Data-Driven Pause Prediction for Speech Synthesis in Storytelling Style Speech

Parakrant Sarkar, 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 the storyteller speech, pauses plays a significant role in introducing suspense and climax. Pauses are used to emphasize keywords, emotion-salient words and separate the phrases in the utterance. The objective of this work is to predict the position and duration of the pauses in the synthesized speech from the text-to-speech system. We analyzed the pause patterns in storyteller speech and classified the pauses into three different categories, that is, short, medium and long pauses. A data driven three stage pause prediction model is proposed. In the first stage, the model is built properly to identify the pause position within an utterance using a set of word-level features. In the second stage, the pauses are classified into three different categories using a set of syllable-level features. In the final stage, a regression predictor is trained to predict the pause duration for each category. We conducted both objective and subjective tests to evaluate the proposed method. The subjective evaluation showed that subjects are perceiving a noticeable difference in the synthesized speech using the proposed method.**

***Keywords*—Storytelling style, Pause prediction, Phrasing, Pause Duration, Breaks, Non-break, Speech synthesis, silences.**

I. Introduction

In natural speech, phrase breaks are realized as silences. Mostly, while speaking people pauses between words. Pausing is carried out to emphasize something relevant to the context, thus making the utterance more intuitive. Quite often, pausing is performed, just to take a breath [1]. Appropriate pausing in the speech can enhance the intelligibility and make the speech more persuasive. In text-to-speech systems, a speech synthesizer should provide phrase breaks in a similar manner. Phrase breaks provides the foundation required by the other prosodic models such as duration, pitch and intensity in speech synthesizer. The procedure of finding out where a synthesizer should insert these pauses is called phrase break prediction or pause prediction. It is basically classifying each of the word boundaries in the text as a break or non-break. A phrase does not possess the grammatical organization of a sentence. It is a syntactic structure with one or more words present within a sentence (i.e. utterance).

Earlier works address the issue of pause prediction in speech synthesis. Style specific pause prediction is shown in [2]. Also, a pause prediction is carried out using speaker specific features in [3]. According to the TOBI scheme [4], there are different types of phrase breaks based on the duration values. In [5], the analysis based on the length of pauses in speech is carried out for different speakers in various contexts. The distribution of pauses in the speech utterance affects the meaning and perception. The pause duration is also a reliable means of discriminating between lexical ambiguity of words [6]. For Indian languages, a set of morpheme tag units [7] are manually identified and used to model phrase breaks for Telugu language. In [8] phrase break prediction is performed for Bengali language. A word terminal (i.e. last syllable of the word) [9] can be used as a feature to predict the phrase breaks. In TTS systems, all duration models treat pauses separately. In Klatt model [10], the Festival based TTS system [11] as well as the Mary TTS system [12] assign 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 [13] but generating the appropriate duration of the pause has not received much attentions.

In a storytelling style speech [14], a storyteller uses expressive, engaging and entertaining style speech. For analysis we recorded the children stories (i.e. story-speech corpus) from a professional storyteller. From the children stories, it is observed that different parts of a story are narrated in different styles. These styles are based on the semantics present at that part of the story. In general, most of the stories in the database begin with introducing the characters present in the story, followed by various events related to the story and finally the story will conclude with a moral. In the context of the storyteller speech, pauses are used for separating phrases, emphasizing keywords and emotion-salient words to introduce suspense and climax in the story.

This paper introduces a data-driven approach to model duration of the phrase break (i.e. pauses) for Hindi language. Our goal is to investigate the pause patterns in storytelling style speech recorded from a professional storyteller. A three stage pause prediction model is proposed to accurately determine the position and duration of the pause. In the first stage from the text, each word boundary is classified as break and non-break. Each of theses break at word boundary is again classified into one of the three categories of pause i.e. long, medium and short based on its duration. A set of syllable features are extracted to learn a regression predictor for each type of pause to predict the duration. Here, we have considered only the breaks or pauses which occur in between sentences. We are not modeling the breaks at the end of an utterance because punctuation mark is used for this purpose.

