Analysis and Modeling Pauses for Synthesis of Storytelling 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*—Generally in Text-to-Speech synthesis (TTS) systems, pause prediction plays a vital role in synthesizing natural and expressive speech. In storytelling style, pauses introduce suspense and climax by emphasizing the prominent words or emotion-salient words in a story. The objective of this work is to analyze and model the pause pattern to capture the story-semantic information. The purpose of this paper is to define a stepping stone towards developing a Story TTS based on modes of discourse. In this work, we base our analysis of the pauses in Hindi children stories for each mode of discourse: narrative, descriptive and dialogue. After grouping the sentences into modes, we analyse the pause pattern to capture the story-semantic information. A three stage data-driven method is proposed to predict the location and duration of pauses for each mode. Both the objective as well as subjective test are conducted to evaluate the performance of the proposed method. The subjective evaluation indicates that subjects appreciates the quality of synthesized speech by incorporating the proposed model.**

***Keywords*—Storytelling style, Pause prediction, Phrasing, Pause Duration, Discourse mode, Speech synthesis.**

I. Introduction

Mainstream data-driven Text-To-Speech (TTS) systems bridge the gap between the text and speech by using an intermediate linguistic representations. This paper will focus on exploiting one of intermediate features associated with the prediction of pauses. A pause prediction problem can be thought as binary classification problem: given a text input, the model should classify each word boundary as being pause or non-pause. In the context of TTS, procedure of finding out where the synthesizer should insert the pauses is called phrase break or pause prediction. It is the first step in modeling the prosody during speech synthesis. Down the line other prosody models such as duration, pitch and intensity uses the information, provided by the pause prediction model. Therefore, if the pause model provides wrong information, it decrease the performance of the other models that are using this information to predict the target prosody. Hence, predicting the accurate position and duration of the pauses relatively improves naturalness as well as expressiveness of the synthesized speech without inducing synthesis artefacts.

Several earlier studies carried on the prediction of pauses using various approaches such as Hidden Markov Models (HMM) [1] or Classification and Regression Trees (CART) [2] or Feed Forward Neural Network (FFNN) [3], Maximum Entropy (ME) [4] and Bayesian approach [5]. The style specific pause prediction is shown in [6]. In [7], authors shows the relevance of speaker specific features for pause prediction. Pauses are categorized into three different types based on the TOBI scheme [8]. In [9], analysis based on the length of pauses in speech is carried out for various speakers in different contexts. It shows the distribution of pauses in speech utterance affects the meaning and its perception. The pause duration is also a reliable means of discriminating between lexical ambiguity of words as shown in [10]. In the Klatt model [11], the Festival based TTS system [12] as well as the Mary TTS system [13] assign a fixed duration to the pauses. The prediction of the position of a pause has been widely explored in speech synthesis for various styles of speech [14] but generating the appropriate duration of the pauses has not received much attentions.

In Indian languages, various works addressed the pause prediction problems are as follows: For Telugu language, a set of morpheme tag units [15] are identified manually and used to model phrase breaks. For Bengali language, phrase break prediction is carried in [3]. Significance of word terminal (i.e. last syllable of the word) in phrase break prediction is shown in [16]. A three stage pause prediction model [17], for Hindi language is used to predict duration as well as the location of the pauses in an utterance. In storytelling style speech, pause prediction has not been widely explored in Indian Languages. Also, predicting the accurate duration of pauses has not received much attention.

In storytelling style speech [18], a storyteller uses a combination of expressive, engaging and entertaining style speech. The act of pausing during narration of a story not only enhances the audience's understanding of the story, but also builds anticipation. Also, pauses are used for separating phrases, emphasizing keywords and emotion-salient words to introduce suspense and climax in the story. For analyzing the story, we recorded the children stories (i.e. story-speech corpus) form a professional female storyteller. It is observed that in children stories, different parts of a story are narrated in different styles. These styles are governed by the story-semantics as shown in [19] and [20], present at that part of the story. Typically a children stories begin with introducing the characters present in the story, followed by the various events related to the characters and finally the story will conclude with a moral.

