Corpus Based Emotional Speech Synthesis in Hindi

Ravi Kalyan Bhakat, N.P. Narendra, and Krothapalli Sreenivasa Rao

Indian Institute of Technology Kharagpur, Kharagpur - 721302, West Bengal, India
{mailtome.rob,narendrasince1987}@gmail.com, ksrao@iitkgp.ac.in

Abstract. This paper explores a unit selection based concatenative approach towards emotional speech synthesis in Hindi. The emotions explored are sad and neutral. The Festival framework is used as the underlying Text-To-Speech (TTS) system. The various steps which are followed to create a new voice in Festival are described here. The developed TTS systems are evaluated by subjective evaluation tests. These tests indicate a significant improvement in the quality of synthesis after necessary prosody modifications. Finally, possible improvements which can be made on the systems are put forward.

Keywords: Emotional Speech Synthesis, Festival, Text to Speech, Unit Selection, Corpus based synthesis, Prosody modification.

1 Introduction

The rapid increase in the scale of human computer interaction in the recent past is evident for all to see. Starting from touch screens to bio metrics, gesture recognition to voice commands, the applications are incredible. Speech applications are a big part of this genre of computing. There has been considerable research on Text-To-Speech (TTS) synthesis. The speech synthesized by some TTS systems almost mimics human speech. Emotional Text-To-Speech (ETTS) synthesis however, is an area of great interest. The prosodic characteristics of speech, i.e. pitch, energy and duration vary with emotions. Incorporating these variations into speech while preserving its naturalness is a major issue. There are various methods that are used for ETTS synthesis. A rule based approach, which works with rules to change the voice parameters and impart emotions to speech was proposed in [1]. There are corpus based approaches which have a database of audio files and transcriptions of various emotions. The audio files are segmented into units and based on the text input, the required units are picked up and concatenated to form the speech [2] [3]. This approach gives us much more natural speech as compared to the rule based approach.

In this paper, we describe the development of corpus based TTS systems in Hindi which synthesize sad and neutral speech. They are syllable based speech synthesizers developed on the Festival TTS platform [4]. The baseline systems are improved by adding positional information to the syllables [2]. A fall-back system is incorporated to address the missing unit problem [2]. The systems are

P. Maji et al. (Eds.): PReMI 2013, LNCS 8251, pp. 390–395, 2013.

further improved by performing appropriate prosody modifications on the synthesized speech. The rest of the paper is organized as follows: Section 2 describes the speech corpora. Section 3 presents an overview of Festival and discusses the steps needed to create a new voice. Section 4 outlines some improvements made on the baseline system. Section 5 discusses the observations made on the systems. Section 6 discusses some possible improvements which can be made on the systems.

2 Details of the Emotional Speech Corpora

A sad and a neutral corpus of 1916 sentences each, serve as our database. The sentences were chosen from works of literature, story books and news articles. The corpora were selected based on maximum unit coverage by an optimal text selection algorithm [2]. The emotions reflected in the sentences were enacted by a professional female artist. The speech files were sampled at 16kHz and stored at 16bit PCM data format as .wav files. The corpora were cleaned by removing noisy and repeated files. The details of the corpora are given below:

1. **Neutral corpus:** Contains 1889 sentences with a duration of 2 hrs 54 mins and has a coverage of 4024 unique syllables.
2. **Sad corpus:** Contains 1915 sentences with a duration of 2 hrs 44 mins and has a coverage of 4107 unique syllables.

3 Building a Voice in Festival

The TTS systems are built on the Festival [5] framework using FestVox [6] and Speech Tools [7] as additional tools. The major issues addressed while building a new voice on Festival are enumerated below.

3.1 Selection of the Basic Unit

The selection of the basic unit in concatenative synthesis is an important decision. A proper choice results high quality speech with a reasonably sized corpus. Phones, syllables, words, phrases and sentences are some of the possible basic units. We have taken the syllable as the basic unit. In most Indian languages, as in Hindi, syllables correspond to written units. This makes the grapheme to phoneme conversion easier. A syllable is a string that can be generated by the regular expression C^*VC^*. Using this, syllables can be extracted from the input text. Figure 1 illustrates this concept with a sample utterance, *raajkumaari*, in Hindi. This consists of four syllables, *raaj*, *ku*, *maa* and *ri*. In the corpus, the speech files are segmented at the syllable level as illustrated.

