Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2015) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

Second International Symposium on Computer Vision and the Internet (VisionNet'15)

Modeling Pauses for Synthesis of Storytelling Style Speech using Unsupervised Word Features

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 the storytelling style speech pauses or phrase breaks play a significant role in introducing suspense and climax in the story. More often pauses are provided by a storyteller to capture the audience's attention by emphasizing keywords, focusing emotion-salient words, and to separate key phrases in an utterance. The goal of the work presented in this paper is to predict the location of pauses, in an utterance synthesized by a Story Text-To-Speech (TTS) system using unsupervised features at word-level. Traditional methods for predicting pauses uses the foremost linguistic features like Parts-of-Speech (POS) tags, chunking information or terminal syllables, etc. These methods presuppose the availability of linguistic knowledge by an automatic tagger or manually annotated corpus. However, this information's are not readily available in case of Indian Languages. Manually annotating the text with this linguistic information is quite hectic and time consuming. Also, these pieces of information's do not capture the co-occurrence statistics of words. Hence, we propose a framework for integrating the Story TTS with proposed pause prediction module. In this module, an unlabeled text corpus is used to extract, the continuous-valued world-level features to model the pause patterns in storytelling speech. A set of story-specific (SS) features are introduced for capturing story-semantic information based on pause pattern. A various combination of pause predictions systems is- proposed such as B, POS, U, POS+SS and U+SS. These systems are evaluated objectively by F-1 Score.

Peer-review under responsibility of organizing committee of the Second International Symposium on Computer Vision and the Internet (VisionNet'15).

Keywords: Storytelling style; Pause prediction; Phrasing; Vector Space model; Latent Semantic Analysis; Singular valued decomposition; Speech synthesis.

Peer-review under responsibility of organizing committee of the Second International Symposium on Computer Vision and the Internet (VisionNet'15).

1. Introduction

In Text-To-Speech (TTS) systems, prosody plays a vital role in synthesizing natural and intelligent speech. Prosody is a concoction of supra-segmental features (like duration, pitch, intensity, co-articulation pattern and pause pattern) which provides additional information in speech that is not found in text. In the context of TTS, the procedure of finding out where the synthesizer should insert pauses is called phrase break or pause prediction. It is the first step, in modeling the prosody during speech synthesis process. Rest of the prosody models (like duration, pitch and intensity) depend on the decision of pause prediction model. Therefore, if the pause model provides wrong information, it decreases the performance of the other models that are using this information. Hence the appropriate prediction of the pause pattern is critical to the overall naturalness of the synthetic speech. In natural speech, pausing is carried out to emphasize something relevant to the context or just to take a breath [1]. Also, providing appropriate pauses in the speech not only enhances the quality but also provide affect [2] in the speech. In context of storytelling style speech [3], a storyteller uses his/her voice in a variety of ways to capture the audience's attention. This include mimicking character's voices, producing various sound effects, use prosody to convey and invoke emotions, and providing appropriate pauses, thus creating an engaging listening experience.

In this paper, we focus our research on the modeling of pause pattern present in storytelling style speech [4]. For analysis, children stories (i.e. story-speech corpus) are recorded from a professional female storyteller. By analyzing the story-speech corpus, it is inferred that, a children story comprise of different parts and each part is narrated in different styles. These styles are biased towards the semantics present at that part of the story. Mostly, a children story begins with introducing the characters present in the story, followed by various events related to the character and finally story will conclude with a moral. In context of storyteller speech, a pause during the narration of a story not only enhances the audience's understanding of the story, but also builds anticipation. A pause can also, be used to introduce suspense and climax in the story. It can also be used for separating phrases, emphasizing keywords and emotion-salient words. The problem of the pause prediction can be thought as a classification problem: given a text, we want to classify each word boundary as being a pause or non-pause.

Earlier works are carried out for predicting the location of the pauses by using machine learning models like Hidden Markov Models (HMM) [5] or Classification and Regression Trees (CART) [6] or Feed Forward Neural Network (FFNN) [7], Maximum Entropy (ME) [8] and Bayesian approach [9]. In Klatt model [10], the Festival based TTS system [11] as well as the Mary TTS system [12] assign fixed duration to the pauses. These approaches uses labeled data with the knowledge of linguistic classes such POS tags, morphological features etc. The training of the classifiers assumes the availability of the hand-labeled training data in large quantities which also includes proper tags provided by a high quality (in terms of accuracy) POS taggers/shallow parsers. Manual annotation is tedious as well as time consuming, and might not be an effective option for the languages where the required resources are low or not readily available.

