Designing Prosody Rule-set for Converting Neutral TTS Speech to storytelling style speech for Indian Languages: Bengali, Hindi and Telugu

Parakrant Sarkar, Arijul Haque, Arup Kumar Dutta, Gurunath Reddy M, Harikrishna D M, Prasenjit Dhara, Rashmi Verma, Narendra N P, Sunil Kr. S B, Jainath Yadav, K. Sreenivasa Rao
School of Information Technology
Indian Institute of Technology Kharagpur, Kharagpur - 721302, West Bengal, India.
Email:{parakrantsarkar, rjlhq05, arupdutta1990, mgurunathreddy, coolhari2006, prasenjitdhara.pd, rashmi9425, narendrasince1987, sunil220552, jaibhu38, ksrao1969 } @gmail.com

***Abstract*—This paper provides a design of prosody rule-set for transforming the neutral speech synthesized by Text-to-Speech (TTS) system to storytelling style speech. The objective of this work is to synthesize storyteller speech from the neutral TTS system for a given story text as input. In this work, neutral TTS refers to TTS system developed using Festival framework with neutral speech corpus. For generating storyteller speech from neutral TTS, we are proposing modifications to various prosodic parameters of neutral synthesized speech. In this work, the prosodic parameters considered for modification are (i) pitch contour, (ii) duration patterns, (iii) intensity patterns, (iv) pause patterns and (v) tempo. We have designed individual rule-sets for the above mentioned prosodic parameters, separately for three Indian languages Bengali, Hindi and Telugu. The rule-sets are designed carefully by analyzing the perceptual differences between synthesized neutral speech utterances and their respective natural (original) spoken utterances, narrated by a storyteller. The designed prosody rule-sets are evaluated using subjective listening tests. The results of the perceptual evaluation indicate that the designed prosody rule-sets play a significant role in achieving the story-specific style during conversion from neutral to storytelling style speech.**

Keywords*—Neutral TTS, Storytelling style, Prosody rule-set, Expressive speech synthesis, emotion-salient words, Story specific prosody generation, story-specific emotion detection, story-specific prosody incorporation.

I. INTRODUCTION

In recent years, various real life applications such as telecommunication services, aids for handicapped persons and talking books for children have drawn the attention to pursue research in the area of expressive speech synthesis. The storytelling application is one of them, in which the story text is given as input to text-to-speech (TTS) system, and the synthesized speech should bear the storytelling style [1]. In an ideal case, story TTS should give a listening experience that is equally alluring as a human storyteller. In order to achieve this, we need more expressive, engaging and entertaining style speech, which cannot be produced by existing TTS systems. Earlier works attempted to improve the quality of the synthesized speech by incorporating basic emotions [2]. In [3], a TTS-based digital storyteller which synthesizes stories using control tags was proposed. These tags were used to control the emotional state of the character as well as its physical gestures. A digital storyteller system, ESPER [4] was developed to achieve all emotion expressivity and render dialogues between two characters in a story. In [5], an interactive virtual storyteller was developed, which takes listeners feedback and modifies the prosody to incorporate emotions in speech synthesized by the neutral TTS. An interactive, fairy-tale storyteller system was discussed in [6], in which emotions were classified as prosody and word semantic. The rule based approaches for incorporating basic emotions in the speech synthesized by the TTS has been carried out in English [4], Dutch [7], Spanish [8], Catalan [9], German [10] and Korean [6] languages. These rule based methodologies were based on sentence level utterance analysis and prosody modifications.

In this work, we have analyzed the storytelling style at phrase-level instead of sentence-level. This is because humans can easily perceive emotions at phrase-level as compared to sentence-level due to presence of emotion-salient words. We have even focused on the analysis of the sub-segments present within a phrase, which is not explored by the existing works. In this paper, we have used neutral TTS systems of three Indian languages: Bengali, Hindi and Telugu. Our goal is to synthesize storytelling style speech from the neutral TTS. The prosody rule-set for each of the emotions considered in this work are derived by carrying out perceptual analysis of the differences in the story told by the human storyteller and the same story synthesized by the neutral TTS. The prosodic parameters used for analysis and modification are (i) pitch contour, (ii) duration patterns, (iii) intensity patterns, (iv) pause patterns and (v) tempo.