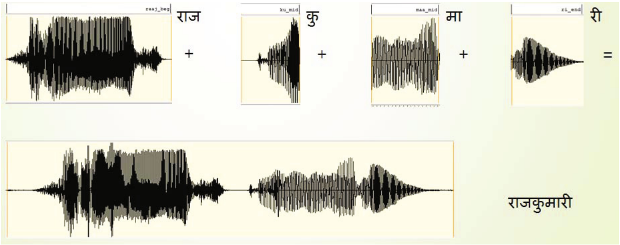

Fig. 1. Component syllables of the utterance *raajkumaari*

3.2 Labelling of the Corpus

The corpus needs to be labelled properly with accurate syllable boundaries for high quality synthesis. From the transcription, a phone level labelling is performed by an ergodic Hidden Markov Model [8] (eHMM) toolkit available on Festival. From the phone level segments, syllable level segments are extracted using syllabification rules. These labels are prone to alignment errors and need to be corrected to get exact syllable boundaries. The labels were manually adjusted to obtain proper syllable boundaries and reduce distortions in the synthesis.

3.3 Prosody Modelling

Prosody models are developed to predict the duration of the syllables using Classification and Regression Trees (CART). Phonetic, contextual and positional features are used in the process. Phonetic features include the positions of the various articulators in the vocal tract while the unit is spoken. Contextual features include the identity of the previous and the next unit. Positional features include the position of the unit in the word.

3.4 Building Unit Clusters

Each of the corpora have a coverage of around 4,100 syllables which are grouped together based on the acoustic differences. This helps in efficient retrieval of the target units. The clustering is done by *wagon* [5] by taking into account the phonological and positional features of the units. The given text input is first parsed and divided into syllables. The prosody models generate the target prosody of the units. From the unit clusters, a set of units is returned for each of the syllables based on the target cost. The unit selection module selects the best sequence of units from these sets which can be concatenated with the minimum possible distortion based on the concatenation cost. These costs can be optimized for better quality of synthesis [9] [10].

4 Improvements Made on the Baseline System

The prosody of the baseline syllable based TTS systems were not satisfactory. Firstly, this was because units from word beginnings were used to create word ends and vice-versa. To address this, each of the syllables were tagged with their position such as *_beg*, *_mid* and *_end*. This ensured that positionally correct units were used during synthesis. Secondly, syllables absent in the database could not be synthesized. To address this, 500 sentences with phone level segmentation were kept in the corpus as a fall back. Thirdly, some neutral sentences had unsatisfactory prosody. The quality was improved by modifying the prosody with algorithms for pitch and duration modification [11] [12], speech normalization and energy correction. Some of these algorithms were integrated into the Festival framework for increased automation. A set of generic rules for prosody modification was generated by evaluating 300 neutral sentences and performing experiments to correct the observed distortions in the syllables. Table 1 outlines the rules which suggest the appropriate prosody modifications for the unsatisfactory units.

Table 1. Some rules generated to facilitate prosody modification of neutral sentences

Syllable	**Problem noted**	**Modification required**
me,ya,te,ti,ke,ki	higher pitch	pitch mod factor = 0.769
ne	higher pitch	pitch mod factor = 0.714
la,se	higher pitch	pitch mod factor = 0.833
kar,par,ab,le,de	lower pitch	pitch mod factor = 1.25
kha	lower pitch	pitch mod factor = 1.11
ka	short duration	duration mod factor = 1.3
ke	short duration	duration mod factor = 1.5
te,la	short duration	duration mod factor = 1.4
ge,pag,mag,te,re, ka,ya,le,me,tha,ne	loud	modify the intensity

5 System Evaluation and Observations

The evaluation of the systems was carried out in four phases. In the first phase, 500 sentences present in the corpus were synthesized by both the sad and neutral systems. Most of them were observed to be identical copies of the corpus files.

In the second phase, the emotional TTS systems were evaluated on intelligibility. A subjective evaluation was carried out with 15 subjects listening to 21 sentences synthesized by both the sad and neutral systems. All the subjects were between 21 to 30 years of age. Their Mean Opinion Score (MOS) were provided based on the scale given in Table 2. The neutral system was rated **3.388/5.00** and the sad system was rated **3.648/5.00**.

In the third phase, the emotion content in the sad sentences was evaluated. The subjects listened to the sad sentences and provided their scores based on the scale in Table 3. The sad system was rated **3.772/5.00**. Sentences which were not actually sad and simulated emotion received lower scores. On the other hand, inherently sad sentences received high scores indicating the natural integration of emotion during synthesis.