To address these shortcomings, various works are directed towards unsupervised method of inducing word representations. This representation is used to substitute the prior knowledge of linguistic classes. In the paper [13], Ney-Essen clustering algorithm is used for automatic induction of the POS. These are generated based on the frequency analysis of the words present in the corpus. For Indian languages, a set of morpheme tag units [14] are manually identified and used to model phrase breaks for Telugu language. In [7] phrase break prediction is performed for Bengali language with inclusion of new phonetic strength feature. A terminal syllable (i.e. rear syllable of the word) [15] can be used as a feature to predict the phrase breaks. In storytelling style speech for Hindi language, a three stage pause prediction model is proposed to accurately determine the position and duration of the pauses [16]. All the machine learning based approaches mentioned so far use a discrete linguistic knowledge of word representation. These representatives need accurate classification of words into a set of discrete classes. The traditional methods use linguistic resources like POS information generated by POS taggers or shallow parsers [17]. However, POS tags are the primary feature used for phrase break prediction. There exits issues like there may be words having more than two POS tags depending on the context it is used, and also this representation does not capture the distributional behavior [18] of the words. These issues are addressed by vector space model approach [19]. Word co-occurrence matrix is formed to capture the distributional behavior of a word. The row of the co-

occurrence matrix denotes the words which are nothing but points in the continuous space. A neural network dictionary learning architecture is proposed in [20] for phrase break prediction.

In this work story-speech corpus is recorded from a professional storyteller. The distributions of pauses based on durations are analyzed. Here, we have considered only the pauses which occur in between sentences. We are not modeling the breaks at the end of an utterance. A supervised prediction of pauses is done using CART. However, we use the Vector Space Model (VSM) features of the words that are attained in an unsupervised fashion. Hence, there is no need to determine the discrete linguistic classes like POS tags/ terminal syllables. In this work, from an unlabeled text corpus, we extract the unsupervised word level features to model the pauses. These features can be used to substitute the existing state of the art linguistic features. Various systems for pause prediction are built: Baseline System (BS), System with full POS Information (POS) and System with unsupervised features (U). A set story-specific features are proposed to capture the story-semantic [4] information. The performance of these models are evaluated by using F-1 score [21].

This paper is organized as follows; Section 2 describes the story-speech corpus. The procedures of extracting the unsupervised features for words are discussed in Section 3. The Section 4, provides details of building the pause prediction model using continuous word features. The details for building the various pause prediction systems are explained in Section 5. The performances of all the systems are discussed in Section 6. Finally, we consolidate the present work with conclusion, future work and acknowledgement.

2. Story Speech Corpus

In general speaking style varies from person to person. It also depends on the nature of the task that narrator is engaged with, such as news reading, extempore, conversation etc. Storytelling style [3] is one of them. A storyteller uses his/her skills with various prosodic transitions while narrating a story to capture the audience's attention. These sets of skills include introducing pauses while narrating the story to induce various story-specific emotions as shown in [4]. A total of 100 children story texts were collected from various story books such as Panchatantra and Akbar-Birbal. The number of sentences in each story varies from 25 to 30. The collected stories comprises of 1960 sentences with 24400 words. These stories are narrated by a professional storyteller, which is recorded in a noise free lab environment. A rigorous and continuous feedback is given to the storyteller for improving the quality of the narrated story in order to maintain high quality. The total duration of the corpus is about 3 hours.

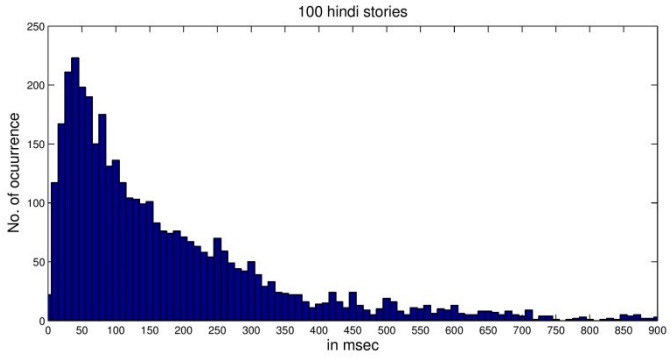

Fig. 1. Histogram of pause duration in story-speech corpus [16]

