptive and dialogue to capture the story-semantic information. Analysis of pause patterns are carried out for children stories in Hindi language. We analyzed the pause patterns and classified pauses into three different categories: short, medium and long pauses for each mode of discourse. A three stage data-driven method is proposed to predict the position and duration of the pauses. We conducted objective test to evaluate the performance of the proposed method at each stage. Also, subjective evaluation is carried out on the final output of the Hindi Story-TTS system. The subjective evaluation connotes that the subjects have perceived an improvement in speech quality in terms of storytelling style.**

***Keywords*—Discourse modes, Storytelling Style, Pause prediction, Speech Synthesis, Phrasing, Pause duration.**

I. Introduction

Prosody plays a significant role in synthesizing natural and intelligible speech from Text-To-Speech (TTS) systems. Supra-segmental features (like duration, pitch, intensity, co-articulation pattern and pause pattern) are concocted to form prosody. This provides an inference of additional information in speech that is not present in the text. A trade off between the prosodic-control and naturalness (in speech) [2] is carried out to mitigate the synthesis artefacts. Also, it results in monotonous quality of the synthesized speech. In the context of TTS, pause prediction is an important step during speech synthesis. Down the line other prosodic models depend upon this information. Therefore, predicting the accurate position and duration of the pauses relatively improves naturalness as well as expressiveness of synthesized speech without inducing synthesis artefacts.

In the context of TTS, procedure of finding out where the synthesizer should insert the pauses is called phrase-break or pause prediction. Pausing in natural speech is carried to out to emphasize something relevant to the context or just to take a breath [3]. Also, providing appropriate pauses in the speech not only enhances the quality but also provides affect [4]. For Indian languages, synthesis of storytelling style is shown in [1] and [5]. In storytelling [6], a storyteller uses their voice in a variety of ways for the better apprehension of the listeners. This include mimicking character's voices, producing sound effects, using prosody to convey and invoke emotions, and providing appropriate pauses, thus creating an engaging and pleasant listening experience. Also, pauses are used for separating phrases, emphasizing keywords and emotion-salient words to introduce suspense and climax in the story.

Several studies carried on the prediction of pauses using various machine learning approaches like Hidden Markov Models (HMM) [7], Classification and Regression Trees (CART) [8], Feed Forward Neural Network (FFNN) [9], Maximum Entropy (ME) [10] and Bayesian approach [11]. Various style-specific pause predictions tasks are shown in [12] and [13]. In Klatt model [14], the Festival based TTS system [15] as well as the Mary TTS system [16] assign fixed duration to the pauses in an utterance. For Indian languages, in [9] phrase break prediction is performed for Bengali. In Telugu, a set of morpheme tag units [17] are identified manually, and used to model phrase breaks. Significance of word terminal (i.e. last syllable of the word) in phrase break prediction is shown in [18]. For Hindi, a three stage pause prediction model [19], is used to predict the position and duration of the pauses in storytelling style speech. Although pause prediction has been widely explored in Indian Language, but predicting accurate duration of pause has not received much attentions in storytelling style.

This paper presents a data-driven approach to model the position and duration of pauses (i.e. phrase breaks) in storytelling style speech for Hindi language. The hypothesis is that using the knowledge of various modes of discourse, story-semantic [1] information can be captured more prominently. In this work, our objective is to investigate the pause pattern in storytelling style speech based on discourse modes [20]. In discourse mode, various prosodic parameters (like pitch, duration, speaking rate, intensity) are studied and modeled in [21], but modeling pauses has not been explored. For analyzing the pause patterns in stories, we recorded the children stories form a professional female storyteller. The rudimentary analysis of the collected story-speech corpus shows a basic pattern in terms of pauses. To capture this pattern we, proposed a three stage pause prediction model. Given an utterance as an

input. In first stage, each word boundaries are classified as pause or non-pause using word-level features. At second stage, if there is a pause after a word, then that pause is classified into one of the three categories: long, medium and short using syllable-level features. Finally, for each category of the pauses, a regression predictor is trained to predict duration. Here, we have only considered the inter-utterance pauses. We are not modeling pause at the end of an utterance because punctuation marks are used for this purpose.

This paper is organized as follows. First, section II, details the story-speech corpus and analysis of pauses based on various modes of discourse. Section III then explains the proposed method for building pause prediction model. In Section IV, evaluate the proposed method by objective and subjective tests. Finally in section V, we consolidate the present work with conclusion, future work and acknowledgement.