978-1-4799-6619-6/15/$31.00 © 2015 IEEE

The paper is organized as follows, the story speech corpus used in this work is described in section II. The proposed method of building the pause prediction model is discussed in section III. Section IV provides the evaluation of the proposed method. Finally in section V, summary and conclusion of the present work are mentioned.

II. Story Speech Corpus

Speaking style varies form person to person [15]. Each individual has his or her style of speaking. Even a person has their own modus operandi when reading a text based on the different contexts and situation. Also, speaking style can alter depending on the task at hand. We carried out the analysis based on the pause patterns in the storytelling style speech [16]. In storytelling style speech, a storyteller provides pauses while narrating the story in order to introduce suspense, climax and various story-specific emotions [14]. In the context of storyteller speech, pauses are used for separating phrases, emphasizing keywords and emotion-salient words.

The story texts comprising of children's tales are collected from the story books like Panchatantra and Akbar-Birbal. A total of 100 story texts are collected. The number of sentences in a 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, delivered 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.

Manual labeling of the phrase breaks for the entire story-speech corpus is quite tedious. For entire corpus force alignment of the speech wave file with the text prompts are performed using the HMM tool [17]. This provides us with the position of pauses and its duration values introduced by the storyteller. Fig. 1 shows the histogram plots of durations of the pauses present in the corpus. There are total of 3804 pauses present in the corpus. It is observed from the histogram that the occurrence of pauses are more with a duration value ranges between 50 to 400 ms. Analysis of the duration of the pauses are carried out based on story semantic information [14]. We divided the pauses into three types present in a story corpus. These are (i) *Long pause* (>250 ms) (ii) *Medium pause* (150 − 250 ms) and (iii) *Short pause* (<150 ms). From the analysis, it is also observed that the pauses with duration value less than 50 ms are not relevant to story semantic information. Whenever a pause with duration (<50 ms) is incurred in the story, listeners may not perceive it as pause. So, all the pauses with duration less than 50 ms in length are ignored for building the models.

Table I shows the different types of pauses present in the story-speech corpus. The mean and standard deviation of the different pause types are also shown. In the story-speech corpus, there are 17% long, 25% medium, 38% short pauses and 20% of the pauses are ignored.

III. Building Pause Prediction Model

This section talks about the procedure followed in building the model for pause prediction. We propose a three stage pause prediction model related to story-specific pause pattern present

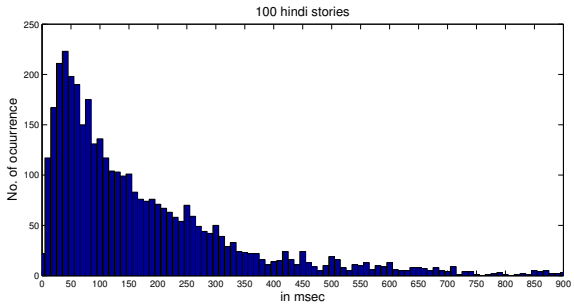

Fig. 1. Histogram of pause duration in story-speech corpus

TABLE I. 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

in the corpus as shown in the Fig. 2. In the first stage, goal is to build the model to properly identify the position of pauses within an utterance using a set of word-level features extracted from the corpus. The second stage deals with the classification of the pauses into three different types (i.e. long, medium and short pauses) using syllable-level features. In the final stage, we model a regression predictor to predict the duration of the pause based on its type. In the following subsections, we will briefly discuss each of the three stages of pause prediction model.

