Multi-stage Children Story Speech Synthesis

A First Seminar Report
Submitted in partial fulfillment of the requirements for the degree of

Master of Science (by Research)

in

Information Technology

by

Harikrishna D M

[Roll No. 12IT72P07]

Under the supervision of

Dr. K. Sreenivasa Rao

School of Information Technology

Indian Institute of Technology, Kharagpur

Kharagpur-721302, India

Contents

1 Introduction

Synthesizing expressive speech involves embedding natural expressions into speech, according to the semantics present in the text. Story synthesis aims at synthesizing story-style speech from the text using text-to-speech (TTS) systems. In order to synthesize story speech, the following approaches may be explored.

1. *Development of TTS systems using story speech corpus:* This framework consists of developing TTS system using story speech corpus recorded from a professional story-teller.
2. *Rule-based story speech synthesis:* Derive the rule-base specific to story genre. After synthesizing the speech by neutral TTS systems, the derived story specific rules can be used to modify the neutral speech into expressive and naturally sounding story-style speech.

Even though the first approach is easier to synthesize story speech, it requires collection of hours of story speech corpus for different languages from professional artists, which is an expensive and laborious task. The second approach derives rule-base by analyzing perceptual differences between actual stories and stories synthesized by neutral TTS systems. This approach seems quite promising as the story specific rules capture the linguistic information and prosody dynamics of the story.

This work is carried out as part of the requirements of the project titled *"Development of Text-to-Speech systems in Indian languages (Phase - II)"*. The basic objective of the project is to synthesize story style speech from a story text using the neutral text-to-speech (TTS) systems developed in Phase - I of the project. Syllable-based unit selection neutral TTS systems were developed for six Indian languages in Phase - I of the project [1].

2 Related Work

In this section, we have discussed some works related to: (i) feature reduction techniques in text classification, (ii) story classification and story TTS system and (iii) text classification in Indian languages. Text classification is a standard problem in Natural Language Processing (NLP). Classifiers like naive Bayes (NB), K-nearest neighbour (KNN), support vector machines (SVM), neural network (NN), decision trees etc. have been used in most of the works on text classification. SVM has shown consistent performance and outperformed other classifiers [2, 3]. Feature reduction techniques like Latent Semantic Analysis (LSA) [4] and Sparse Representation [5] are explored for text classification.

Within the overall context of a storytelling TTS application, a perceptual study to identify emotions in children stories is carried out in [6]. In [7], children stories are analyzed to perform various tasks like identification of characters, personality attributes of character like age and

gender. In [8], top-n item-to-item recommendation algorithm is used to define clusters of similar stories. In [9], a story classifier is developed to identify a paragraph as story or not using the standard keyword-based features, linguistic features and a new set of semantic features. Their proposed semantic features are based on subject, verb and object triplet's aggregation and generalization.

Over the last decade, there has been significant growth in Indian NLP. In [10], an ontology and hybrid based approach for classification of Punjabi text documents are proposed. Marathi articles are classified using NB, centroid and KNN classifiers in [11]. In [12], Kannada web pages are classification using NB and Maximum Entropy classifiers. In [13], manually collected Kannada sentences from Kannada Wikipedia are classified. Artificial Neural Network (ANN) and Vector Space Model (VSM) are used for classifying Tamil documents in [14]. In [15], Telugu news articles are classified into four categories: Politics, Sports, Business and Cinema using NB classifier. In [16], language independent, corpus-based machine learning techniques are used for text categorization in ten major Indian Languages.

The existing works are mostly limited to text classification in the domains such as news, sports, etc. But, none of the works attempted the story classification in Indian languages. In this work, we are attempting story classification in view of synthesizing story speech.

3 Scope of the Work

Generating an expressive, naturally sounding, story like speech from text using a neutral TTS system is a highly challenging task. We have attempted this task by dividing it into sub-tasks such as: (i) identifying whether the given text is related to story or not, (ii) identifying the story genre from the story text, (iii) identifying emotions specific to story genres, (iv) deriving prosody modification factors (rules) for story specific emotions and (v) synthesizing the story style speech from neutral TTS by incorporating the derived prosody modification factors. In summary, the objective of this work is to generate the expressive story-style speech using neutral TTS system, from the given story text and prosody modification rules for the corresponding story type.