The entire story-corpus is annotated manually to obtain the proposed story-specific (SS) features. Annotation task is carried out by four native Hindi speaking experts. The features are emotion (sad, anger, happy, fear, neutral) of the current word, whether the word is a content or functional and class of the story (fable, legendary, folk-tales). Each annotator's task is to label the words into one of the emotions and whether the word is a content or functional. Also, at story-level, each story is annotated with one of the story class. The Fleiss Kappa (k) gives the inter-annotator

agreement are 0.77, 0.72 and 0.70 respectively. The (k) values above 0.65 or so can be considered to be substantial. For the entire story-speech corpus, manual annotation of the pauses is cumbersome. Hence, we followed a forced alignment of the speech wave file with text prompts are performed using the HMM tool [22], and a CLUSTERGEN [23] voice is built with Festival [11] framework. The position and duration information of pauses introduced by the storyteller is obtained. The Fig. 1 shows the distribution of the durations of pauses present in the corpus. A total of 3804 pauses are present in the corpus. These are the pauses present within an utterance. Punctuation marks realize pauses present at an end of the utterance. Hence, pauses that occur at the end of an utterance are ignored for modeling. Also, it is observed that the majority of the pauses are distributed within the duration value ranges from 50 to 400 ms. Moreover, we noted that the pauses with duration value less than 50 ms are not relevant to the story-semantic information. Whenever a pause having a duration value (<50 ms) is incurred in the story, listeners may not perceive it as a pause. Hence, we ignored the pauses for building models with a duration value less than 50 ms. Table 1 shows the mean and standard deviation of the pauses and ignored pauses present in the corpus. In the story-speech corpus, there are 77.82% and 22.18% pauses and ignored pauses respectively.

Table 1. Pause information in the story-speech corpus

Pause Type	Mean(ms)	StdDev(ms)	%in original
Pause	225.30	188.14	77.82
Ignored Pause	31.57	12.80	22.18

1.1. Story Text-to-speech system

The goal of story TTS system is to synthesize storytelling style speech from the neutral TTS for a given story text as input. For various Indian Languages, story TTS [4] are developed. An additional set module adds the existing neutral TTS. These modules include (i) Story-specific emotion detection (SSED) module, (ii) Story-specific prosody generation (SSPG) module and (iii) Story-specific prosody incorporation (SSPI) module. The Fig. 2 shows the entire process of story synthesis. The raw story text is given 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 with each phrase 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 SSPG module specifies the desired modification factors associated with various prosodic parameters (like pitch, duration, intensity, tempo and pause pattern). The SSPI module incorporates the modification factors in the synthesized neutral speech. In this work, we added story-specific pause prediction (SSPP) module, which is an extension of the SSPG module. It predicts the proper position of the pauses based on story semantics. In this paper, we are focusing on predicting the location of the pauses using unsupervised features from an unlabeled story-text corpus. The total duration of the corpus is about 3 hours.

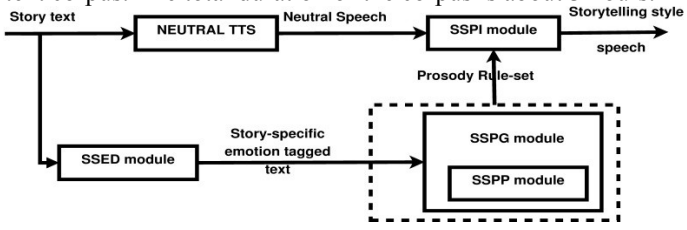

Fig. 2. Block diagram of Story Synthesis

3. Unsupervised Word Features

This section discusses the procedure followed in extracting the unsupervised continuous-valued word features. The primary motivation for this idea comes from the fact that the words can be distributionally [18] represented. Latent Semantic Analysis (LSA) [24] is carried out which is based on characterizing words with which it co-occurs in the body of the text. The following steps elaborate the process of the unsupervised word-level feature representation. m unique word types extracted from the story-speech corpus. n is the most frequently occurring words (also termed as feature words) present in the corpus where n << m. An m × 2n co-occurrence matrix C is computed where: C_{ij} counts the number of times the i^{th} word types occurs with the j^{th} feature word as the left context. C_{ij+n} counts the number of times the i^{th} word types occurs with the j^{th} feature word as the right context. Following this procedure, each of the m unique words is mapped to the points in 2n continuous dimensional space, as defined by the rows of the co-occurrence matrix. Hence, each vector has a dimension of 2n. The calculated co-occurrence matrix C is sparse. A typical Principal Component Analysis (PCA) is applied using Singular Valued Decomposition [25] to project the co-occurrence matrix into the dense lower dimensional form. The transformation is as follows:

$$C_{m\times 2n} = U_{m\times r} \cdot D_{r\times r} \cdot V^{T}_{r\times 2n}$$

Where, columns of the U and V matrices are the left and right singular vectors and D is the square diagonal matrix whose diagonal entries are the best r singular values of C. The matrix U is the dense and lower dimensional representation of the matrix C. Also, each vector represents a r dimensional feature vector corresponding to a word in the latent space.