Rest of the paper is organized as follows. The details of baseline TTS systems used in this work are given in Section II. Section III discusses the design of proposed prosody rule-sets for converting neutral speech into storytelling style speech. Perceptual evaluation of synthesized story speech using listening tests is provided in Section IV.The overall work has been summarized and concluded in the section V.

II. BASELINE TTS SYSTEMS

In this work, baseline TTS systems refer to syllable based TTS [11], 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)*. We considered three TTS systems corresponding to Bengali, Hindi and Telugu languages for generating storytelling style speech, which were developed by IIT Kharagpur, IIT Madras and IIIT Hyderabad, respectively. Baseline TTS systems were developed using approximately four hours of speech corpus recorded by a professional female speaker across three languages. The text is chosen from various sources such as newspapers and story books.

III. DESIGN OF PROSODY RULE-SET FOR GENERATING STORYTELLING STYLE SPEECH FROM NEUTRAL SYNTHESIZED SPEECH

This section discusses the process involved in deriving the prosody rule-set for transforming the neutral synthesized speech to storytelling style speech. One can generate storytelling style speech using Festival-based TTS systems with story speech corpus. However, collecting story speech corpus of 4 hours duration in each language by professional storytellers is a hectic task. Besides that, manually labeling the huge speech corpus is quite tedious. Therefore, we want to make use of existing neutral TTS systems with some additional modules to synthesize storytelling style speech. With this modification, we can make use of existing TTS systems for generating neutral speech as well as storytelling speech.

The additional modules that need to be added to neutral TTS for synthesizing storytelling speech are: (i) Story-specific emotion detection (SSED) module, (ii) Story-specific prosody generation (SSPG) module and (iii) Story-specific prosody incorporation (SSPI) module. The whole process of story synthesis is shown in Fig. 1. The function of SSED module is to detect the story-specific emotions associated to each of the phrases in the story text, based on story semantic information. SSPG module consists of prosody rule-sets for each of the story-specific emotions considered. Based on the story-specific emotion detected by SSED module, appropriate prosodic rules are activated, and the desired modification factors associated with various prosodic parameters are specified by SSPG module. The prosody modification factors suggested by SSPG module are incorporated on synthesized neutral speech by using prosody modification methods present in SSPI module. In this paper, we mainly focus on SSPG module (shown in Fig. 2) which deals with design and derivation of prosodic rules for each story-specific emotion.

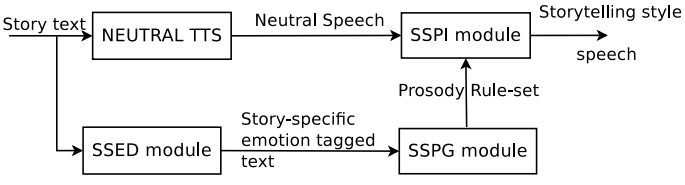

Fig. 1. Block diagram of story synthesis

For designing the prosodic rules for different story-specific emotions, first we need to have some reference, indicating the high quality, naturally spoken stories narrated by a human storyteller. With respect to these reference stories, we have to derive the prosodic rules for transforming the synthesized neutral speech to storytelling style speech. In this regard, we have collected story speech corpus indicating the reference stories, which represent our desired target.

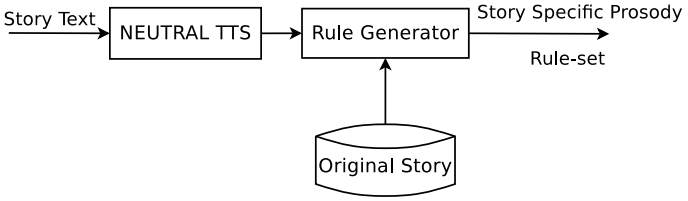

Fig. 2. Block diagram of story-specific prosody generation module

A. Story Speech Corpus

In this work, we have collected story speech corpus in three Indian languages: Bengali, Hindi and Telugu. Story speech corpus in each language is collected from a professional radio artist. For maintaining high quality in the collected story speech corpus, continuous feedback is given to storyteller for improving the quality of narrated story. In each language, the number of stories recorded are approximately 125, and the duration of the corpus is about 3 hours. Most of the stories are children's stories taken from the Internet and story books like Panchatantra and Akbar-Birbal. The number of sentences in a story vary from 15 to 25. The recorded story speech corpus is manually labeled at syllable level. The speech corpus is organized into wave files at sentence and phrase-level. Along with wave files, the supported transcription files are also available. Each wave file along with transcription is also tagged with one of the five story-specific emotions. The phrases are manually annotated by listening to wave files based on the presence of emotion-salient words. The story speech corpus is recorded at 48 kHz with 16-bit resolution and later down sampled to 16 kHz for analysis and processing.