In order to derive story specific prosody modification factors, first we have to identify the story genre information from the given story text. With this motivation, we are exploring classification of children stories into different genres based on story text. In this work, story classification for Hindi and Telugu stories is explored. We are classifying children stories into three story genres: *fable, folk-tale and legend.* For story classification, linguistic based features like Part-of-speech (POS) and Keyword based features like *Term frequency (TF)*, *Term frequency inverse document frequency (TFIDF)* are explored.

The primary research objectives of our work are given as follows:

1. To classify children stories into different genres based on text.
2. To predict the emotion from story text.
3. To derive prosody modification factors for story specific emotions.
4. To synthesize story speech using mark-up language and evaluation of the system.

4 Work Done

4.1 Story Classification Framework

Figure 1 shows the overall framework for story classification. Short stories are collected from blogs and story books. Story corpora are cleaned and stopwords are removed. Using the shallow parser, lemmatization and POS tagging are carried out for the entire story corpora. Feature vectors are computed using the combination of POS and Keyword features with different weighting schemes. Output class labels are predicted using various classifiers. Each block in the framework is explained in the following subsections.

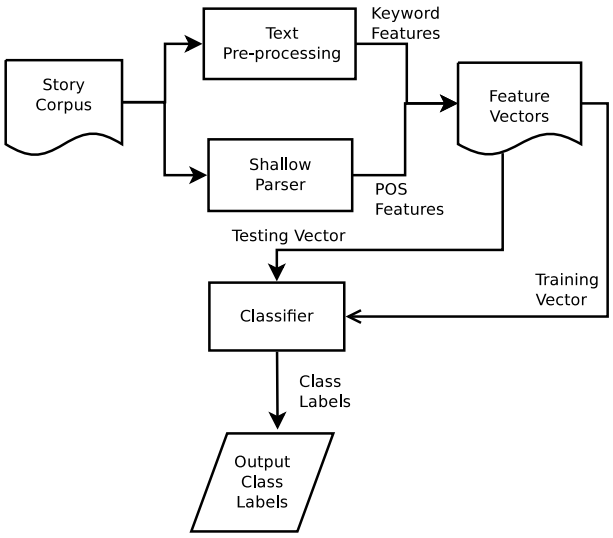

Figure 1: Flow diagram of Story Classification Framework

4.1.1 Database Preparation

Hindi and Telugu story corpora consisting of 300 and 150 short stories respectively, are collected from Blogs[1], Panchatantra and Akbar-Birbal books. No standard story corpora are available for Indian languages. Details of the Hindi and Telugu story corpora are presented in Table 1.

Table 1: Details of Hindi and Telugu Story Corpora

Story genre	Hindi		Telugu	
	# Stories	# Words	# Stories	# Words
Fable	100	50344	50	6668
Folk-tale	100	46900	50	6144
Legend	100	35991	50	8540

4.1.2 Text Pre-processing and POS Tagging

Text pre-processing is done to clean the story corpus which includes stripping multiple white spaces, removing special symbols and numbers. Furthermore, Hindi and Telugu shallow parser[2] developed by IIIT Hyderabad are used for POS tagging and lemmatization i.e. to convert each word into its root word. No standard stopword lists for Hindi and Telugu are available. In this work, a list of 164 and 138 stopwords respectively, for Hindi and Telugu are prepared and used.

4.1.3 Feature Extraction

Linguistic-based features such as *POS density (PD)* and Keyword-based features like *Term frequency (TF), Term frequency inverse document frequency (TFIDF)* are explored. Different combinations of POS and keyword-based features are considered for evaluation. All stories in the collection can be regarded as a *document-term matrix (DTM)*, where each row represents a story and each column represents a term in the collection. "R" statistical programming language is used for feature extraction.

- **Term Frequency (TF)**: Frequency of terms in a story are calculated. TF measure explains the importance of a word within a story genre.
- **Term Frequency Inverse Document Frequency (TFIDF)**: For a term, weight is assigned as the product of TF and IDF. IDF is calculated as

$$idf(t_i) = log\frac{N}{n_i}$$

where N is the total number of stories and n_i is the number of stories in the corpus that contains word t_i. TFIDF measure gives importance of a word across story genre.