This study presents a data-driven approach to model position and duration of pauses for Hindi language. It is know

that, while narrating a story, narrator is not self-experiencing the emotions, s/he is trying to simulate the emotions in order to engage the listeners of the story. A storyteller uses various prosodic variations based on the story-semantic information present in that part of the story. This motivated us to analyze, the pauses present in the storytelling style speech based on story-semantics. The story-semantics are analyzed based on the discourse modes is shown in [21]. In discourse mode, various prosodic parameters (such as pitch, duration, speaking rate, intensity) are studied and modeled in [22], but modeling pause has not been explored. Our objective is to investigate the pause pattern in storytelling style speech based on discourse modes. As shown in [17], an extension to the three stage pause prediction model is proposed to model the pauses. Initially, the sentences in a story are divided into one of the three modes of discourse. For each mode of discourse, three stage model is used. In first stage, word boundary is classified as pause or non-pause using word-level features. At second stage, pause is classified into one of the three categories: long, medium and short pause using syllable-level features. Finally, for each category of pause, a regression predictor is trained to predict duration. Here, we have only considered the pauses which occur in between an utterance. We are not modeling pause at the end of an utterance.

II. Story Speech Corpus

The text comprising of childern's stories are collected from the story books like Panchatantra and Akbar-Birbal. A total of 100 story texts are collected. The number of sentences in each story approximately varies from 20 to 25. The story text 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 total duration of the corpus is about 3 hours.

A. Analysis of Story Discourse Modes

There are various discourse modes such as narrative, descriptive, argumentative, explanatory and dialogue. In context of storytelling style speech, narrative, descriptive and the dialogue are relevant as shown in [21] and [22]. Most of the stories in the corpus, begin with introducing the characters present in story, followed by various events related to the story and finally story will conclude with a moral. A storyteller uses various discourse modes while narrating a story based on story-semantic information. Generally, narrative mode is used to depict the listener/reader about the actions taking place in story. The descriptive mode depicts 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.

The entire story-texts are annotated manually based on the three discourse modes. At sentence level, text of the story was entrusted by four experts on text classification. The task of the annotator is to classify the sentences as descriptive, dialogue and narrative mode. Table I, shows the details of total number of sentences classified into various discourse modes. The inter-annotator agreement is given by Fleiss Kappa (κ) $= 0.70$ and can be considered to be substantial. Each story is manually classified into three classes (i.e. fable, legendary, folk-tales) with $\kappa = 0.72$. Also, sentences in each story are manually tagged for story-specific [19] emotion (anger, sad, fear, happy) with $\kappa = 0.774$.

TABLE I. Total #sentences in discourse modes

	Descriptive	Narrative	Dialogue
#sentences	547	1134	279

B. Analysis of Pauses

Following the similar line of work as shown in [17], the duration of the pauses are classified into three types as present in a story corpus. These are (i) *Long pause* (>250 ms) (ii) *Medium pause* ($150 - 250$ ms) and (iii) *Short pause* (<150 ms). A pause with a duration value (<50 ms) is not perceived as a pause by the listeners. So, pauses with a duration value less than 50 ms are not considered for building the models. The Table II shows the three different types of pauses present among the various discourse modes in story-speech corpus. The mean and standard deviation for different pause types are shown in column 2 and 3. The column 4 shows the various percentages of the pauses present in each category. Also, 20% ($<50ms$) of the pauses are not considered for building the models.

Manually annotating the entire story-speech corpus, for pauses is quite tedious. Hence, we performed an force alignment of the speech wave file with the text prompts using the HMM tool [23] by building a CLUSTERGEN [24] voice within the Festival [12] and Festvox [25] frameworks. This provides the information regarding the position of pauses and its duration values. The Figure 1, Figure 2, and Figure 3 shows the histogram plots of pauses based on descriptive, narrative and discourse modes respectively.

TABLE II. Pause Pattern in Story-speech Corpus

Description: Discourse Mode			
Pause Type	**Mean (ms)**	**StdDev (ms)**	**% in original**
Long Pause	447.59	240.18	6.15
Medium Pause	194.57	30.75	6.94
Short Pause	128.68	59.73	6.96
Narrative: Discourse Mode			
Pause Type	**Mean (ms)**	**StdDev (ms)**	**% in original**
Long Pause	413.00	179.89	13.14
Medium Pause	197.61	27.89	11.48
Short Pause	93.97	30.19	23.84
Dialogue: Discourse Mode			
Pause Type	**Mean (ms)**	**StdDev (ms)**	**% in original**
Long Pause	492.28	211.00	4.85
Medium pause	196.44	27.22	2.44
Short Pause	92.24	28.97	4.20

There are total of 3804 pauses present in the corpus which comprises of pauses in each mode 987:descriptive, 2283:narrative and 537:dialogue. From the histogram plots, it is observed that the occurrence of pauses are more with a duration value ranges between 50 to 500 ms across all three mode of discourse.