B. Design of story-specific Prosody Rules

The main goal of designing a story-specific prosody rule-set is to process the neutral synthesized speech to obtain the storytelling speech. Here rules are designed by careful perceptual analysis of differences between neutral synthesized speech and its corresponding reference story speech. From story speech corpus, it is observed that different parts of a story are narrated in different styles mostly 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 general, the beginning and ending parts of a story are almost devoid of emotions, while the rest of the story is not. In our story speech corpus, we have observed 5 different story-specific emotions: anger, fear, happy, sad and neutral. We have considered five different prosodic parameters (pitch contour, duration, intensity, tempo and pause duration) for deriving the rule sets as well as incorporating them in neutral synthesized speech. Among these five prosodic parameters pitch contour, duration, and intensity are used together to capture the story-specific emotions. In emotionless story segments, tempo is considered to represent the general storytelling style. Pauses of different durations are used to generate suspense and climax.

The differences between the prosodic parameters of synthesized neutral speech to reference story speech are formulated in the form of rules, specific to each storytelling style. To quantify the variation between neutral and story speech, we have analyzed 20 phrases from each category of storytelling style. The 20 phrases chosen from story speech corpus and their respective neutral speech utterances synthesized from existing TTS systems are perceptually analyzed pairwise. In the analysis, it is observed that the synthesized neutral speech is quite intelligible, but the desired natural storytelling style is completely missing. This is mainly due to the inherent characteristics of the speech corpus which is used for building the neutral TTS. The speech corpus of the neutral TTS is recorded in a monotonous style, by giving equal emphasis to all the words and syllables present in sentences. The sentences in the corpus are independent, and the adjacent sentences may not be semantically related. As a result, the story synthesized by existing neutral TTS systems is found to be unnatural.From the general observation, it is found that storytelling speech has more dynamics in prosodic parameters (such as pitch, duration ans intensity) compared to synthesized neutral speech. Incorporation of variations in prosodic parameters in the synthesized neutral speech may provide the desired natural storytelling style.

For quantifying the prosodic variations, first we determined the absolute difference between neutral and story speech utterances, and we compensated the difference by modifying the prosodic parameters of neutral speech according to the target story speech. From the perceptual analysis it is observed that, even though by replacing the prosody parameters of neutral speech by the target story-specific prosody, the modified speech does not sound like target story speech. The main reason may be due to dominance of speaker specific prosody over story-specific prosody. Since the speakers of neutral and story speech corpus are different, we cannot simply replace prosodic parameters of neutral speech by story speech. The inherent prosodic patterns of one speaker is very different from another. But, while listening to stories narrated by any professional storyteller, we will appreciate the style of story. From this we can intuitively say that there exists some common style while narrating stories across the storytellers beyond speaker specific characteristics. We need to extract that story-specific style beyond speaker specific characteristics in the form of prosody rule-set for each story-specific emotion.

In order to derive a set of prosodic rules, we modified the pitch contour, duration and intensity patterns of neutral speech iteratively with a trial and error approach. For each trial, we used to listen to both modified and target speech utterances. Based on difference in perception, the modification factors are adjusted manually. The iteration process will be terminated whenever we are satisfied with the quality of modified speech (i.e. perceptual difference between target and modified speech utterances is minimum). In the process of designing rule-set, we have observed that suggesting modification factors uniformly across the phrase may not provide the desired style effectively. In general, it is observed that humans impose prosody non-uniformly along the phrase, while speaking the utterance. Therefore, in this study we have analyzed the prosodic aspects of the phrase at three levels (i) initial words of the phrase, (ii) middle words of the phrase and (iii) final words of the phrase. By analyzing the prosodic aspects separately for initial, middle and final parts of the phrase, we are ensuring the presence of non-uniform prosody across the phrase and retaining the same in our prosody rule-set design.