[1] http://telugubalalu.blogspot.in/

[2] http://ltrc.iiit.ac.in/showfile.php?filename=downloads/shallow_parser.php

– **POS Density (PD)**: POS tags used here are: Noun (NN), Proper Noun (NNP), Spatial and Temporal Nouns (NST), Pronoun (PRP), Finite Verb (VM), Auxiliary Verb (VAUX), Post Position (PSP), Particles (RP), Adjective (JJ) and Quantifiers (QF). For each document, PD is used as a feature vector. It is calculated as

$$PD = \sum_{p \in P} \frac{count(p)}{Total\ words\ in\ Document}$$

where P = NN, VM, PRP, VAUX, NNP, NST, PSP, RP, JJ and QF.

4.1.4 Classifiers

Five different combinations of features are explored in this work: PD, TF, TFIDF, TF + PD and TFIDF + PD. The Performance of feature selection methods is evaluated using three classifiers: NB, KNN and SVM. Classifier performance is evaluated using 10-fold cross validation. In this work, we used WEKA as a framework combined with LibSVM package for the implementation of these classifiers.

4.1.5 Evaluation Measures

The performance of the classifier is evaluated using *Precision (P), Recall (R), F-measure (F) and Accuracy*. Macro F1 measure is also used as a metric to evaluate the performance. Macro F1 computes a simple average of individual F-measures over classes. McNemar's statistical significance test is used to compare cross-classifier performance.

$$P = \frac{No.\ of\ stories\ correctly\ classified\ as\ class\ \text{“}x\text{”}}{No.\ of\ stories\ classified\ as\ class\ \text{“}x\text{”}}$$

$$R = \frac{No.\ of\ stories\ correctly\ classified\ as\ class\ \text{“}x\text{”}}{Actual\ No.\ of\ stories\ of\ class\ \text{“}x\text{”}}$$

$$F = \frac{2 \times P \times R}{(P + R)}$$

$$Accuracy = \frac{No.\ of\ stories\ correctly\ classified}{Total\ No.\ of\ stories}$$

$$Macro\ F1 = \frac{\sum_{i \in C} F_i}{\mid C \mid}$$

Where C is the set of predefined classes and F_i is the F-measure for the i^{th} class in C.

4.2 Story Classification using Keyword based Features

For moderate collection of stories, the document-term matrix is likely to have thousands of columns (terms). It is observed that most of the entries in DTM are zeros. For effective representation, we can reduce the dimension of the feature vector using appropriate feature dimension reduction techniques. We have experimented with two feature reduction techniques: (i) *Sparse Term Removal* and (ii) *Latent Semantic Analysis (LSA)*.

- **Sparse Term Removal:** It reduces the higher dimensional document-term matrix to a low dimension matrix by removing the sparse terms. Given a sparsity value, terms having sparseness value more than specified value are removed from document-term matrix. We experimented with different sparseness factors (0.7, 0.75, 0.8, 0.85, 0.9, 0.95)
- **Latent Semantic Analysis:** LSA is a method for dimensionality reduction in which original document-term vector space is transformed into a lower dimensional space which captures the behaviour of implicit higher-order structure. LSA is based on the singular value decomposition (SVD) which is commonly used in dimension reduction for document classification. We have experimented with different values of k i.e. (25, 50, 75, 100, 125, 150) and (15, 30, 45, 60, 75, 90) for Hindi and Telugu, respectively.

Tables 2 and 3 represent macro F1 measure for story classification using feature reduction techniques for Hindi and Telugu, respectively. Term frequency (TF) is used as a feature to calculate the macro F1 measure. Dimensions of the document-term matrix are listed in the fourth row of the tables. By increasing the sparseness factor, the most frequently repeated terms in story corpora are included in DTM. Increasing the sparseness factor beyond a threshold can add noisy terms, which do not contribute for identifying the story genre and thus decreases the performance. The poor performance of LSA can be attributed to the failure of capturing the behaviour of implicit higher-order structure by lower dimensional document-term matrix. Out of these two feature reduction techniques, the highest F1 measure is achieved using sparseness factor of **0.9**.