4. Pause Prediction Model using continuous word features

This section discusses about the procedure followed in building the various models for SSPP module. From the story-speech corpus, as discussed in the Section 3 for each word feature extraction is done. These feature representation is used for modeling the pauses using CART model. The reasoning behind choosing CART as the classification model because the implementation of Story TTS followed the Festival framework and CART is readily available in the Speech Tools (available http://www.cstr.ed.ac.uk/projects/speech_tools/). A CART is modeled with these set of unsupervised features to predict whether a word boundary should marked as pause(P) or non-pause (NP) for an utterance. Now we elaborate the feature extraction from the story-speech corpus. There are a T = 24400 total number of words in the corpus. A dictionary is created by finding out the m = 3579 unique word types. These words are tailed with n most frequently occurring word types. In this case, we have conducted the experiments with different values of n and determined n = 300 suits the best. The co-occurrence matrix C of dimension m × 2n is calculated, in which the column denotes the n most frequently occurring words and rows represents the m unique words. An m × 600 co-occurrence matrix is formed, then this matrix is reduced to an m × 50 dimensional matrix U, using Singular Valued Decomposition. Hence, the dictionary consists of m unique words and each word comprises of r = 50-dimensional vector. *UNK* tag is provided for the word present in the dictionary having occurrence frequency one. We followed a ten-fold cross-validation for training and testing. The entire data is divided into ten sets. For each fold, 9 sets are used for training and 1 set is used for testing. Co-occurrence matrices are computed from both the 9 sets (for training), and 1 set (for testing). Contextual information is addressed by a five-gram window to build the feature vector for a word. Experiments are carried out with three, five and seven-gram window respectively, out which five-gram model gave the best performance. The previous two words, current word and following two words are considered in the five-gram model. The r = 50-dimensional vector of these words are concatenated together to produce a fixed length of 250 size feature vector. Hence, for training these characteristics are extracted from the total number of words present in the training set. For testing, similar way features are obtained from the testing set. A notable problem arises; in this method is how to handle the unseen words at test time. In this case, a similar approach is followed as shown in [26]. Here a portion of the training set is taken in which, words having a frequency of occurrence, only once. These words are marked with a special token *UNK*. Whenever an unseen word is encountered, we will randomly assign feature from these *UNK* token word types.

For training, features are computed for all the word types including the unseen word types. At test time, all the hidden words are randomly mapped to the *UNK* token and corresponding feature vector is obtained.

5. System Built

This section describes different systems built for prediction of the pauses in storytelling style speech. For training, we followed a ten-fold cross-validation technique for the following regimes. The features are extracted based on the five-gram window. The pause prediction system is as follows:

1.2. BS: Baseline systems

The system built with a set of essential features related to punctuation mark ,.!?. Whenever in the body of the text the punctuation marks come, a pause is provided in the speech. The word-level features used for training the Baseline System are as follows:

- Position of the current word from the beginning and ending of the utterance.
- Total number of words/syllables/phones in the utterance.
- Total number of phones/syllables in the current, previous and following two words.
- Total number of syllables/phones in the current, previous and following two words.

1.3. POS: System with full POS Information

The POS information of words collected from a Hindi POS tagger (available: http://nltr.org/snltr-software/) developed at IIT Kharagpur. The POS information obtained from the automatic POS tagger can be divided in 27 tags which includes (such as noun, proper noun, adjectives, adverbs, quantifiers, verbs, auxiliary verbs, conjunction, punctuation mark etc). For each word, these POS tags are used as a feature. Hence, a POS system for pause prediction is built using the full POS information along with all the features used for baseline system.