II. Story Speech Corpus

A total of 100 story texts comprising of children's tales are collected from story books like Panchatantra and Akbar-Birbal. The average number of sentences in each story, approximately varies from 20 to 25. The corpus covers 1960 sentences with 24400 words. The stories are recorded in a noise free studio environment by a professional female storyteller. For maintaining the high quality in the collected story-speech corpus, continuous feedback is given to storyteller for improving the quality of narrated story. The duration of the corpus is about 3 hours approximately.

A. Analysis of Discourse Modes

There are various discourse modes such as narrative, descriptive, argumentative, explanatory and dialogue [20]. Based on the analysis of the utterances present in collected story corpus, we concluded that there are only three different kinds of discourse modes (i.e. narrative, descriptive and dialogue). From the preliminary analysis, it is also observed that different parts of story are narrated in different styles. These styles are based on the semantics present at that part of story. In general, most of the stories, begin with introducing the characters present in story, followed by various events related to the story and finally story will conclude with a moral. Most of the stories are fictional. While narrating a story, as shown in [20], narrative, descriptive and dialogue modes are more prominent. As the story progress, one event after another, narrative mode is used to depict the listener/reader about the actions taking place in story. The descriptive mode shows the various activities that the main character is experiencing. Dialogue mode is used for any type of conversation taking place between any two characters. Generally, a greater amount of the text comprises of narrative mode. A storyteller uses his/her skills to add various expressive registers at sentence-level while narrating a story.

For Hindi children stories text classifications are shown in [22] and [23]. Similar approached is followed for manually annotating the story-corpus based on the three discourse modes. At sentence-level, text of the story was entrusted by four native Hindi speakers on text classification. They have been trained separately and work independently in order to avoid any labeling bias. In order to make the task of the annotation more focused, various discourse modes are annotated from the point of view of the text. Each annotator's task is to label the sentence with one of the modes of discourse (i.e. descriptive, dialogue and narrative). Similarly, at sentence-level annotators are also instructed to label with one of the emotions (i.e. anger, sad, fear, happy). At story-level, each story is annotated with one of the story-class (i.e fable, legendary, folk-tales). Table I, shows the details of total number of sentences classified into various discourse modes. Also, Table II, shows the total number of the sentences labeled with one of the emotions. Similarly, Table III displays, the total number of the stories labeled with one of the story-class. The inter-annotator agreement is given by Fleiss Kappa (κ). The κ values above 0.65 or so can be considered to be substantial. The κ values are 0.73, 0.67 and 0.69 for the three annotation tasks: discourse, emotion and story-class respectively.

TABLE I. Sentence-level Discourse modes Classification

	Descriptive	Narrative	Dialogue
#sentences	547	1134	279

TABLE II. Sentence-level Emotion Classification

	Happy	Sad	Fear	Angry
#sentence	245	172	110	90

TABLE III. Story-class classification

	Fable	Legendary	Folk-tales
#stories	35	19	46

B. Analysis of Pauses

The position and duration information of pauses are marked automatically by following the force-alignment methodology. The pause marking are also manually adjusted to mitigate the errors that has incurred during the automatic process. At the end of the process each word boundary in text corpus is labeled as pause or non-pause. Fig.1 shows histogram plot of durations of the pauses present in story-corpus. There are 3804 total number of pauses in the corpus. It is observed that distribution of pauses are more with duration value ranges between 50 to 400 ms. By analyzing distribution of the duration of pauses we categorised the pauses into three types: (i) *Long pause* ($>$250 ms) (ii) *Medium pause* ($150 - 250$ ms) and (iii) *Short pause* ($<$150 ms). Also, it is noticed that pauses with duration value less than 50 ms are not relevant. Since, listeners are not be able to perceive it as pause. So, we ignored all the pauses with duration value less than 50 ms in training the regression predictor. The Table IV shows the mean and standard deviation of different types of pauses present in the corpus. There are 17% long, 25% medium, 38% short pauses, out which 20% of the pauses are ignored.