Table 2: Macro F1 measure for story classification using feature reduction techniques for Hindi

Classifiers	Full Story	Dimension Reduction Techniques											
		Sparseness Factor						LSA					
		0.7	0.75	0.8	0.85	0.9	0.95	25	50	75	100	125	150
	300×6608	300×78	300×104	300×143	300×182	300×366	300×681	300×25	300×50	300×75	300×100	300×125	300×150
NB	0.71	0.81	0.83	0.84	0.86	**0.89**	0.84	0.4	0.4	0.41	0.41	**0.43**	0.42
KNN	0.61	0.71	0.73	0.74	0.75	**0.77**	0.73	0.62	0.63	0.63	0.67	**0.68**	0.65
SVM	0.62	0.79	0.82	0.85	0.86	**0.91**	0.82	0.32	0.37	0.41	0.46	**0.48**	0.47

Table 3: Macro F1 measure for story classification using feature reduction techniques for Telugu

Classifiers	Full Story	Dimension Reduction Techniques											
		Sparseness Factor						LSA					
		0.7	0.75	0.8	0.85	0.9	0.95	15	30	45	60	75	90
	150×4539	150×17	150×29	150×49	150×88	150×232	150×582	150×15	150×30	150×45	150×60	150×75	150×90
NB	0.76	0.78	0.8	0.81	0.83	**0.86**	0.8	0.64	0.66	**0.67**	0.61	0.56	0.54
KNN	0.46	0.68	0.7	0.72	0.73	**0.75**	0.71	0.63	0.65	**0.71**	0.63	0.58	0.46
SVM	0.81	0.84	0.85	0.87	0.89	**0.94**	0.87	0.44	0.51	**0.58**	0.56	0.55	0.52

4.3 Story Classification using POS Features

For analysis, top 10 POS tags in terms of frequency of occurrence are selected. Details of the POS tags selected in this work are presented in Table 4. For the investigation of the effect of linguistic information on story classification, we carried out experiments with different combinations of POS tags. The POS groups based on syntactic category are listed in Table 5. Stories are classified based on these POS tag-sets and macro F1 measures for different POS tag-sets are shown in Table 6. Set 2 has the highest macro F1 measure compared to the other POS tag-sets. This result shows the importance of nouns, adjectives and quantifiers for story classification. It is also observed that the performance of POS features are not good as compared to keyword based features.

Table 4: POS distribution across story genres

POS Tags	Hindi			Telugu		
	Fable	Folk-tale	Legend	Fable	Folk-tale	Legend
NN	10975	9985	7277	2539	2386	2957
VM	9298	8439	6098	1919	1730	2377
PSP	6788	6249	4898	104	110	131
PRP	5286	4910	3761	615	557	769
VAUX	4278	3735	2817	40	38	48
JJ	1691	1698	1420	264	217	238
NNP	1534	1497	1554	22	152	516
RP	1456	1353	1011	45	38	86
NST	1035	764	584	275	178	283
QF	635	530	503	61	40	75

Table 5: Different sets of POS tags

Set	POS Tags
Set 1	$\{NN, NNP, NST, PRP, JJ, QF, VM, VAUX, PSP, RP\}$
Set 2	$\{NN, NNP, NST, PRP, JJ, QF\}$
Set 3	$\{NN, NNP, NST, PRP, VM, VAUX\}$
Set 4	$\{NN, NNP, NST, PRP, PSP, RP\}$
Set 5	$\{NN, NNP, NST, PRP\}$
Set 6	$\{JJ, QF, VM, VAUX\}$

Table 6: Macro F1 measures for different sets of POS tags

Set	Hindi			Telugu		
	NB	KNN	SVM	NB	KNN	SVM
Set 1	0.48	0.4	0.45	0.55	0.47	0.56
Set 2	**0.49**	**0.43**	**0.5**	**0.56**	**0.55**	**0.58**
Set 3	0.48	0.4	0.48	0.55	0.51	0.57
Set 4	0.48	0.38	0.47	0.54	0.52	0.56
Set 5	0.45	0.4	0.46	0.53	0.51	0.56
Set 6	0.42	0.33	0.39	0.38	0.38	0.36

4.4 Story Classification using Concatenation of Keyword and POS Features

A detailed performance measures for story classification using the concatenation of keyword and POS features is given in Table 7. PD feature corresponds to POS density of Set 2 POS tags (See Table 5). Story classification accuracy using keyword and POS features is shown in Figure 2. Tables 8 and 9 summarizes the McNemar's statistical significance test results for different combinations of features and cross-classifier performance respectively. The meaning of symbols used in the Tables 8 and 9 are as follows: (i) " $\gg$ " means *P-value* ≤ 0.01, which is extremely statistically significant, (ii) " $>$ " means $0.01 <$ *P-value* ≤ 0.05, which is statistically significant and (iii) " $\sim$ " means *P-value* > 0.05, which is not statistically significant. From the tables, it is evident that adding PD features to TF and TFIDF improves the F-measure slightly. Among the three classifiers explored in this work, the performance of SVM is better than NB and KNN in terms of F-measure for both Hindi and Telugu stories.