1.4. U: System with unsupervised features

A system with unsupervised features is trained by extracting, continuous valued features as described in section III. These features are obtained from the untagged text of story-speech corpus. Along with the unsupervised characteristics, we will also include the story-specific (SS) features like emotion (sad, anger, happy, fear) of the current word, whether the word is a content or functional, class of the story (fabled, legendary, folk-tales). These story-specific features are hand annotated from the story-speech corpus. We also explored combining these unsupervised features with story-specific features. The hypothesis behind the addition of story-specific features along with the unsupervised features may add some complementary information related to story-semantics [4].

6. Pause Prediction Model using continuous word features

In the SSPP module, we built pause prediction model for each of the three systems (i.e. BS, POS & U) by performing 10-fold cross validation. For each system, we calculate the F-1 measure [21] for pause and non-pause prediction to evaluate the model. The F-1 test is the harmonic mean of recall and precision. A good model gives a higher F-1 score close to 1.00. Ideally, high recall signifies that the model is predicting all the pauses that are present in test data. Also, high precision means that the model does not wrongly predict the pause as non-pause. The results of the evaluation in terms of F1-score, measured on the various systems built as shown in the Table 2. The F1-score of the different methods predicting the word boundary as a pause are 0.56 for baseline system (BS), 0.75 for system with full POS Information (POS) and 0.65 for system with unsupervised features (U). Similarly, the F1-score of the methods predicting the word boundary as a non-pause i.e. BS, POS and U are 0.86, 0.91 and 0.88 respectively. An observation can be drawn from the results that, the POS system performance is better as compared to BS and U

systems. Also, using the unsupervised word features improves the performance of the U system compared to BS. There is an absolute increase of 0.11 (F1), 0.13 (P) and 0.04 (R) is noted for predicting pause in the U as compared with BS. Hence, instead of using the features of POS system, we can opt for the unsupervised features with little or no degradation in term of performance as compared to BS system. The F1-score of the systems (i.e. U+SS, POS+SS) with the story-specific (SS) features shown in the column 4 and 5 of the Table 2. The addition of SS features with the systems exhibits an improvement in the F1 score for both pause and non-pause prediction. There is an absolute increase of 0.15(F1 for pause) and 0.02(F1 for non-pause) for U+SS system as compared with the BS. Similarly, a total increase of 0.18(F1 for pause) and 0.07(F1 for non-pause). Hence, the addition of the SS features helps in improving the performance of the system. Also, intuitively we can say that adding these SS features captures the story-semantic information. Hence, in SSPP module, we can use the U or U+SS system for predicting the accurate position of the pauses. Besides to CART, we have explored different non-linear classifiers such as ANN and SVM for classification. The order of classifiers in terms of performance CART <ANN <SVM.

Table 2. Performance (in terms of F-1 score, P: Precision, R: Recall) of various Systems (BS, POS, U) for predicting Pause and Non-pause.

Systems	**Pause**			**Non-Pause**		
	R	**P**	**F1**	**R**	**P**	**F1**
BS	0.66	0.48	0.58	0.83	0.91	0.86
POS	0.69	0.81	0.75	0.94	0.89	0.91
U	0.70	0.61	0.65	0.91	0.87	0.88
U+SS	0.72	0.68	0.71	0.88	0.89	0.88
POS+SS	0.69	0.79	0.74	0.95	0.91	0.93

7. Conclusion

In this paper, we described a continuous valued feature representation for words used for pause/non-pause prediction. These set of features utilized in the storytelling style speech for Hindi Story TTS. We analyzed the pause patterns using the Story-speech corpus. Three pause prediction systems i.e. BS, POS, and U built with different feature sets using CART model. The BS system is made by using basic positional and contextual features with respect to a word/syllable/phone. The POS system uses a full knowledge of POS tags provided by an automatic tagger/shallow parser along with the features of BS system. The pause prediction system, i.e., U built with using the unsupervised features generated by using Latent Semantic analysis. The POS system outperforms the other two BS and U systems in terms of the F1 measure. The increasing order of the performance of the systems: BS < U < POS. Most of the performance of the systems improves by adding the linguistic knowledge such as Parts-of-speech (POS tags). Also, instead of using the linguistic knowledge, one can make use of unsupervised continuous valued features. Also, the inclusion of the SS features improves the performance of the systems (for both U and POS), which helps in capturing the story-semantic information. Possible extensions to the current work are as follows. Apart from Hindi, the current pause prediction study can be extended to other Indian languages such as Bengali, Telugu, Tamil, Marathi, and Malayalam. The unsupervised word features along with the story-specific features can be used to integrate with the proposed Story TTS systems [4] for accurate pause prediction and also for improving the quality of synthesized story speech. Further studies can be performed to analyze the pause patterns present at paragraph level [27] for storytelling style speech. Also, pauses can be explained based on the discourse modes [28].