TABLE IV. Pause Pattern in the story-speech corpus

Pause Type	Mean(ms)	StdDev(ms)	%in original
long pause	455.07	125.99	17
medium pause	210.12	32.33	25
short pause	92.97	29.81	38

C. Story Text-to-speech system

The aim of story TTS system is to synthesize storyteller speech from the neutral TTS system for a given story text

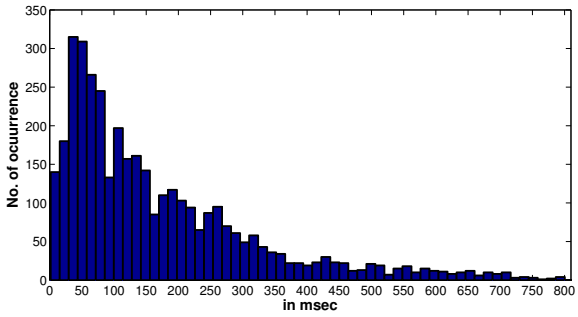

Fig. 1. Histogram of pause duration in story-speech corpus

as input. A set of additional modules are added with the existing neutral TTS systems as explained in [1]. The modules are (i) Story-specific emotion detection (SSED) module, (ii) Story-specific prosody generation (SSPG) module and (iii) Story-specific prosody incorporation (SSPI) module. The entire process of story synthesis framework is shown in Fig.2. The raw story text is given as input to a neutral TTS to synthesize a neutral speech. The text is also parsed by SSED module to detect the story-specific emotions associated to each phrases within a sentence. The SSPG module consists of prosody rule-sets for each of the story-specific emotions. Based on emotions detected, appropriate prosodic rules are activated, and the desired modification factors associated with various prosodic parameters (like pitch, duration, intensity, tempo and pause pattern) are specified by SSPG module. The modification factors are incorporated on the synthesized neutral speech by SSPG module. In this work, we added story-specific pause prediction (SSPP) module, which is an extension of the SSPG module. It predicts the proper position and duration of pauses in the synthesized speech based on story-semantic knowledge.

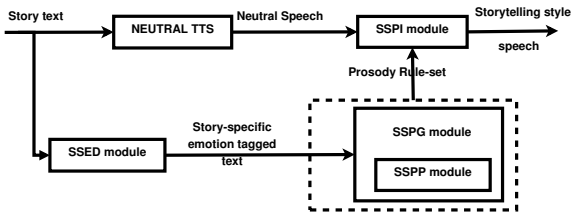

Fig. 2. Block diagram of Story Synthesis Framework

III. Building SSPP Module: Pause Prediction Model

This section discusses the procedure followed in building the SSPP module for pause prediction. A three stage pause prediction model is proposed as shown in the Fig. 3. In the first stage, goal is to model the position of pauses within an utterance using word-level features. The seconds stage deals with the classification of pauses into one of the three different types (i.e. long, medium and short) based on syllable-level features. In final stage, a regression predictor is trained to predict the duration of pause based on its type.

The CART trees are modeled with a set of features for each stage of SSPP module. From the corpus, 90% of the stories are

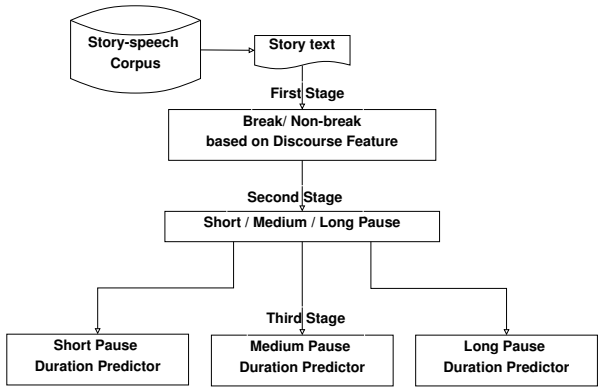

Fig. 3. Three stage pause prediction model [19]

used for training and 10% is used for testing. The models are trained by following a 10-fold cross validation technique. We have used the Wagon tool[1] for building CART tree present in the speech tools. A five-gram window is followed in our current study to extract the features. The five-gram window is represented by the previous two words, current word and following two words. For testing, we are using the stories that are not used for training. The prior information such as discourse and linguistic (POS, terminal syllable) information are readily available (manually annotated) for the stories used for testing.

The set of features that we have considered for modeling at each stage are as follows: (i) *First stage*: position of the current word from the beginning and ending of an utterance, total number of syllables/phones in words, total number of words/syllables/phones in an utterance, parts-of-speech (POS)[2] of word, emotion (sad, anger, happy, fear) of the word, discourse mode of the utterance, whether the word is a content or functional word and class of the story (fable, legendary, folk-tales). (ii) *Second stage*: Terminal syllable [18] of word, total number of phones in the terminal syllable, position of the vowel in the terminal syllable, number of segments(i.e. consonants) before and after the vowel in terminal syllable. (iii)*Third Stage*: Positional, contextual, morphological and linguistic (terminal syllable) features at syllable-level as discussed in [24].