The Keyword and POS features are based on terms (words) present in a story. Moreover, while analyzing the story corpora, it is noted that some of the terms are common for both fable and folk-tale. The terms in legend story are easily distinguishable from fable and folk-tale. NB is a probabilistic learning method. It is based on Bayes theorem and the story genre will be assigned to the class having *maximum a posteriori probability*. KNN classifies the story based on the similarities between features i.e terms used in story. The poor performance of KNN can be due to the noisy terms in the stories. SVM performs classification by constructing hyperplane in a high dimensional space that maximizes the margin between classes. SVM is

resilient to noise because the classifier performance depends on the support vectors. Hence SVM has the best accuracy.

Table 7: Performance measures for story classificaiton using concatenation of keyword and POS features

Story Genre	Features	Hindi									Telugu								
		NB			KNN			SVM			NB			KNN			SVM		
		P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
Fable	PD	0.46	0.65	0.54	0.47	0.70	0.57	0.46	0.48	0.48	0.56	0.62	0.59	0.48	0.72	0.58	0.59	0.96	0.73
	TF	0.93	0.88	0.9	0.89	0.59	0.71	0.94	0.91	0.92	0.95	0.8	0.87	0.68	0.7	0.69	0.94	0.94	0.94
	TF + PD	0.93	0.90	**0.91**	0.89	0.68	**0.77**	0.95	0.93	**0.94**	0.98	0.82	**0.89**	0.78	0.76	**0.77**	0.98	0.96	**0.97**
	TFIDF	0.89	0.44	0.59	0.86	0.56	0.68	0.92	0.9	0.91	0.86	0.74	0.8	0.64	0.64	0.64	0.92	0.92	0.92
	TFIDF + PD	0.9	0.75	**0.81**	0.88	0.66	**0.75**	0.94	0.92	**0.93**	0.93	0.78	**0.85**	0.72	0.68	**0.7**	0.94	0.92	**0.93**
Folk-tale	PD	0.63	0.35	0.45	0.38	0.31	0.34	0.52	0.41	0.46	0.46	0.72	0.56	0.55	0.30	0.39	0.58	0.22	0.32
	TF	0.87	0.87	0.87	0.66	0.84	0.74	0.96	0.9	0.93	0.75	0.92	0.83	0.86	0.76	0.8	0.96	0.92	0.94
	TF + PD	0.87	0.90	**0.89**	0.75	0.86	**0.8**	0.97	0.92	**0.94**	0.76	0.94	**0.84**	0.8	0.82	**0.81**	0.98	0.94	**0.96**
	TFIDF	0.76	0.76	0.76	0.65	0.82	0.73	0.94	0.89	0.91	0.74	0.84	0.78	0.78	0.72	0.75	0.94	0.9	0.92
	TFIDF + PD	0.82	0.8	**0.81**	0.7	0.83	**0.76**	0.94	0.9	**0.92**	0.79	0.86	**0.82**	0.76	0.78	**0.77**	0.96	0.92	**0.94**
Legend	PD	0.59	0.39	0.47	0.49	0.34	0.40	0.54	0.54	0.54	0.87	0.40	0.55	0.75	0.60	0.67	0.72	0.62	0.67
	TF	0.87	0.93	0.9	0.85	0.9	0.87	0.85	0.94	0.89	0.91	0.86	0.88	0.72	0.8	0.76	0.92	0.96	0.93
	TF + PD	0.96	0.96	**0.96**	0.84	0.92	**0.88**	0.9	0.96	**0.93**	0.96	0.88	**0.92**	0.84	0.84	**0.84**	0.92	0.98	**0.95**
	TFIDF	0.64	0.96	0.77	0.82	0.9	0.86	0.86	0.92	0.88	0.82	0.82	0.82	0.68	0.74	0.71	0.9	0.94	0.91
	TFIDF + PD	0.74	0.88	**0.8**	0.84	0.91	**0.87**	0.87	0.93	**0.9**	0.81	0.88	**0.84**	0.77	0.8	**0.78**	0.9	0.96	**0.93**