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).

References

1. N. Braunschweiler, L. Chen, *Automatic detection of Inhalation Breath Pauses for Improved Pause Modelling in HMM -TTS*, in: 8th ISCA Speech Synthesis Workshop, ISCA, Barcelona, Spain, 2013, pp. 1– 6.
2. E. C. O. Alm, *Affect in Text and Speech*, Ph.D. thesis (2008).
3. M. Theune, K. Meijs, D. Heylen, R. Ordelman, *Generating Expressive Speech for Storytelling Applications*, IEEE Transactions on Audio, Speech & Language Processing 14 (4) (2006) 1137–1144.
4. P. Sarkar, A. Haque, A. Dutta, G. Reddy, M. Harikrishna, P. Dhara, R. Verma, P. Narendra, B. S. Sunil, J. Yadav, 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), 2014, pp. 473- 477.
5. P. Taylor, A. W. Black, *Assigning Phrase Breaks from Part-of-speech Sequences*, Computer Speech & Language 12 (2) (1998) pp. 99 -117.
6. K. Yoon, *A Prosodic Phrasing Model for a Korean Text-to-speech Synthesis System*, Computer Speech & Language 20 (1) (2006) pp. 69-79
7. K. Ghosh, 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, Vol. 306 of Communications in Computer and Information Science, Springer Berlin Heidelberg, 2012, pp. 118 – 129.
8. S. Kim, J. Lee, B. Kim, 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, USA, September 17-21, 2006.
9. P. Zervas, M. Maragoudakis, N. Fakotakis, G. Kokkinakis, *Bayesian Induction of Intonational Phrase Breaks*, Eurospeech (2003).
10. D. Klatt, *The KLATTALK Text-to-Speech Conversion System*, in: Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP ’82., Vol. 7, 1982, pp. 1589–1592.
11. A. W. Black, P. Taylor, *The Festival Speech Synthesis System: System Documentation*, Tech. Rep. HCRC/TR-83, Human Communciation Research Centre, University of Edinburgh, Scotland, UK (1997).
12. M. Schroder, *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, A. Black, *Data-driven Phrasing for Speech Synthesis in Low-resource Languages*, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012, pp. 4013 – 4016.
14. N. S. Krishna, H. A. Murthy, *A New Prosodic Phrasing Model for Indian Language Telugu*, in: INTERSPEECH, ISCA, 2004.
15. A. Vadapalli, P. Bhaskararao, 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, ISCA, Barcelona, Spain, 2013, pp. 189 –194.
16. P. Sarkar, K. S. Rao, *Data-Driven Pause Prediction for Speech Synthesis in Storytelling Style Speech*, in: NCC, IEEE, 2015.
17. http://ltrc.iiit.ac.in/showfile.php?filename=downloads/shallow parser.php
18. H. Schutze, *Distributional part-of-speech tagging*, in: In Proc. of 7th Conference of the European Chapter of the Association for Computational Linguistics, 1995.
19. O. Watts, J. Yamagishi, S. King, *Unsupervised Continuous-Valued Word Features for Phrase-Break Prediction without a Part-of- Speech Tagger*, in: INTERSPEECH, ISCA, 2011, pp. 2157–2160.
20. A. Vadapalli, 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, pp. 41–45.
21. C. J. V. Rijsbergen, *Information Retrieval*, 2nd Edition, Butterworth-Heinemann, Newton, MA, USA, 1979.
22. K. Prahallad, A. Black, 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, 2006, pp. I–I.
23. 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.
24. T. K. Landauer, P. W. Foltz, D. Laham, *An Introduction to Latent Semantic Analysis*, Discourse Processes (25) (1998) 259–284.
25. http://en.wikipedia.org/wiki/singular value decomposition.
26. O. Watts, *Unsupervised Learning for Text-to-speech Synthesis*, Ph.D. thesis, University of Edinburgh (2012).
27. K. Prahallad, A. R. Toth, A. W. Black, *Automatic Building of Synthetic Voices from Large Multi-Paragraph Speech Databases*, in: INTERSPEECH, ISCA, 2007, pp. 2901–2904.
28. R. Montao, F. Alas, 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, 2013, pp. 191–196.