IV. Evaluation of Proposed Model

The proposed three stage pause prediction model is evaluated by conducting both objective and subjective tests. We calculated the F-1 measure [25] to evaluate our models at first and second stage. The F-1 measure is the harmonic mean of recall and precision. A good model gives a higher F-1 score close to 1.00. Ideally, a model should provide high recall and high precision. High recall guarantees that the model, predicts as many break as there are in the actual. Similarly, high precision signifies that the model, predicting the pauses in the wrong places should be less. The third stage is evaluated using: average prediction error (μ), standard deviation (σ) and correlation coefficient ($\gamma_{x,y}$).

[1]http://www.cstr.ed.ac.uk/projects/speech_tools/

[2]Automatic POS tagger is used, developed by IIT Kharagpur for Hindi. http://nltr.org/snltr-software/

A. Objective Evaluation

We built the model at each stage by performing 10-fold cross validation, in which the data is divided into 10 different sets. Nine sets are used for training the model and one set is used for testing. For each model, accuracy is given by the average F-1 measure across all 10 sets.

1) Accuracy of first stage of pause prediction model: The average recall, precision and F-1 measure at first stage is given in the Table V. Two different types of models are built: first by considering all the features that are explained in section III and second considering all the features excluding the discourse feature. The values shown within and outside of parenthesis are obtained by the first and second model respectively. An analysis of the performance of pause shows an absolute increase of 0.02 (F-1) and 0.03 (precision). Similarly, performance of the non-pause shows an absolute decrease of 0.02 (F-1). It can be inferred, by adding discourse as an additional feature adds complementary information to the information captured by the first model. This complementary information helps in capturing the story-semantic knowledge.

TABLE V. PERFORMANCE MEASURES FOR PAUSE (P) AND NON-PAUSE (NP) PREDICTION

	Recall	**Precision**	**F-1 Score**
NP	0.86(0.89)	0.94(0.94)	0.89(0.91)
P	0.74(0.68)	0.784(0.81)	0.76(0.74)

2) Accuracy of second stage of pause prediction model: The average accuracy (i.e. F-1 measure) of the model in second stage are calculated, as shown in Table VI. The average F-1 measures for long is 0.69, medium is 0.58 and short is 0.60. The performance of the models for classifying medium and short pauses are less compared to long pauses. One possible explanation here is that, syllable-level features may not be sufficient enough to categories the pause, leading to a weak model.

TABLE VI. SHORT, MEDIUM AND LONG PAUSE PREDICTION ACCURACY

	Recall	**Precision**	**F-1 Score**
long pause	0.75	0.64	0.69
medium pause	0.60	0.58	0.58
short pause	0.56	0.65	0.60

3) Accuracy of third stage of pause prediction model: At this stage, three different CART models are trained for short, medium and long pauses. The trees are evaluated by calculating objective measures [24]. These include average prediction error (μ), standard deviation (σ) and correlation coefficient ($\gamma_{x,y}$) between the actual and predicted pause duration values.

The objective measures in terms of the average of actual pause duration values $\bar{x}$, and average of predicted pause duration values $\bar{y}$ are shown in TableVII. From the table, we can infer that the average prediction error for long pause is significantly high compared to medium and short pause. The high prediction error is reasonable as the average actual pause duration for the long pause is much higher 347.96 ms. The high prediction error of 78.46 ms does not significantly change the nature of the pause as medium or short. The correlation coefficient ($\gamma_{x,y}$) for each of the CART models are better.

TABLE VII. PERFORMANCE OF CART MODELS FOR LONG(L), MEDIUM(M) AND SHORT(S) PAUSES USING OBJECTIVE MEASURES (μ, σ AND $\gamma_{x,y}$)

	$\bar{x}$ **(in ms)**	$\bar{y}$ **(in ms)**	μ **(in ms)**	σ **(in ms)**	$\gamma_{x,y}$
L_CART	347.96	372.92	78.46	57.31	0.73
M_CART	208.43	199.30	26.19	17.02	0.67
S_CART	87.33	84.99	24.51	17.81	0.72

B. Subjective Evaluation

Listening test is conducted to show the significance of the proposed pause prediction model. The test is performed on 20 synthesized sentences before and after incorporating the three stage pause prediction model in SSPP module. 20 research scholars in the age group of 20-35 participated in the test. To evaluate the naturalness and intelligibility of the synthesized speech, degradation mean opinion scores (DMOS) [26] are calculated. Each subjects have to give a score on a five-point scale (1: very poor, 2: poor, 3: fair, 4: good and 5: excellent). Table VIII shows the DMOS score for 20 sentences before (using festival default pause model) and after (using proposed three stage pause prediction model) modifications. The statistical significance for the pairs of DMOS score are tested using hypothesis testing [27]. The levels of confidence achieved for all the transitions are high (99.5% for naturalness and 95% for intelligibility respectively) using Student-T distribution. The results of subjective test shows that three stage pause prediction model performs better than festival default pause model. Also, the evaluation indicates that subjects have perceived a noticeable difference in the synthesized speech by incorporating the proposed model.