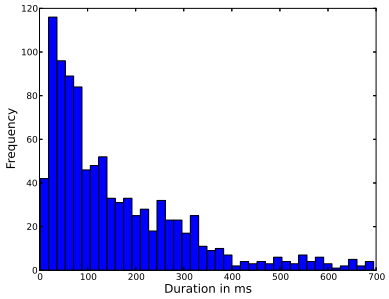

Fig. 1. Distribution of Pause Duration based on Description (Discourse type)

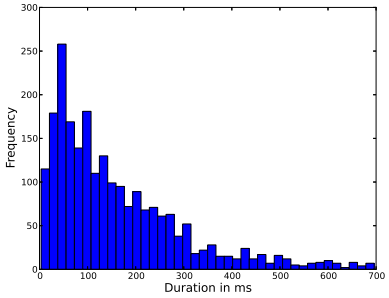

Fig. 2. Distribution of Pause Duration based on Narrative (Discourse type)

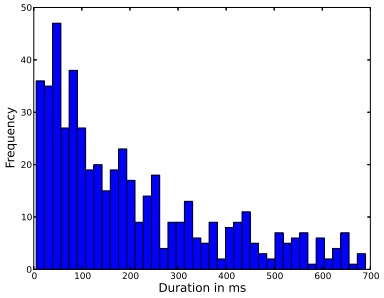

Fig. 3. Distribution of Pause Duration based on Dialogue (Discourse type)

III. Building Pause Prediction Model

This section discusses about the procedure followed in building the model for pause prediction. The sentences present in a story are divided into three categories based on discourse modes (i.e. descriptive, narrative and dialogue) as shown in the Figure 4. Given a story input, we will classify each sentence present in the story into one of three discourse modes. For each discourse mode, we follow a three stage pause prediction model as shown in the Figure 5. At first stage, objective is to classify each word boundary as break (i.e. pause) or non-break (i.e. non-pause) within an utterance. For this purpose, as set of word-features are exploited from the corpus. At second stage, pauses are classified as one of the three different categories (i.e. long, medium and short pauses) by exploiting the syllable-level features. Finally, at third stage we model a regression predictor to predict the duration of pauses.

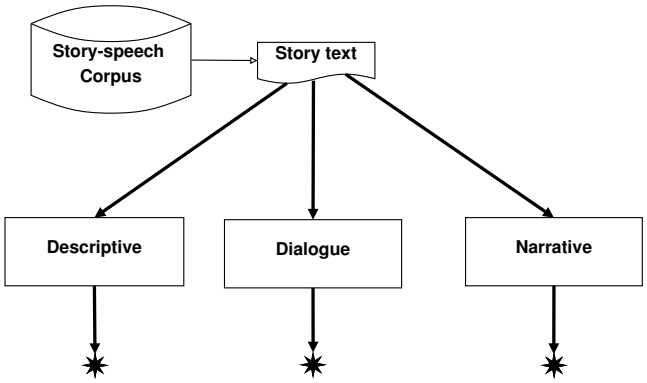

Fig. 4. Classifying Story text into three modes of Discourse

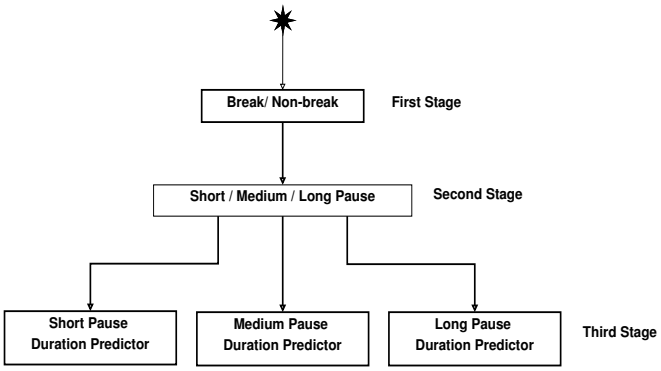

Fig. 5. Three stage pause prediction model [17]

At each stage CART trees are used to modeled with a specific set of features to attain the desired goals. From the story-speech corpus, 90% of the stories are used for training and 10% is used for testing. The models are trained by following a 10-fold cross validation technique. In this technique, entire training set is divide into 10 sets. Out of 10 sets, nine sets are used for training the model and one set is used for testing. In a similar manner, the procedure is repeated for 10 times and the average performance of the models are reported. The CART trees are built using the Wagon tool[1] present in the *Speech Tools* [25]. The reasoning behind choosing the CART tree is that, the models can be easily integrated to the existing Story TTS systems [19] framework. An empirically determined *five-gram* window is followed in our current study to capture the contextual information. It 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, linguistic (POS, terminal syllable) and story-specific informations are readily available (i.e. manually annotated) for the stories used for testing.