1) Pitch Contour: In contrast to neutral speech, the dynamics of pitch (temporal variations in pitch) contour is very crucial in storytelling speech. The pitch contours of story speech are found to have rising or falling patterns (or both) with respect to neutral speech. Even mean pitch of story speech is also found to be higher or lower when compared to neutral speech based on specific emotion. To formulate the rise-fall patterns of story speech, we explored various non-linear functions to model the pitch patterns of target story speech. Among various non-linear functions, sinusoidal function seems to be more promising. Hence we formulated the pitch patterns for the target story speech using sine functions. Based on the trends observed in pitch patterns, formulation based on sine function is derived for anger, fear, happy and sad emotions. In general, pitch patterns in anger, fear and happy emotions are observed to consist of rise patterns. For sad emotion, pitch patterns are observed to have both rise and fall patterns compared to neutral speech. Sine function is chosen to characterize the intonation trend in four emotions. Even though same function is used to model all four emotions, variations among the four emotions are characterized separately using two pairs of constants namely a and b. The formula for converting the pitch contour of neutral speech to the desired story-specific style for the four emotions are given below.

$$y'(t) = y(t)[1 + a * sin(\frac{(t - t1)}{(t2 - t1)} * b * \pi)] \qquad (1)$$

where $y(t)$ and $y'(t)$ are the original pitch value and the modified pitch value at time t; $t1$ and $t2$ are start and end time instants of a speech segment where pitch modification is supposed to be perform; a controls desired maximum shift in the pitch contour from the average pitch and b is the constant used for determining whether the sine function is constantly increasing or rising and then falling.

The values of a and b for the four story-specific emotions are given in columns 2, 5 and 8 of Table I. These values are derived by analyzing 20 phrases for each story-specific emotion collected from the story speech corpus. For each phrase, both neutral utterance synthesized by neutral TTS and story speech utterance narrated by a storyteller are analyzed perceptually. The neutral utterance is modified by trial and error approach with different values of a and b in equation 1. The values of a and b are finalized for that phrase, whenever the modified neutral speech and original story speech utterances sound perceptually similar. Likewise, a and b values are determined for the remaining phrases. Now from these 20 sets of a and b values, a single set of values are derived by conducting a listening test. Subjects are made to listen to each of these 20 phrases, and the a and b values corresponding to the phrase most preferred by them in terms of emotional content and accuracy are taken as the final values of a and b for that emotion. This procedure is repeated for computing a and b values for all four emotions at initial, middle and final words of the phrase within a sentence.

2) Duration: In order to efficiently impose the dynamics of story speech prosody on neutral speech, pitch, duration

and intensity should be modified to appropriate levels. To obtain duration modification factors for four story-specific emotions, variations of duration patterns between neutral and story speech utterances are analyzed. For deriving the duration modification factors, we followed the same procedure similar to pitch contour modification. Here we are scaling each part i.e. initial, middle and final words of a phrase with some constant modification factors. These factors are shown in columns 3, 6 and 9 of Table I.

3) Intensity: We observed in the original story that the storyteller increases intensity between 2-6 dB relative to the average intensity of the neutral speech synthesized by TTS system. The incorporation of desired intensity is carried out after modifying the neutral speech with suitable pitch and duration modification factors. The desired intensity modifications are done by adding x dB to the average intensity for initial, middle and final words of the phrase. The x values are given in columns 4, 7 and 10 of Table I.

4) Tempo (Speaking Rate): Tempo has a specific relevance for storytelling style speech. From the perceptual analysis, we found that the tempo has to be varied gradually for those phrases, which does not exhibit the story-specific emotion across three different languages. The duration of the initial, middle and final part of these phrases are modified. The modification factors are shown in column 3, 6 and 9 of Table I.

5) Pauses: The duration of pauses is more in the original story as compared to story synthesized by neutral TTS. In the context of storyteller speech, pauses are used for separating phrases, emphasizing keywords and emotion-salient words to introduce suspense and climax in the story. In this work, we considered three types of pauses according to their durations. These are (i) *Long pause* (>250 ms) (ii) *Medium pause* (>150 ms) and (iii) *Short pause* (<150 ms)

IV. EVALUATION OF PROPOSED PROSODY RULE-SETS

In this work, the proposed prosody rule-sets are evaluated subjectively using degradation mean opinion score (DMOS) [12] and story-specific emotion recognition tests. Listening tests based on DMOS were conducted by using synthesized story speech samples and their respective original stories spoken by storyteller. The quality of synthesized story speech depends on the quality of the original story speech produced by storyteller. The score obtained for synthesized story speech is therefore normalized to that of natural story speech. This is referred to as degradation mean opinion score.

