Unit Selection Using Linguistic, Prosodic and Spectral Distance for Developing Text-to-Speech System in Hindi

K. Sreenivasa Rao, Sudhamay Maity, Amol Taru, and Shashidhar G. Koolagudi

School of Information Technology
Indian Institute of Technology Kharagpur
Kharagpur - 721302, West Bengal, India
ksrao@iitkgp.ac.in, friendsudha@gmail.com, amol.taru@gmail.com, koolagudi@yahoo.com

Abstract. In this paper we propose a new method for unit selection in developing text-to-speech (TTS) system for Hindi. In the proposed method, syllables are used as basic units for concatenation. Linguistic, positional and contextual features derived from the input text are used at the first level in the unit selection process. The unit selection process is further refined by incorporating the prosodic and spectral characteristics at the utterance and syllable levels. The speech corpora considered for this task is the broadcast Hindi news read by a male speaker. Synthesized speech from the developed TTS system using multi-level unit selection criterion is evaluated using listening tests. From the evaluation results, it is observed that the synthesized speech quality has improved by refining the unit selection process using spectral and prosodic features.

Keywords: Text-to-speech, unit selection, linguistic features, prosodic features and spectral features.

1 Introduction

In the concatenative speech synthesis approach, the speech is generated by concatenating the segments of natural speech waveforms corresponding to the sequence of sound units that are derived from the input text [1]. Earlier, the concatenative speech synthesis is performed by concatenating the sequence of sound units (phones or diphones or syllables), where the unique versions of the sound units are stored in the database. After the concatenation of basic sound units, the prosodic information will be incorporated using appropriate signal processing techniques. This method introduces distortion due to the manipulation of sound units by signal processing techniques. To overcome this distortion, corpus based (data driven) concatenation approach is proposed. In this approach, the database consists of huge labeled speech corpus, having the multiple replicas of the basic sound units. Since the database has multiple candidates for each sound unit, there should be a mechanism to choose the sequence of sound units in an accurate way, such that it requires minimal signal manipulation.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 531–536, 2009.

In this paper we are proposing multilevel unit selection criterion for choosing the sequence of units from the speech corpus. At the first level the unit selection is performed using the linguistic, positional and contextual features derived from the text. In the second level the unit selection is performed on the units selected in the first level using spectral and prosodic features separately. At the final level (3rd level), the unit selection is performed on the units selected at the second level, by combining spectral and prosodic features together.

There are some earlier attempts in the research of developing TTS systems for Indian languages. At IIT Madras, TTS system for Hindi was developed in early 90's using parametric approach [2]. N. Sridhar Krishna *et al.*, have proposed duration and prosodic phrasing models for developing the TTS system in Telugu [3]. Samuel Thomas *et al.*, have developed natural sounding TTS system in Tamil using syllable like units [4]. S. P. Kishore *et al.*, have proposed data-driven speech synthesis approach using syllables as basic sound units for developing TTS in Hindi [5]. At TIFR Mumbai, TTS system for Indian accent English was developed for browsing the web [6]. Sreekanth *et al.*, have developed festival based TTS system for Tamil [7]. Speech synthesizers in Hindi and Bengali were developed at IIT Kharagpur for visually challenged people [8].

The paper is organized as follows: The details of the development of speech corpora are discussed in section 2. In section 3, description of the base line TTS system using the proposed unit selection approach is provided. The proposed approach of unit selection process and performance of the TTS system by incorporating the proposed unit selection criterion is analyzed using listening tests are discussed in section 4. In the final section the summary of the paper is given along with future work that can improve the performance of the system further.

2 Speech Corpus

Speech corpus used in this work consists of Hindi broadcast news data read by male speaker. Duration of speech corpus is about 1 hour. The speech signal was sampled at 16 kHz, and each sample is represented as a 16 bit number. database is organized at sentence, word and syllable levels. Each of the syllables is labeled by 24 features representing the linguistic context and production constraints [9]. These features represent positional, contextual and phonological information of the syllable. The list of the features representing the syllable is given in the Table 1. Along with the linguistic features derived from the text, the syllables are also labeled with the prosodic (i.e., pitch, duration and energy) and spectral information.