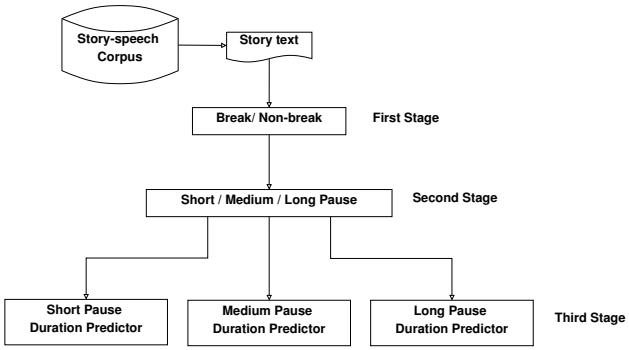

Fig. 2. Three stage pause prediction model

A. First stage of pause prediction model

The basic goal in the first stage is to automatically predict the position of the pauses in an utterance. A set of word level features are extracted from the text. A CART tree is modeled with these set of features to predict whether a word boundary should be marked as break (B) or non-break (NB). 10-fold cross validation technique is followed. In this technique, entire corpus is divide into 10 sets. Out of 10 sets, nine sets are used for training the CART Tree and one set is used for testing. The average performance of the CART trees are calculated which are trained on each set. We have used the Wagon tool[1] for building the CART tree present in the speech tools.

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

A five-gram window is followed in our current study to extract the features at word-level. The five-gram window is represented by the previous two words, current word and following two words. Positional, structural, morphological and emotion features are derived from words. In morphological feature, a Hindi POS tagger[2] developed at IIT Kharagpur is used to extract the Parts of speech (POS) information of the word. Also, phonetic strength [8] feature for the words are extracted manually from the corpus. Generally, while speaking people give stress in a word. This provides the information of the phonetic strength of a word. It is used as a binary feature. This feature helps in noting the emphasis, contrast and citation of a word. During the analysis of pattern in pauses, we noticed that the presences of a pause after an emotional salient word is more prominent as compared to a neutral word. Also, from the POS information, it is inferred that the words having adjective, auxiliary verbs and adverbs have a higher probability of having a pause at its word boundary. So, to automatically extract the emotion feature at word-level POS tags of the words are used. The word-level features used for training the CART trees are as follows:

1) Positional:
   - Position of the current word from the beginning and ending of the utterance.
   - Total number of words in the utterance.
2) 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.
3) Morphological:
   - POS of current, previous and following two words.
   - Phonetic strength of current word.
4) Emotion
   - Emotion of the current word.

All the features used in the feature-set may not be significant in predicting accurately the locations of the pauses in an utterance. So, the CART tree is trained with stepwise option. This helped to select the most optimal features for building the CART tree. Also, we empirically determined the stop value to be 40. Hence, this stage accurately predicts the positions of the pauses in an utterance.

B. Second stage of pause prediction model

The objective of this stage is to classify the pauses into three different types i.e long, medium and short pauses. Similar to first stage, CART model is trained by extracting the features at syllable-level. The features related to terminal syllables are extracted in this stage. After predicting the position of the pauses in the utterance, each pause is classified into one of the three types of pause. Positional, structural, morphological and emotion set of features are used in modeling. These set of features are described as follows:

1) Morphological:
   - Terminal syllable of the current, previous and following two words.
2) 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) 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.
4) Emotion
   - Emotion of the current word.

The CART tree is trained with stepwise option to know the optimal set of features. Also, we experimented with various stop values and decided to use a value of 15. Hence, this stage classify the pause into three different types (i.e. long, medium and short pause).

C. Third stage of pause prediction model

In this final stage, we predict the duration of the pause based on its type. For each of the three different types of pauses, three different CART trees are trained (i.e. CART_long, CART_medium and CART_short). The set of features used in this stage are similar to the second stage. As the CART trees are trained with stepwise option, the least contributing feature is emotion feature of the word. So, we decided not to use this feature for training. Also, CART trees are trained with various stop values and we decided to choose 15 items in every leaf node. Finally, this stage predicts the duration of the pause based on its type.

IV. Evaluation of the three stage pause prediction model

In this work, the proposed three stage pause prediction model is evaluated using both objective and subjective measures. The objective measure consists of percentage of correct prediction for the breaks and non-breaks. The model should provide high recall i.e. predicting as many breaks as there are in the actual (or reference) data. Also should have high precision i.e. predicting pauses in the wrong places should be less. We can calculate the F-1 measure [18] to evaluate our models at each 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. In addition to the objective measures, we conducted a subjective listening tests to determine the significance of the phrasing model in the Hindi TTS system.