TABLE VIII. MEAN OPINION SCORE FOR BASLINE AND PROPOSED METHOD

Pause Model	**Naturalness**	**Intelligibility**
Festival Default	2.95	2.80
Proposed	3.25	3.05

V. CONCLUSION AND FUTURE WORK

In this paper, an extension of the SSPG module i.e. SSPP module for Hindi Story TTS system is carried out. In the SSPP module, we proposed three stage data-driven pause prediction model to learn the pause pattern present in storytelling style speech based on three discourse modes. Also, we concluded from the results of the first stage, the discourse modes can capture story-semantic information. We analyzed the pause patterns using the story-speech corpus. Based on analysis, three different categories (i.e. short, medium and long pauses) are considered. The goal of the work is accurately determine the position and duration of pauses in an utterance at each word boundary. First stage properly identifies the position of pauses within an utterance. In the second stage, each pause is classified into three different types. In the third stage, for each type of pause, a regression predictor is trained to predict the duration value. The CART models are evaluated both by conducting objective and subjective measures. The results of perceptual evaluation indicates that the proposed method is effective in imposing pauses in synthesized speech utterance.

Possible extensions to the current work are as follows. In the second stage, new features can also be explored, to

improve performance of the model. Various prosodic parameter analysis (pitch, duration, intensity, tempo) based on discourse modes can be carried out for storytelling style speech. Our current study is based on the synthesizing one utterance at a time. Further studies can be performed to analyze the pause patterns present at paragraph level [28] for synthesizing a story. The use of supervised linguistic features like (POS, terminal syllable etc) are not readily available in the case of Indian Languages. Manual labeling of the text with these linguistic information is quite hectic and time consuming. Studies can be carried out by following the work as shown in [29] and [30]. We can use unsupervised features that can capture the co-occurrence statistics of word distribution. Hence, a framework for extracting the continuous valued word representation from an unlabeled text corpus can also be proposed and integrated with the existing Hindi Story TTS. Apart from Hindi, the current pause prediction study can be extended to other Indian languages such as Bengali, Telugu, Tamil, Marathi, Malayalam. In addition to CART, different nonlinear classifiers can be explored.

VI. 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 for the listening tests.

References

[1] P. Sarkar, A. Haque, A. Dutta, G. Reddy, M. Harikrishna, P. Dhara, R. Verma, P. Narendra, B. S. Sunil, J. Yadav, and K. S. Rao, "Designing Prosody Rule-set for Converting Neutral TTS Speech to storytelling style speech for Indian Languages: Bengali, Hindi and Telugu," in *Seventh International Conference on Contemporary Computing (IC3), Noida*, Aug 2014, pp. 473–477.

[2] M. Schroder, "Emotional Speech Synthesis–a Review," in *Proceedings of Eurospeech*, Aalborg, Denmark, 2001, pp. 561–564.

[3] N. Braunschweiler and L. Chen, "Automatic detection of inhalation breath pauses for improved pause modelling in HMM -TTS," in *8th ISCA Speech Synthesis Workshop*. Barcelona, Spain: ISCA, August 31 - September 2, 2013 2013, pp. 1– 6.

[4] C. Alm, "Affect in Text and Speech," Tech. Rep., 2008.

[5] R. Verma, P. Sarkar, and K. Rao, "Conversion of neutral speech to storytelling style speech," in *Advances in Pattern Recognition (ICAPR), 2015 Eighth International Conference on, ISI Kolkata, India*, Jan 2015, pp. 1–6.

[6] M. Theune, K. Meijs, D. Heylen, and R. Ordelman, "Generating Expressive Speech for Storytelling Applications," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 14, no. 4, pp. 1137–1144, 2006. [Online]. Available: http://dblp.uni-trier.de/db/journals/taslp/taslp14.html

[7] P. Taylor and A. W. Black, "Assigning phrase breaks from part-of-speech sequences," *Computer Speech & Language*, vol. 12, no. 2, pp. 99–117, 1998.

