IOSR Journal of VLSI and Signal Processing (IOSR-JVSP)
Volume 5, Issue 6, Ver. II (Nov -Dec. 2015), PP 31-35
e-ISSN: 2319 – 4200, p-ISSN No. : 2319 – 4197
www.iosrjournals.org

Approach of Syllable Based Unit Selection Text- To-Speech Synthesis System for Marathi Using Three Level Fall Back Technique

Sangramsing N. Kayte[1], Monica Mundada[1], Dr. Charansing N. Kayte[2], Dr.Bharti Gawali*

*1,*Department of Computer Science and Information Technology Dr. Babasaheb Ambedkar Marathwada University, Aurangabad*
2Department of Digital and Cyber Forensic, Aurangabad, Maharashtra

Abstract: *A text-to-speech synthesis system is one that is capable of producing intelligible and natural speech corresponding to any given text. A popular approach to speech synthesis is unit selection synthesis (USS). The current work focuses on developing a USS system for Marathi. Literature suggests that syllable is a suitable unit for Indian languages. Creating a database that covers all the syllables of Marathi is tedious and expensive, and the footprint size of the system would be in the order of GBs. Therefore, to reach a compromise between the quality and the footprint size, the current work proposes to use a database containing all the phonemes, consonant-vowel (CV) units, and the most frequently occurring syllables of Marathi. This way, given a text, it is first decomposed into syllables. If a particular syllable is not available in the database, it is broken down to CV units and phonemes. The appropriate speech units are then chosen from the database and concatenated to produce a speech utterance. The performance of the system will be evaluated subjectively by the mean opinion score.*

Keywords*: Speech Synthesis, Unit Selection, Text-to-Speech, Mean Opinion Score*

I. Introduction

Speech Synthesis is the artificial generation of speech signal from text. Speech synthesis system has mainly two parts;first part converts speech to linguistic specification (grapheme-to-phoneme conversion). The second part generates the speech waveform. An unrestricted text-to-speech system is expected to produce a speech signal, which is corresponding to the given text[1]. A popular approach to speech synthesis are Unit Selection Synthesis (USS), Hidden Markov model-based speech synthesis[2][3]. A successful concatenative speech synthesis technique is the unit selection synthesis. Unit selection speech synthesis is used to synthesize speech for the given input text. Festival framework is requiredfor synthesizing the voice for unit selection speech synthesis approach. It involves the concatenation of appropriate pre-recorded speech units, for the given text, based on the target and concatenation costs. The target cost identifies the units in the database that best match the required specification and the concatenation cost identifies the units that join smoothly. The speech units can be words or sub-word units such as phonemes, di-phones, syllables, etc. The quality of speech synthesized varies based on the size of the unit. If the units are longer, naturalness is better-preserved and the number of concatenation point are less. However, the amount of data required to train the synthesis system increases, by increasing the unit-size, thereby increasing the footprint size of the system. Earlier phoneme based system and CV based systems are developed, and the result shows that phoneme system performs better since the concatenation points are less. For smaller units, phoneme is selected as the best unit, but there are more sonic glitches in the phoneme based system. For building a system using syllable as a unit, more amount of training speech data are required, creating a huge database is tedious and not a time consuming process and also more memory space is required for building such a system. If less amount of data are used then some syllables [2][4]may not be present in the database. Hence there are some issues in maintaining the good quality of the synthesized Speech and naturalness has to be considered to a greater extent for a better synthesized system. In the current work, the unit selection speech synthesis systems for Marathi are developed with less amount of speech data with syllable as the major unit. The given text is first decomposed to syllables. If the particular syllable in the given input is not present in the internal database, then the syllable unit is broken down into CV units and phonemes. If the particular CV is not present in the database, then it picks the phoneme unit from the database for developing the synthesis system. Finally the appropriate speech units are chosen from the entire database and the units are concatenated which are used to produce a speech utterance. The given text is synthesized by combining the pre-recorded units from the database. The performance of the synthesized voice is analyzed by obtaining the mean opinion score. The creation of the database is tedious process, so in this

approach with less amount of training data, the system generates a speech which is highly intelligible and natural [5] and quality is also maintained. The paper is organized as given below: Section II describes the speech corpus for building the synthesis system. Section III describes the Unit Selection systems developed for Marathi in detail. Section IV describes the performance analysis for the developed systems. Section V describes the conclusion of the given work [12-20]..

II. Speech Corpus

The speech corpus consists of one hour of recorded Marathi speech data for training the system. The data are recorded from a native Marathi speaker. The recorded speech has the sampling rate of 16 kHz. Recording was done in a noise free environment at laboratory using a unidirectional carbon microphone. An M-Audio Mixer was used to suppress the noise and was set in mono mode to capture only the speaker's voice and the sampling rate was set to be 160000 KHz. For recording, Audacity software was used. Precautions were taken to maintain a constant energy in the recorded speech. Festival supports different forms of audio files such as ulaw, snd, aiff (audio interchange file format) and riff (resource interchange file format chunks) format [6].