For the sake of completeness, we describe various features associated to each stage of building the models are as follows:

1) First stage of pause prediction model
   a) Positional:
      - Position of the current word from the beginning and ending of the utterance.

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

- Total number of words in the utterance.

b) Structural:
- Total number of phones in the current, previous and following two words.
- Total number of syllables in the current, previous and following two words.
- Total number of phones in the utterance.

c) Morphological:
- POS [2] of current, previous and following two words.
- Phonetic strength (stressed or not) of current word.

d) Story-specific
- Emotion (sad, anger, happy, fear) of the current word in the utterance.
- Class of the story (fable, legendary, folktales)
- Whether the word is a content or functional word.

2) Second stage of Pause Prediction Model.

a) Positional
- Total number of phones in the terminal syllable of the current, previous and following two words.
- Position of the current word from the beginning and ending of the utterance.

b) Morphological:
- Terminal syllable of the current, previous and following two words.

c) Structural
- Position of the vowel in the terminal syllable
- Number of segments (i.e. consonants) before and after the nucleus (i.e. vowel) in the terminal syllable.

3) Third stage of pause prediction model

a) Positional
- Syllable position in the word.
- Syllable position from the beginning and end of an utterance.

b) Structural
- Syllable identity as shown in [26]: segments of the syllable i.e. consonants and vowels.
- Position of vowel in the syllable.
- Number of segments (i.e. consonants) before and after the nucleus (i.e. vowel) in the syllable.

In the testing, the stories with discourse annotated labels are given input to the pause prediction model. The model will segregate the sentences based on labels for various discourse modes. For each of the sentence belonging to a specific mode, three stage pause prediction model is followed to predict the position and duration of pauses.

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

IV. Evaluation of the three stage pause prediction model

This section focuses on evaluating the three stage pause prediction model using objective measures. The objective measure manifests the percentage of correct prediction for the pause and non-pause. At first and second stage, models are evaluated by calculating F-1 measure [27]. The $F1$ measure is the harmonic mean of recall and precision. A good model gives a higher $F1$ score close to 1.00. Ideally, a model should provide high recall and high precision in order get a higher $F1$ measure. High recall guarantees that the model, predicts as many pauses as there are in the actual (test data). Similarly, high precision guarantees 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 ($\gamma_{x,y}$).

A. Objective Evaluation

1) Accuracy of first stage of pause prediction model: The model at this stage is built by performing 10-fold cross validation. For each model, we calculate the F-1 measure for both pause and non-pause. The accuracy is given by average F-1 measure across all 10 sets. The average F-1 measure, at first stage is shown in Table III for three different modes of discourse. The column $2-4$ shows the recall, precision and F-1 score respectively.

TABLE III. Recall, Precision and $F-1$ measure for Pause and Non-pause prediction

Descriptive			
	Recall	**Precision**	**F-1 Score**
Non-pause	0.978	0.837	0.902
Pause	0.454	0.88	0.60
Dialogue			
	Recall	**Precision**	**F-1 Score**
Non-pause	0.952	0.872	0.91
Pause	0.569	0.793	0.663
Narrative			
	Recall	**Precision**	**F-1 Score**
Non-pause	0.953	0.856	0.902
Pause	0.552	0.806	0.655

Based on the results obtained in Table III, among the three modes of discourse, dialogue mode shows better performance (F1 score) in predicting the position of pauses followed by narrative and descriptive modes. This shows the relevance of the pauses at dialogue mode in storytelling style speech. Also, the performance of the systems are quite less. One of the possible explanation here is that there is less availability of the training data.

2) Accuracy of second stage of pause prediction model: Similarly, 10-fold cross validation is performed at second stage. We calculated the average F-1 measure for the model across all 10 sets. The F-1 measure is shown in the Table IV for three different discourse modes. The column $2-4$ shows the recall, precision and F-1 score. In both descriptive as well as dialogue mode, long pause type has the maximum F-1 measure as compared with other medium and short pause types. Similarly in narrative mode, short pause type has higher F-1 measure as compared to other pause types.