A. Objective Evaluation

1) Accuracy of first stage of pause prediction model: 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 the model. For each model, we calculate the F-1 measure for breaks and non-breaks. The accuracy is given by the average F-1 measure across all 10 sets. Ideally, a model

[2]http://nltr.org/snltr-software/

should predict all breaks actually present in the test data (high recall) and also does not wrongly predict the breaks as non-breaks (high precision). The average F-1 measure at first stage is shown in the Table II. The average F-1 measure for break is 0.74 and for non-break is 0.91.

TABLE II. BREAK AND NON-BREAK PREDICTION ACCURACY (F-1 SCORE)

	Recall	Precision	F-1 Score
Non-break	0.89	0.94	0.91
Break	0.68	0.81	0.74

We also calculated the average accuracy of the first stage based on different testing sets. For training and testing we used 1764 (i.e. 9 sets) and 196 (i.e. 1 set) sentences. On an average the pause prediction model predicts the locations of the 639 breaks (B) with 84.91% and 1801 no-breaks (NB) with 87.45% accuracy.

2) Accuracy of second stage of pause prediction model: Similarly, 10-fold cross validation is performed. We calculated the average accuracy (i.e. F-1 measure) of the model in second stage across all 10 sets. The F-1 measure is shown in the table III. The average F-1 measures for long is 0.67, medium is 0.51 and short is 0.58.

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

	Recall	Precision	F-1 Score
long pause	0.73	0.62	0.67
medium pause	0.50	0.52	0.51
short pause	0.53	0.63	0.58

3) Accuracy of third stage of pause prediction model: We evaluated the CART trees for three different pause types i.e. short, medium and long. This evaluation is done by prediction accuracy. The prediction accuracy is evaluated by means of objective measures [19] such as average prediction error (μ), standard deviation (σ) and correlation coefficient ($\gamma_{x,y}$) between the actual and predicted duration values. The formula 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

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

where σ_X, σ_Y are the standard deviations for the actual and predicted duration values respectively, and $V_{X,Y}$ is the correlation 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 IV. From the table, we can observe 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 high 347.96 ms. The high prediction error of 76 ms does not significantly change the nature of the pause as medium. The correlation coefficient ($\gamma_{x,y}$) for each CART trees are better.

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

	$\bar{x}$ (in ms)	$\bar{y}$ (in ms)	μ (in ms)	σ (in ms)	$\gamma_{x,y}$
CART_long	347.96	368.71	76.13	55.05	0.65
CART_medium	208.43	199.30	26	17.03	0.66
CART_short	87.33	91.01	34.05	22.19	0.77

B. Subjective Evaluation

In subjective evaluation listening tests are conducted to show the significance of the proposed pause prediction model. The listening tests are performed on 10 synthesized sentences before and after incorporating the three stage pause prediction model in the baseline Hindi TTS. The baseline TTS systems refer to syllable based TTS [20], developed using Festival framework with neutral speech corpus. These TTS systems were developed as part of Department of Information Technology (DIT) sponsored project *Development of Text to Speech Systems in Indian Languages (Phase-I)*. 20 research scholars in the age group of 20-35 participated in the test. The subjects are instructed to judge the quality of the synthesized speech on a 5 point scale for each sentence. The details of the five point scales are mentioned in Table V:

TABLE V. INSTRUCTION TO SUBJECTS FOR MEAN OPINION SCORE METHOD

Score	Quality
1	Poor quality with very low intelligibility
2	Poor quality but intelligible
3	Good quality and intelligible
4	Very good speech quality but less naturalness
5	As good as natural speech

Table VI shows the mean opinion scores for the following cases:

- MOS1: Mean opinion score for the baseline Hindi TTS system with no pause prediction model.
- MOS2: Mean opinion score for the baseline Hindi TTS with three stage pause prediction model.