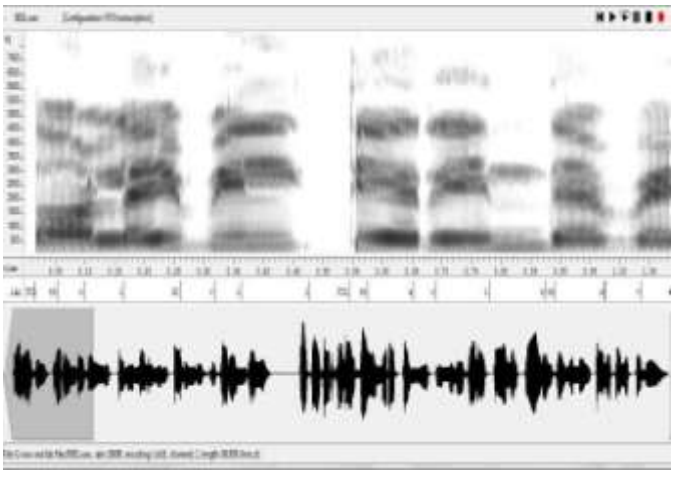
Segmentation

Speech segmentation is the process of identifying the boundaries between words, syllables, or phonemes in spoken natural languages. The lowest level of speech segmentation is the breakup and classification of the sound signal into a string of phones. The lab files are required for the corresponding wave files. The lab files are generated by segmenting the data. The data are segmented by using forced-viterbi algorithm. Initially five minutes of data are manually segmented, which contains of 30 sentences. The manual segmentation is done at phoneme level for the representations waveforms and the spectrograms, HTK transcriptions, TIMIT transcriptions, etc. Models are generated for all the phonemes and the lab files are generated. Forced-viterbi alignment are performed to segment the rest of the data. The following steps are used to segment the data:

1) By using 5 minutes of data and the corresponding time aligned phonetic transcriptions, context-independent phoneme models are trained.
2) Using these models and the phonetic transcriptions, the speech data are segmented using forced-Viterbi alignment procedure.
3) Using the obtained phonetic transcription (phone-level label files), new context-independent phoneme models are trained. 4) Steps 2 and 3 are repeated for times.

5) After iterations, the resultant HMMs are used to segment the entire speech data, again. These boundaries are considered as final boundaries. Finally the lab files are obtained by itearatively performing forced-viterbi alignment for the speech data.

Figure 1 Segmented wave file



From the phoneme level lab file obtained, the CV and syllable lab files are obtained by using the script which gives the phonemes as CV units and syllable units. Finally phoneme, CV and syllable lab files has been created and the segmentation is checked manually.

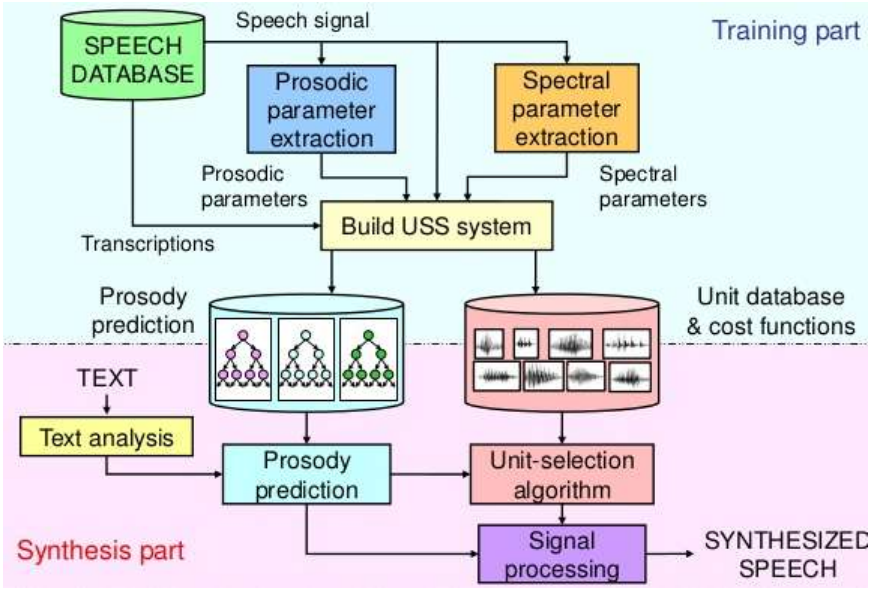
III. Unit Selection Speech Synthesis

In the current work, the phoneme based system, CV based system, syllable based system and syllable with fall back systems were developed using Unit Selection Speech Synthesis to compare the performance of all

the systems. Concatenative synthesis generates speech by connecting natural, prerecorded speech units. These units can be words, syllables, half-syllables, phonemes, diphones or trip hones [7]. The unit length affects the quality of the synthesized speech. With longer units, the naturalness increases, less concatenation points are needed, but more memory is needed and the number of units stored in the database becomes very numerous. With shorter units, less memory is needed, but the sample collecting and labeling techniques become more complex.