TABLE IV. SHORT, MEDIUM AND LONG PAUSE PREDICTION ACCURACY (F-1 SCORE)

Descriptive			
Pause Type	**Recall**	**Precision**	**F-1 Score**
long	0.56	0.46	0.51
medium	0.48	0.39	0.43
short	0.56	0.47	0.50
Dialogue			
	Recall	**Precision**	**F-1 Score**
long	0.50	0.72	0.59
medium	0.53	0.65	0.58
short	0.40	0.55	0.46
Narrative			
	Recall	**Precision**	**F-1 Score**
long	0.37	0.48	0.42
medium	0.53	0.46	0.49
short	0.73	0.46	0.56

3) Accuracy of third stage of pause prediction model: At this stage, for each of three different categories of pauses i.e. short, medium and long, we built three CART trees to predict the duration. Each of the CART trees are evaluated based on the prediction accuracy. We followed the similar approach as shown in [26], such as average prediction error (μ), standard deviation (σ) and correlation coefficient ($\gamma_{x,y}$) between the actual and predicted duration values. The formulas used to compute objective measures are 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 duration values of pauses respectively, and e_i is the error between the actual and predicted duration values. The deviation in error is d_i and N is the number of observed duration values of the syllables. The correlation coefficient is given by

$$\gamma_{X,Y} = \frac{V_{X,Y}}{\sigma_X . \sigma_Y}, \tag{3}$$

TABLE V. PERFORMANCE OF CART TREES FOR LONG, MEDIUM AND SHORT PAUSES USING OBJECTIVE MEASURES (μ,σ AND$\gamma_{x,y}$)

Descriptive					
	$\bar{x}$ **(in ms)**	$\bar{y}$ **(in ms)**	μ **(in ms)**	σ **(in ms)**	$\gamma_{x,y}$
CART_long	468.24	486.93	80.89	100.92	0.78
CART_medium	201.03	189.97	14.39	12.95	0.76
CART_short	88.26	93.48	13.76	7.37	0.73
Narrative					
	$\bar{x}$ **(in ms)**	$\bar{y}$ **(in ms)**	μ **(in ms)**	σ **(in ms)**	$\gamma_{x,y}$
CART_long	402.87	397.99	104.90	77.26	0.69
CART_medium	198.09	198.87	10.60	9.79	0.66
CART_short	93.73	92.69	9.65	7.14	0.69
Dialogue					
	$\bar{x}$ **(in ms)**	$\bar{y}$ **(in ms)**	μ **(in ms)**	σ **(in ms)**	$\gamma_{x,y}$
CART_long	472.47	463.31	52.96	61.93	0.71
CART_medium	193.31	200.33	7.02	7.01	0.95
CART_short	87.56	89.03	9.74	10.07	0.76

$$\text{where } V_{X,Y} = \frac{\sum_i |x_i - \bar{x}|.|y_i - \bar{y}|}{N} \tag{4}$$

Here $\sigma_X \& \sigma_Y$ are the standard deviations for the actual and predicted duration values respectively, and $V_{X,Y}$ is the correlation coefficient between the actual and predicted pause duration values. The results of the objective measures in terms of, average of actual pause duration values $\bar{x}$, and average of predicted duration values $\bar{y}$ of pauses are shown in Table V.

From the Table V, we can notice that the average prediction error for long pause is significantly high compared to medium and short pause for descriptive and dialogue modes. The high prediction error is reasonable as the average actual pause duration for the long pause is also high. Hence, high prediction error does not significantly change the nature of the long pause as medium or short. The correlation coefficient ($\gamma_{x,y}$) for each CART trees at various discourse modes are better as shown in column 6.

B. Subjective Evaluation

Subjective evaluation is conducted to show the significance of the proposed pause prediction model based on the discourse modes. This pause prediction model is integrated with the earlier proposed Story TTS [19] for Hindi Language. These TTS systems were developed as part of Department of Information Technology sponsored project *Development of Text to Speech Systems in Indian Languages (Phase-II)*. For listening test, two stories that are not used for building the models are used as input to our proposed pause model. The listening tests are conducted on 20 subjects within an age group of 20-35, having Hindi as their mother tongue. The subjects are instructed to judge the quality of the synthesized speech on a five point scale for each utterance. The five-point scales are mentioned as 1: very poor, 2: poor, 3: fair, 4: good and 5: excellent. Table VI shows the mean opinion scores for the following cases: **MOS1**: Mean opinion score for the Story Hindi TTS system with no pause prediction model. **MOS2**: Mean opinion score for the Story Hindi TTS with proposed pause prediction model.