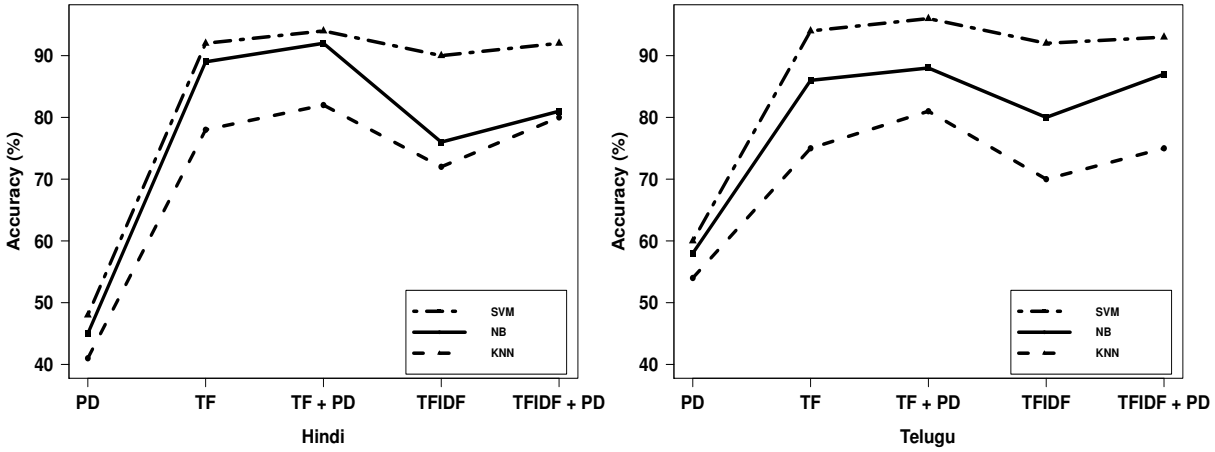

Figure 2: Story classification accuracy using concatenation of keyword and POS features

Table 8: Statistical significance test results for different combination of features

Classifier	Hindi		Telugu	
	TF + PD vs TF	TFIDF + PD vs TFIDF	TF + PD vs TF	TFIDF + PD vs TFIDF
NB	$>$	$\sim$	$>$	$\sim$
KNN	$\sim$	$\sim$	$\sim$	$\sim$
SVM	$>$	$>$	$>$	$>$

Table 9: Statistical significance test results for cross-classifier performance

Classifier A	Classifier B	Hindi					Telugu				
		PD	TF	TF + PD	TFIDF	TFIDF + PD	PD	TF	TF + PD	TFIDF	TFIDF + PD
NB	KNN	$\sim$	$\gg$	$\gg$	$>$	$\sim$	$\sim$	$\gg$	$\gg$	$\gg$	$\gg$
SVM	KNN	$\sim$	$\gg$	$\gg$	$\gg$	$\gg$	$\sim$	$\gg$	$\gg$	$\gg$	$\gg$
SVM	NB	$\sim$	$\sim$	$\sim$	$\gg$	$\gg$	$\sim$	$\gg$	$>$	$\gg$	$>$

5 Summary and Conclusion

In this work, Hindi and Telugu children stories are classified into three genres: fable, folk-tale and legend. The significant contributions of this work are: (i) story classification using keyword based features, (ii) story classification using POS features and (iii) story classification using concatenation of keyword and POS features. Hindi and Telugu story corpora consisting of 300 and 150 short stories are used for experimentation. NB, KNN and SVM classifiers are used to evaluate the story classification performance using 10-fold cross validation. The effectiveness of classifiers is measured using Precision, Recall and F-measure. McNemar's statistical significance test are conducted to compare cross-classifier performance. Results showed that using linguistic information boosts the performance of story classification significantly. From the experiments conducted on the analysis of POS tags, it is observed that POS tag set consisting of nouns, adjectives and quantifiers have the highest accuracy compared to the other tag sets. In most of the cases, the highest performance is achieved by TF + PD features and SVM models outperformed the other models in terms of classification accuracy.