3 Text-to-Speech System

Block diagram of the TTS system using the proposed unit selection approach is shown in Fig. 1. Text normalization module converts abbreviations, numbers etc., into the spoken equivalent text. Syllabification module derives the sequence of syllables from the normalized text. Syllabification module first identifies vowels, and

Table 1. List of the factors and features representing the linguistic context and production constraints of the syllable

Factors	Features
Syllable position in the phrase	Position of syllable from beginning of the phrase Position of syllable from end of the phrase Number of syllables in the phrase
Syllable position in the word	Position of syllable from beginning of the word Position of syllable from end of the word Number of syllables in the word
Word position in the phrase	Position of word from beginning of the phrase Position of word from end of the phrase Number of words in the phrase
Syllable identity	Segments of the syllable (consonants and vowels)
Context of the syllable	Identity of the previous syllable Identity of the following syllable
Syllable nucleus	Position of the nucleus Number of segments before the nucleus Number of segments after the nucleus

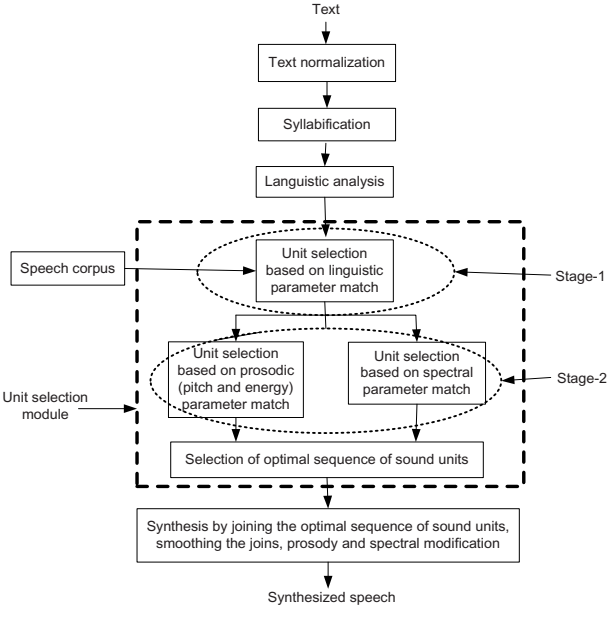

Fig. 1. Block diagram of the TTS system using the proposed unit selection approach

then determines number of syllables present in text. Each syllable is derived by associating consonants to the appropriate vowel using linguistic rules. Linguistic analysis module derives 24 features (see Table 1) for each syllable. The first stage of the unit selection module derives the multiple realizations of the given sound units, which satisfy the matching criterion based on 24 features, derived from linguistic analysis module. These sound units are further fed to spectral and prosodic parameter analysis modules. The second stage of the unit selection module (i.e. spectral and prosodic analysis module) derives the unique sequence of sound units from its

multiple realizations by minimizing the spectral and prosodic distances, between the adjacent units, to minimize the join cost. The final output sequence of sound units given by unit selection module is concatenated by taking care of smoothening the joints between successive units. After simple concatenation, prosodic and spectral manipulations are carried out according to the predicted prosody from the models. Finally, the speech synthesized after prosodic manipulation is the desired speech for the given input text.

4 Proposed Unit Selection Criterion

The basic goal of unit selection module is to derive the optimal sequence of sound units from the speech corpus by minimizing the cost function. Here the cost function may be derived from the three components (1) Linguistic match, (2) Spectral match and (3) Prosodic match. In this work we have proposed two stage unit selection criterion. At the first stage, for the derived sequence of sound units, multiple realizations are chosen using linguistic match criterion. The linguistic match is carried out by matching the 24 dimensional feature vector of each syllable derived from the text to the syllables present in the corpus [9]. At this stage we have considered five syllables for each target syllable. These five units correspond to the top five matched units with respect to the target unit. Unit selection based on the linguistic match is illustrated with an example sentence *"bhaarat ke pradhaanmanthri ne kahaa"*. The text analysis module derives the desired sequence of syllables for the above text as: *bhaa, rat, ke, pra, dhaan, man, thri, ne, ka, haa*. In this utterance there are 10 syllables. The identity of the first syllable is *"bhaa"*, the context of this syllable is represented by the syllable identities of the preceding and the following syllables. In this case the preceding syllable is absent, and the following syllable is *"rat"*. In view of positional information, the preset syllable is at the beginning of the word, beginning of the utterance and the word position is one. Here, for the target syllable *"bhaa"*, all the units of *"bhaa"* in the corpus are chosen. Then based on syllable context some realizations of *"bhaa"* are filtered out. Likewise, the selection process follows the sequence of filtering the units. The filtering process will be terminated when the number of realizations of the unit reaches to 5.

In the proposed unit selection approach spectral and prosodic matches are performed on the realizations of the target units derived from the first stage of unit selection (i.e., linguistic match). Spectral matching between adjacent units is very crucial in view of minimizing perceptual distortion. Therefore while searching units, appropriate weightage has to be given to the spectral distances between the adjacent units. From the first stage of unit selection module, we get roughly 5 realizations of each unit. For performing spectral match between adjacent units, we need to compute spectral distances between the five realizations of the present unit and the five realizations of the following unit (i.e., total of 15 spectral distances). Among the five realizations of present unit and the five realizations of following units, a pair of units is selected based on the minimal spectral distance measures.