For listening tests, three stories are considered from each language, which are not used for deriving the prosody rule-sets. The stories contain minimum of five phrases in each of the five story-specific emotions. For evaluation purpose, five fragments per emotion are considered, and it results to 25 fragments from five emotions. A fragment consists of one or two sentences, out of which one phrase is associated to one of the five story-specific emotions. For DMOS test, we collected a total of 75 fragments, out of which 25 fragments from the following three sources: (i) neutral synthesized speech, (ii)story style synthesized speech (modified neutral speech by incorporating prosody rules) and (iii) original stories spoken by storyteller. These 75 fragments were shuffled randomly and supplied to each subject for evaluation. The subjects evaluated the quality of speech in terms of naturalness and story style on a five-point scale (1: very poor, 2: poor, 3: fair, 4: good and 5: excellent). In addition to this, the subjects were also asked to identify the emotion associated to each fragment. For each language, 10 research scholars in the age group of 20-35 participated in the listening tests.

From the obtained scores, DMOS is computed for each emotion category for three Indian languages: Bengali, Hindi and Telugu. In addition to this, the subjects were also asked to recognize the emotion associated to each fragment of story-style synthesized speech. This is termed as Recognition Accuracy (RA) and it is computed for each of the five emotions. DMOS is computed for both neutral speech (NS) synthesized by TTS systems and modified speech (MS) obtained by applying the prosody rule-set. RA is computed only for modified speech. The computed DMOS and RA values are given in columns 2-7 and 8-10 of Table II respectively. From the results, it is observed that both DMOS and RA are better in case of synthesized story style speech, compared to synthesized neutral speech. From the results, we can observe that anger and sad have better RA and DMOS compared to other emotions. The incorporation of appropriate intensity patterns in all the emotions have resulted in a better performance for anger and sad emotions compared to others. This is because intensity plays a major role in anger and sad emotions compared to other emotions, while the other parameters like duration and pitch have almost uniform contributions across all emotions.

The demos of the story TTS systems in the three Indian languages i.e. Hindi, Bengali and Telugu are available at our website $sitspeech.iitkgp.ac.in$.

TABLE II. RESULT OF SUBJECTIVE EVALUATION

	DMOS						Recognition Accuracy (%)		
	Bengali		Hindi		Telegu				
Emotion	NS	MS	NS	MS	NS	MS	Bengali	Hindi	Telegu
Anger	3.03	3.13	3.05	3.77	3.02	3.2	54	64	56
Happy	2.82	2.45	2.4	2.6	2.9	2.84	46	38	40
Fear	2.71	2.91	2.95	2.6	2.83	2.92	26	36	44
Sad	3.08	3.16	3.03	3.16	2.94	3.0	52	58	54
Neutral	3.03	3.17	3.48	3.54	3.4	3.5	76	84	80

V. SUMMARY AND CONCLUSION

In this paper, we have designed prosody rule-sets for converting neutral speech synthesized by TTS system to story style speech. In this study, we considered five prosodic parameters, namely, pitch, duration, intensity, tempo and pause durations for analysis and modification. The prosody rules were derived by analyzing the difference between prosodic parameters of neutral and storyteller speech utterances. The perceptual evaluation results indicated that the proposed prosody rule-sets were effective in imposing the story style on neutral speech utterances. The design of prosody rule-sets may be extended to other Indian languages. Further studies may be carried out to examine the commonalities and differences among the prosody rule-sets across the languages. Prosody rules can be refined for improving the quality of storyteller speech. The existing prosody modification methods could be improved to enhance the quality of synthesized story speech.