6 Future Work

1. *Story classification using partial story information:* Exploring story classification by dividing stories into parts based on story semantics and also to find the optimal number of sentences required to identify story genres.
2. *Emotion prediction from story text:* Exploring Keyword, POS and story specific features for predicting emotion from story text.

3. *Deriving prosody rules:* Deriving prosody rules (modification factors) specific to emotions and story genres. The prosody modification factors can be derived carefully by analyzing the perceptual differences between synthesized neutral speech utterances and their respective utterances narrated by a storyteller.
4. *Synthesis of story speech using mark-up language:* Story-specific prosody rules can be effectively incorporated using SABLE mark-up language. The quality and naturalness of the synthesized story speech can be evaluated using subjective tests.

Publications

Conference:

1. **Harikrishna D M** and K. Sreenivasa Rao, "Classification of Children Stories in Hindi Using Keywords and POS Density," in *International Conference on Computer Communication and Control (IC4)*, Indore, 2015.

Acknowledgements

We are thankful to the Department of Information Technology, Govt. of India for supporting this research work, *Development of Text-to-Speech synthesis for Indian Languages Phase II*, Ref. no. 11(7)/2011HCC(TDIL)

References

[1] H. A. Patil, T. B. Patel, N. J. Shah, H. B. Sailor, R. Krishnan, G. Kasthuri, T. Nagarajan, L. Christina, N. Kumar, V. Raghavendra *et al.*, "A syllable-based framework for unit selection synthesis in 13 Indian languages," in *Oriental COCOSDA held jointly with Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*. IEEE, pp. 1–8, 2013.

[2] T. Joachims, *Text categorization with support vector machines: Learning with many relevant features*. Springer, 1998.

[3] Y. Yang and X. Liu, "A re-examination of text categorization methods," in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, pp. 42–49, 1999.

[4] A. Moldovan, R. I. Bot, and G. Wanka, "Latent semantic indexing for patent documents," *International Journal of Applied Mathematics and Computer Science*, vol. 15, pp. 551–591, 2005.

[5] T. N. Sainath, S. Maskey, D. Kanevsky, B. Ramabhadran, D. Nahamoo, and J. Hirschberg, "Sparse representations for text categorization," in *INTERSPEECH*, pp. 2266–2269, 2010.

[6] C. O. Alm and R. Sproat, "Perceptions of emotions in expressive storytelling," in *INTERSPEECH*, pp. 533–536, 2005.

[7] E. Iosif and T. Mishra, "From speaker identification to affective analysis: A multi-step system for analyzing children stories," *European Chapter of the ACL (EACL)*, pp. 40–49, 2014.

[8] P. V. Lobo and D. M. De Matos, "Fairy tale corpus organization using latent semantic mapping and an item-to-item top-n recommendation algorithm," in *Language Resources and Evaluation Conference (LREC)*, 2010.

[9] B. Ceran, R. Karad, A. Mandvekar, S. R. Corman, and H. Davulcu, "A semantic triplet based story classifier," in *International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2012.

[10] Nidhi and V. Gupta, "Domain based classification of Punjabi text documents using ontology and hybrid based approach," in *Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing, COLING*, 2012.

[11] M. Patil and P. Game, "Comparison of Marathi text classifiers," *Association of Computer Electronics and Electrical Engineers*, vol. 4, pp. 11–22, 2014.

[12] N. Deepamala and P. R. Kumar, "Text Classification of Kannada webpages using various pre-processing agents," in *Recent Advances in Intelligent Informatics*. Springer, 2014.

[13] R. Jayashree and M. K. Srikanta, "An analysis of sentence level text classification for the Kannada language," in *International Conference of Soft Computing and Pattern Recognition (SoCPaR)*, 2011.

[14] K. Rajan, V. Ramalingam, M. Ganesan, S. Palanivel, and B. Palaniappan, "Automatic classification of Tamil documents using vector space model and artificial neural network," *Expert Systems with Applications*, vol. 36, pp. 10914–10918, 2009.

[15] K. N. Murthy, "Automatic categorization of Telugu news articles," *Department of Computer and Information Sciences*, 2003.

[16] K. Raghuveer and K. N. Murthy, "Text categorization in Indian languages using machine learning approaches," in *Indian International Conference on Artificial Intelligence*, 2007.