In this work prosodic match is estimated using the differences in average pitch and energy between the adjacent units. For deriving the unique sequence of sound units based on prosodic match, prosodic distance is estimated between the five realizations of the present and the following units. The optimal pair is selected based on the minimum distance criterion. For implementing the combined spectral and prosodic matching to select the sequence of units, distances between all possible pairs of units derived from the first stage of unit selection, need to be computed. The optimal sequence is derived by minimizing combined distance derived from prosodic and spectral matches together.

Performance of proposed multilevel unit selection process is evaluated by conducting listening tests on synthesized speech samples. Speech samples are synthesized using concatenative speech synthesizer, by implementing the proposed unit selection methodology. Listening tests are conducted using 25 research scholars in the age group of 25-35 years. Four sets of sentences are synthesized using the proposed unit selection process at different levels: (1) The first set of sentences are synthesized by concatenating the sound units derived from unit selection module without implementing spectral and prosodic matching (i.e., unit selection based on the linguistic match only). (2) The second set of sentences are synthesized by sequence of sound units from unit selection module using linguistic match and spectral match (i.e., without prosodic match). (3) The third set of sentences are synthesized by deriving sound units from unit selection module using linguistic match and prosodic match (i.e., without spectral match). (4) The fourth set of sentences are synthesized by deriving sound units using linguistic, spectral and prosodic matches together.

The tests are conducted in the laboratory environment by playing the speech signals through headphones. In the test, subjects were asked to judge the perceptual distortion and quality of the speech on a 5-point scale for each of the sentences. Each listener has to give the opinion scores for each of the five utterances in all four cases (altogether 20 scores) mentioned above. Mean opinion scores (MOS) indicating the quality of synthesized speech are given in Table 2. Different approaches for selecting units are given in the first column of Table 2. The second column in Table 2 shows the MOS for speech quality. The obtained MOS's indicate that the synthesized speech quality has improved by using the proposed multilevel unit selection criterion for choosing optimal sequence of sound units from speech corpus.

Table 2. Mean opinion scores for the quality of synthesized speech for different unit selection approaches

Unit selection approach	Mean opinion score (MOS)
Linguistic match	2.1
Linguistic match + Spectral match	2.5
Linguistic match + Prosodic match	2.6
Linguistic match + Spectral match + Prosodic match	2.8

5 Summary and Conclusion

Multilevel unit selection methodology was proposed to develop the TTS system to enhance the quality of synthesized speech. Hindi broadcast news speech corpus was used for developing the baseline Hindi TTS system. Linguistic, spectral and prosodic features were explored in the unit selection module to choose the optimal sequence of sound units from their multiple realizations. The efficiency of proposed unit selection approach was evaluated by developing Hindi TTS system and carrying out perceptual analysis on synthesized speech. The perceptual analysis showed that quality of synthesized speech was improved by performing the unit selection process using linguistic, spectral and prosodic features in a combined way. The performance of proposed TTS system can be enhanced by manipulating spectral and prosodic features of the concatenated units using the predicted prosody and spectral models.

References

1. Hunt, A.J., Black, A.W.: Unit selection in a concatenative speech synthesis system using a large speech database. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Atlanta, Georgia, USA, May. 1996, vol. 1, pp. 373–376 (1996)
2. Yegnanarayana, B., Murthy, H.A., Sundar, R., Ramachandran, V.R., Kumar, A.S.M., Alwar, N., Rajendran, S.: Development of text-to-speech system for Indian languages. In: Proc. Int. Conf. Knowledge Based Computer Systems, Pune, India, December 1990, pp. 467–476 (1990)
3. Krishna, N.S., Murthy, H.A.: A new prosodic phrasing model for Indian language Telugu. In: INTERSPEECH 2004 - ICSLP, October 2004, vol. 1, pp. 793–796 (2004)
4. Thomas, S., Rao, M.N., Murthy, H.A., Ramalingam, C.S.: Natural sounding TTs based on syllable-like units. In: Proc. 14th European Signal Processing Conference, Florence, Italy (September 2006)
5. Kishore, S.P., Kumar, R., Sangal, R.: A data-driven synthesis approach for indian languages using syllable as basic unit. In: Int. Conf. Natural Language Processing, Mumbai, India (December 2002)
6. Sen, A., Vijaya, K.S.: Indian accent text to speech system for web browsing, Sadhana (2002)
7. Sreekanth, M., Ramakrishnan, A.G.: Festival based maiden TTS system for Tamil language. In: Proc. 3rd Language and Technology Conf., Poznan, Poland, October 2007, pp. 187–191 (2007)
8. Basu, A., Sen, D., Sen, S., Chakrabarthy, S.: An Indian language speech syn- thesizer: Techniques and its applications. In: National Systems Conference, IIT Kharagpur, Kharagpur, India (2003)
9. Rao, K.S., Yegnanarayana, B.: Modeling durations of syllables using neural networks. Computer Speech and Language 21, 282–295 (2007)