TABLE I. PROSODIC MODIFICATION FACTORS FOR CONVERTING PHRASE FROM NEUTRAL STYLE TO STORY-SPECIFIC STYLE. (NC: NO CHANGE)

	Modification Factors								
	Bengali			Hindi			Telugu		
Word Position	Pitch	Duration	Intensity(dB)	Pitch	Duration	Intensity(dB)	Pitch	Duration	Intensity(dB)
Anger									
Initial words	a=0.1, b=0.1	0.95	+2	a=0.2, b=0.3	0.78	+4	a=0.25, b=0.4	0.95	+3
Middle words	NC	NC	NC	NC	NC	NC	a=0.2, b=0.4	0.80	+4
Final words	a=0.1, b=0.1	0.95	+2	a=0.3, b=0.5	0.88	+4	NC	NC	NC
Happy									
Initial words	NC	NC	NC	NC	NC	NC	a=0.2, b=0.7	0.9	+2
Middle words	a=0.23, b=0.73	0.95	+1	a=0.2, b=0.7	0.9	+2	a=0.2, b=0.7	1.05	+2
Final words	a=0.23, b=0.63	1.149	+1	a=0.2, b=0.6	1.15	+2	NC	NC	NC
Fear									
Initial words	NC	NC	NC	NC	NC	NC	a=0.35, b=0.5	0.85	+2
Middle words	a=0.185, b=0.585	1.098	+2	a=0.2, b=0.6	1	+2	a=0.4, b=0.5	0.95	+3
Final words	a=0.185, b=0.585	0.981	-2	a=0.2, b=0.6	1.1	-2	NC	NC	NC
Sad									
Initial words	NC	NC	NC	NC	NC	NC	a=0.1, b=0.9	1.02	+2
Middle words	a=0.18, b=0.6	1.095	-2	a=0.2, b=0.6	1.15	-2	a=0.1, b=0.9	1.25	+2
Final words	a=0.18, b=0.6	1.005	-3	a=0.3, b=0.3	1.05	-3	NC	NC	NC
Neutral									
Initial words	NC	0.88	NC	NC	0.80	NC	NC	0.90	NC
Middle words	NC	NC	NC	NC	NC	NC	NC	NC	NC
Final words	NC	1.12	NC	NC	1.20	NC	NC	1.10	NC

Acknowledgements

The authors would like to thank the Department of Information Technology, Government of India, for funding the project, *Development of Text-to-Speech synthesis for Indian Languages Phase II*, Ref. no. 11(7)/2011HCC(TDIL). The authors would like to thank all consortium partners for providing the TTS systems developed in Phase-I. The authors would also like to thank all the participants for the listening tests.

REFERENCES

[1] 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

[2] M. Schröder, "Emotional speech synthesis: a review." in *INTERSPEECH*, P. Dalsgaard, B. Lindberg, H. Benner, and Z.-H. Tan, Eds. ISCA, 2001, pp. 561–564. [Online]. Available: http://dblp.uni-trier.de/db/conf/interspeech/interspeech2001.html

[3] A. Silva, M. Vala, A. Paiva, and R. A. Redol, "The storyteller: Building a synthetic character that tells stories," in *in Proc. Workshop Multimodal Communication and Context in Embodied Agents*, Montreal,QC, Canada, May/June 2001, pp. 53–58.

[4] J. Y. Zhang, A. W. Black, and R. Sproat, "Identifying speakers in children's stories for speech synthesis," in *INTERSPEECH*, Geneva, Switzerland, 2003.

[5] A. Silva, G. Raimundo, C. de Melo, and A. Paiva, "Recent improvements to the ibm trainable speech synthesis system," in *ICASSP*, vol. 1, 2003, pp. I708– I711.

[6] H. J. Lee, "Fairy tale storytelling system: Using both prosody and text for emotional speech synthesis," *Communications in Computer and Information Science*, vol. 310, pp. 317–324, 2012.

[7] S. J. Mozziconacci, "Speech variability and emotion: production and perception (ph.d thesis)," Ph.D. dissertation, Technical University Eindhoven, 1998.

[8] J. M. Montero, J. M. Gutirrez-Arriola, S. E. Palazuelos, E. Enrquez, S. Aguilera, and J. M. Pardo, "Emotional speech synthesis: from speech database to tts." in *ICSLP*. ISCA, 1998. [Online]. Available: http://dblp.uni-trier.de/db/conf/interspeech/icslp1998.html

[9] I. I. Sanz, F. Alías, J. Melenchón, and M. A. Llorca, "Modeling and synthesizing emotional speech for catalan text-to-speech synthesis," in *ADS*, 2004, pp. 197–208.

[10] M. Schröder, "Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions." in *ADS*, ser. Lecture Notes in Computer Science, vol. 3068. Kloster Irsee, Germany,: Springer, 2004, pp. 209–220. [Online]. Available: http://dblp.uni-trier.de/db/conf/ads/ads2004.html

[11] 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.

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