TABLE VI. MEAN OPINION SCORE FOR BASLINE AND PROPOSED METHOD

Approach	Naturalness	Intelligibility
MOS1	2.95	2.85
MOS2	3.45	3.35

The significance of the differences in the pairs of the MOS1 and MOS2 is tested using hypothesis testing [21]. 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 (99.5% for naturalness and 95% for intelligibility respectively). From the subjective test, we can conclude that subjects perceived notable improvements in naturalness and intelligibility with incorporation of proposed phrasing models in the Hindi TTS system for storytelling style speech.

V. CONCLUSION

In this paper, we proposed an approach to model the pattern of the pauses present in storytelling style speech for Hindi TTS system. We analyzed the pause patterns using the story-speech corpus. Based on the analysis, three different categories (i.e. short, medium and long pauses) are considered. A three stage pause prediction model is proposed to learn the pause patterns in storytelling style speech. In this work, prediction of proper position and duration of pauses in an utterance is performed at word boundaries. In the first stage, a model is developed to properly identify the position of pauses within an utterance. In the second stage, the pauses are classified into three different types using syllable-level features. In the final stage, a regression predictor is modeled to predict the duration of these pauses based on its type. The CART tree models are evaluated at each stages by using objective measures. Subjective listening test are also performed after incorporating the proposed method in baseline Hindi TTS. The results of the perceptual evaluation indicates that the proposed method is effective in imposing the pauses in the synthesized speech utterance.

Possible extensions to the current work are as follows. In addition to CART, different nonlinear classifiers can be explored. Further studies can be performed to analyze the pause patterns present at paragraph level [22] for storytelling style speech. Apart from Hindi, the current pause prediction study can be extended to other Indian languages. The present pause prediction method can be integrated with our previously proposed Story TTS systems [14] for accurate pause prediction and also for improving the quality of synthesized story speech.

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] 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.

[2] A. Parlikar and A. Black, "Modeling Pause-Duration for Style-Specific Speech Synthesis," in *INTERSPEECH*. ISCA, 2012. [Online]. Available: http://dblp.uni-trier.de/db/conf/interspeech/interspeech2012.html

[3] 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. [Online]. Available: http://www.isca-speech.org/archive/ssw7/ssw7_162.html

[4] 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. [Online]. Available: http://dblp.uni-trier.de/db/conf/interspeech/icslp1992.html

[5] F. Goldman-Eisler, "The distribution of pause durations in speech," *Language and Speech*, vol. 4, no. 4, pp. 232–237, 1961. [Online]. Available: http://las.sagepub.com/content/4/4/232.abstract

[6] 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. [Online]. Available: http://scitation.aip.org/content/asa/journal/jasa/123/5/10.1121/1.2934182

[7] N. S. Krishna and H. A. Murthy, "A New Prosodic Phrasing Model for Indian Language Telugu," in *INTER-SPEECH*. ISCA, 2004. [Online]. Available: http://dblp.uni-trier.de/db/conf/interspeech/interspeech2004.html

[8] 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.

[9] 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.

[10] 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.

[11] 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. [Online]. Available: http://www.cstr.ed.ac.uk/projects/festival.html

[12] 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.

[13] 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.

[14] 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.

[15] J. Hirschberg, "A corpus-based approach to the study of speaking style," in *Prosody: Theory and Experiment*, ser. Text, Speech and Language Technology, M. Horne, Ed. Springer Netherlands, 2000, vol. 14, pp. 335–350. [Online]. Available: http://dx.doi.org/10.1007/978-94-015-9413-4_12

[16] 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

[17] 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.

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

[19] 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. [Online]. Available: http://dx.doi.org/10.1016/j.csl.2006.06.003

[20] M. N. Rao, S. Thomas, T. Nagarajan, and H. A. Murthy, "Text-to-speech Synthesis using syllable like units," in *Proceedings of National Conference on Communication (NCC)*, IIT Kharagpur, 2005, pp. 227–280.

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

[22] 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. [Online]. Available: http://dblp.uni-trier.de/db/conf/interspeech/interspeech2007.html