TABLE VI. MEAN OPINION SCORE FOR BASLINE AND PROPOSED METHOD

Approach	**Naturalness**
MOS1	2.85
MOS2	3.35

The statistical significance between the differences in the pairs of *MOS1* and *MOS2* is tested using hypothesis testing [28]. The level of confidence for the observed increment was obtained by using sample variances and values of Student-t distribution. The levels of confidence achieved for all the transitions are high (95%) for naturalness. From the subjective test, we can conclude that subjects perceived notable improvements in naturalness and intelligibility with incorporation of proposed phrasing or pause model based on discourse modes in Hindi Story TTS.

V. CONCLUSION

In this paper, we proposed an framework to model the pause pattern in storytelling style speech for Hindi language. The pauses are analyzed based on three modes of discourse (i.e descriptive, narrative and dialogue). We proposed three stage data-driven pause prediction model to learn the pause pattern at each discourse modes. From the analysis of the corpus, pauses are categorised into three type (i.e. short, medium and long pauses) for each mode of discourse. In three stage pause model,

first stage properly identifies the position of pauses within an utterance. At second stage, each pause is classified into three different categories. 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 at each stage for three modes of discourse. The results of perceptual evaluation indicates that the proposed method is effective in imposing pattern of the pauses in synthesized speech utterance.

Form these experiments we showed, that the discourse mode can capture story-semantic information present in story. This idea can be also used to model the other prosodic parameters like pitch, duration, intensity, tempo in storytelling style speech. At second stage, new features can be explored to improve the F-1 score of the models for three modes of discourse. In this work, discourse modes are manually annotated. Automatic discourse, prediction can also be proposed from an uttrance which may mitigate the tedious task of annotators. In addition to CART, different nonlinear classifiers can be explored. Further studies can be performed to analyze the pause patterns present at paragraph level [29] for storytelling style speech. Apart from Hindi, the current pause prediction study can be extended to other Indian languages. In earlier proposed, Story TTS [19], story-specific prosody rules can be derived at each discourse-level for the story-specific emotions.

ACKNOWLEDGMENTS

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

[2] 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.

[3] 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.

[4] 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.

[5] P. Zervas, M. Maragoudakis, N. Fakotakis, and G. Kokkinakis, "Bayesian Induction of Intonational Phrase Breaks," Eurospeech, 2003.

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

[7] K. Prahallad, E. V. Raghavendra, and A. W. Black, "Learning Speaker-Specific Phrase Breaks for Text-to-Speech Systems," in *The Seventh ISCA Tutorial and Research Workshop on Speech Synthesis, Kyoto, Japan, September 22-24, 2010*, 2010, pp. 162–166.

[8] K. E. Silverman, M. E. Beckman, J. F. Pitrelli, M. Ostendorf, C. W. Wightman, P. Price, J. B. Pierrehumbert, and J. Hirschberg, "TOBI: a standard for labeling English prosody," in *ICSLP*. ISCA, 1992.

[9] F. Goldman-Eisler, "The distribution of pause durations in speech," *Language and Speech*, vol. 4, no. 4, pp. 232–237, 1961.

[10] R. Dhillon, "Using pause durations to discriminate between lexically ambiguous words and dialog acts in spontaneous speeech," *The Journal of the Acoustical Society of America*, vol. 123, no. 5, pp. 3425–3425, 2008.

[11] 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.

[12] 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.

[13] 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.

[14] A. Parlikar and A. Black, "Data-driven phrasing for speech synthesis in low-resource languages," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 4013 – 4016.

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

[16] 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.

[17] P. Sarkar and K. S. Rao, "Data-driven pause prediction for speech synthesis in storytelling style speech," in *National Conference on Communication*. IEEE, Feb-Mar 2015.

[18] 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.

[19] 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)*, Aug 2014, pp. 473–477.

[20] 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*, Jan 2015, pp. 1–6.

[21] 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.

[22] 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.

[23] K. Prahallad, A. Black, and R. Mosur, "Sub-Phonetic Modeling For Capturing Pronunciation Variations For Conversational Speech Synthesis," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1, May 2006, pp. I–I.

[24] A. W. Black, "CLUSTERGEN: a statistical parametric synthesizer using trajectory modeling," in *INTERSPEECH 2006 - ICSLP, Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, September 17-21, 2006*, 2006.

[25] A. W. Black and K. Lenzo, *Building voices in the festival speech synthesis system*, 2002.

[26] 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.

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

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

[29] 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.