Table 2. Criteria for MOS scores on intelligibility of synthesized speech

MOS Score	Criteria
1	Very poor intelligibility, not understandable at all
2	Poor intelligibility, but some parts can be understood well
3	Average intelligibility
4	Good quality of speech, a few distortions
5	Very good quality of speech, can be understood easily

Table 3. Criteria for MOS scores on emotion content

MOS Score	Criteria
1	No emotion at all
2	Emotion evident in few parts of the sentence
3	Emotion present, but not perceived on first few hearings
4	Good quality of emotional speech, few unnatural parts
5	Very good quality of emotional speech, almost natural

From the above evaluations, it was noted that some neutral sentences had unwanted prosodic variations. Analysis revealed that there were some pitch and energy fluctuations that caused the sentences to sound unnatural. These issues were addressed by the prosody modification algorithms discussed in Section 4.

The fourth phase of system evaluation was conducted on the neutral system after the prosody modification algorithms were applied on the unsatisfactory sentences. A subjective evaluation was performed with 15 subjects listening to 10 modified sentences and providing their scores based on the scale in Table 2. They were rated **3.98/5.00** as compared to the earlier rating of 3.388/5.00. This indicates the effectiveness of the prosody modifications that were performed.

6 Summary and Future Work

In this work, systems which synthesize sad and neutral speech in Hindi were built on the Festival TTS system with syllables as basic units. Positional tagging and a fallback system system were used to improve the baseline TTS systems. Prosodically unsatisfactory sentences were modified using appropriate prosody modification algorithms to enhance their quality. Subjective evaluation of the systems were carried out and the results were discussed.

A better tagging approach with sentence level positional information may be explored for better synthesis. Selecting better units during synthesis can result in the lesser prosodic variations. This can be achieved by predicting the prosody of the target syllables in a better manner using some machine learning techniques. Incorporating the customized cost parameters into the TTS systems would help to increase the quality of synthesis. Better prosody modification techniques can be explored and may be useful in improving the synthesized speech.

The proposed prosody modification rule set can be extended to make it more comprehensive. Integrating those generic rules in the speech synthesis process will help increase the automation and quality of synthesis.

References

1. Murray, I.R., Arnott, J.L.: Implementation and testing of a system for producing emotion-by-rule in synthetic speech. Speech Communication 16(4), 369–390 (1995)
2. Narendra, N.P., Rao, K.S., Ghosh, K., Vempada, R.R., Maity, S.: Development of syllable-based text to speech synthesis system in Bengali. International Journal of Speech Technology 14(3), 167–181 (2011)
3. Iida, A., Campbell, N., Higuchi, F., Yasumura, M.: A corpus-based speech synthesis system with emotion. Speech Communication 40(12), 161–187 (2003)
4. Clark, A.J.R., Richmond, K., King, S.: Festival 2 - Build your own general purpose unit selection speech synthesiser. In: Proceedings of 5th ISCA Workshop on Speech Synthesis (2004)
5. Black, A.W., Taylor, P., Caley, R.: The Festival Speech Synthesis System, System documentation, edn. 1.4, for Festival Version 1.4.3 (2002)
6. Black, A.W., Lenzo, K.A.: Building Synthetic Voices. Language Technologies Institute, Carnegie Mellon University (2007)
7. King, S., Black, A.W., Taylor, P., Caley, R., Clark, R.: Edinburgh Speech Tools Library, System Documentation, edn. 1.2, for 1.2.3. Centre for Speech Technology, University of Edinburgh (2003)
8. Rabiner, L., Juang, B.H.: An introduction to hidden markov models. IEEE ASSP Magazine 3(1), 4–16 (1986)
9. Narendra, N.P., Rao, K.S.: Syllable specific unit selection cost functions for text-to-speech synthesis. ACM Transactions on Speech and Language Processing 9(3), 5:1–5:24 (2012)
10. Narendra, N.P., Rao, K.S.: Optimal weight tuning method for unit selection cost functions in syllable based text-to-speech synthesis. Applied Soft Computing 13(2), 773–781 (2013)
11. Rao, K.S., Yegnanarayana, B.: Prosody modification using instants of significant excitation. IEEE Transactions on Audio, Speech and Language Processing 14(3) (May 2006)
12. Rao, K.S., Prasanna, S.R.M., Yegnanarayana, B.: Determination of instants of significant excitation in speech using Hilbert envelope and group delay function. IEEE Signal Processing Letters 14(10) (October 2007)