[8] K. Yoon, "A Prosodic Phrasing Model for a Korean Text-to-speech Synthesis System ," *Computer Speech & Language*, vol. 20, no. 1, pp. 69 – 79, 2006.

[9] K. Ghosh and K. Sreenivasa Rao, "Data-Driven Phrase Break Prediction for Bengali Text-to-Speech System," in *Contemporary Computing - 5th International Conference, IC3 2012, Noida, India, August 6-8, 2012. Proceedings*, ser. Communications in Computer and Information Science. Springer Berlin Heidelberg, 2012, vol. 306, pp. 118 – 129.

[10] S. Kim, J. Lee, B. Kim, and G. G. Lee, "Incorporating second-order information into two-step major phrase break prediction for korean," in *INTERSPEECH 2006 - ICSLP, Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, September 17-21, 2006*, 2006.

[11] P. Zervas, M. Maragoudakis, N. Fakotakis, and G. Kokkinakis, "Bayesian induction of intonational phrase breaks," Eurospeech, 2003.

[12] A. Parlikar and A. Black, "Modeling Pause-Duration for Style-Specific Speech Synthesis," in *INTERSPEECH*. ISCA, 2012.

[13] A. Parlikar and A. W. Black, "A grammar based approach to style specific phrase prediction," in *in Interspeech*, 2011, pp. 2149–2152.

[14] D. Klatt, "The KLATTALK Text-to-Speech Conversion System," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '82.*, vol. 7, May 1982, pp. 1589–1592.

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

[16] M. Schröder, "The German Text-to-Speech synthesis system MARY A tool for research, development and teaching," in *International Journal of Speech Technology*, 2001, pp. 365–377.

[17] N. S. Krishna and H. A. Murthy, "A New Prosodic Phrasing Model for Indian Language Telugu," in *INTERSPEECH*. ISCA, 2004.

[18] A. Vadapalli, P. Bhaskararao, and K. Prahallad, "Significance of word-terminal syllables for prediction of phrase breaks in Text-to-Speech systems for Indian Languages," in *8th ISCA Speech Synthesis Workshop*. Barcelona, Spain: ISCA, August 31– September 2, 2013 2013, pp. 189 – 194.

[19] P. Sarkar and K. S. Rao, "Data-driven Pause Prediction for Speech Synthesis in Storytelling Style Speech," in *Communications (NCC), 2015 Twenty First National Conference on, IIT Bombay, India*, Feb 2015, pp. 1–5.

[20] J. Adell, A. Bonafonte, and D. E. Mancebo, "Analysis of prosodic features towards modelling of emotional and pragmatic attributes of speech," *Procesamiento del Lenguaje Natural*, vol. 35, 2005.

[21] R. Montao, F. Alas, and J. Ferrer, "Prosodic analysis of storytelling discourse modes and narrative situations oriented to Text-to-Speech synthesis," in *8th ISCA Workshop on Speech Synthesis*, Barcelona, Spain, August 2013, pp. 191–196.

[22] D. M. Harikrishna and K. S. Rao, "Classification of Children Stories in Hindi Using Keywords and POS Density," in *International Conference on Computer Communication and Control (IC4)*, September 2015.

[23] D. M. Harikrishna and K. S. Rao, "Children Story Classification based on Structure of the Story," in *Fourth International Symposium on Natural Language Processing (NLP'15)*, August 2015.

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

[25] C. J. V. Rijsbergen, *Information Retrieval*, 2nd ed. Newton, MA, USA: Butterworth-Heinemann, 1979.

[26] S. King, "Degradation MOS and word error rate for text to speech synthesis systems." private communication.

[27] R. Hogg and J. Ledolter, *Engineering statistics*, ser. Mathematics & statistics. Macmillan, 1987.

[28] K. Prahallad, A. R. Toth, and A. W. Black, "Automatic Building of Synthetic Voices from Large Multi-Paragraph Speech Databases," in *INTERSPEECH*. ISCA, 2007, pp. 2901–2904.

[29] O. Watts, J. Yamagishi, and S. King, "Unsupervised Continuous-Valued Word Features for Phrase-Break Prediction without a Part-of-Speech Tagger," in *INTERSPEECH*. ISCA, 2011, pp. 2157–2160.

[30] A. Vadapalli and K. Prahallad, "Learning continuous-valued word representations for phrase break prediction," in *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, 2014, pp. 41–45.