The unit selection is based on two cost functions [12-20].

- – Target cost, Ct (ui, ti) is an estimate of the difference between a database unit, ui and the target, ti which it is supposed to represent.
- Concatenation cost, Cc (ui-1, ui) , is an estimate of the quality of a join between consecutive units, ui-1 and ui
- The architecture of unit selection system is shown in the below figure:



The following systems are developed using unit selection speech synthesis:

Phoneme based System: The phoneme based system is developed using the phoneme level lab files. In the word "SangraM". It is transcribed as /s/ /a/ /n/ /g//r/ /a/ /M/. The utterances are generated for each of the labels and the corresponding text. Phoneme is the smaller unit in the speech corpus, hence the concatenation points are more in the phoneme based system. Finally the system is built with one hour of speech data. The phone-set features have been defined for all the phonemes in the Marathi language. Using the festival framework, the system is developed and the performance of the system is also analyzed.

CV-Based System: The CV-based system was developed using CV level lab files. The label files for the CV-based system are obtained from the phoneme-level label files by combining two successive phonemes if a vowel follows a consonant. The CV-based system contains vowels (V), consonants(C), and consonant-vowel (CV) units. For example, the word "sangraM" is split as /s/ /a/ /n/ /g//r/ /a/ /M/.

Syllable based system:

The syllable based unit selection speech synthesis systems [8] are developed. More amount of training data are required to develop a synthesis system with syllable as a unit. In this system, there are less number of examples so some of the syllable units are not synthesized. The syllable level lab files are generated from the CV lab files, whereas the CV lab files are obtained from the phoneme lab files by concatenating the consonants followed by a vowel. Finally these lab files with corresponding text and wave files are required for building the synthesis system.

Syllable with three level fall back:

The syllable based system with fallback is similar to syllable based system but in addition the CV and phoneme lab files are also considered. The systems are built using all the sub word units such as phoneme, Consonant vowel (CV) and syllable units. The syllables which occur frequently and contains more number of examples are sorted out, and for these syllables alone the syllable units are picked. In case if particular syllable is not present in the database the syllables are decomposed into CV units. If the particular CV is not present in

the database finally it is fall back to phoneme units and the corresponding word has been synthesized and the speech was generated. In this system, if the syllable units are not present in the database, then also the text is synthesized by fall back to CV units and phonemes.Hence the syllable with fallback requires only less amount of data but can synthesize all the syllable units and generate the corresponding speech voice. Therefore we introduce a system which contains all the units and the text is synthesized. Therefore we infer that this systems outperforms the phoneme, CV and syllable based systems [9]. The steps for building voice in festival framework are as follows:

- The features for each of the units has to be mentioned and is used for the pronunciation of the particular unit, the features are vowel/ consonant/ nasal/ fricative etc.
- The utterance are created for all the sentences used in the entire database.
- Letter-to-sound rules are used to break a sentence into required sub-word units, to generate the initial utterances.
- Extracting the pitch marks and building LPC coefficients.
- Post processing is done to tune the pitch marks. Pitch marking plays a vital role in the extraction of Mel cepstral coefficients, because synchronous framing was used for Festival.[9][10]
- The units in the database are clustered using Classification and Regression Trees (CART) [6]. The number of questions has been defined to classify the units into clusters. If the number of questions are greater than the number of units in a cluster are less and tree becomes deeper.
- Testing of the voice for out domain sentences.

To synthesize a new sentence or paragraph, the text is given as the input, the system splits the text into the required sub-word units and identifies the most suitable unit to be concatenated based on the target cost and the concatenation cost. The utterance are created for the corresponding text with the units selected and the final speech is synthesized from the utterance.

IV. Performance Analysis

MOS is calculated for subjective quality measurement. It is calculated for the synthesized speech using the Unit selection synthesis. It was counseled to the listeners that they have to score between 01 to 05 (Excellent – 05 Very good – 04 Good – 03 Satisfactory – 02 Not understandable-01) for understandable. The MOS scores were collected from 10 native listeners. 30 wave files were synthesized and the Mean Opinion Score was obtained for each of the wave files[6] [11].

Table 1: MOS Obtained for all the Systems

Syllable with Fall back System	Phoneme based System	CV based System	Syllable based System
4	4.5	4.7	3.5

The testing is done by synthesizing sentences from out domain, which is not present in the training data. The corresponding wave files are synthesized form the festival system for the sentences or paragraphs. From the MOS score obtained, the syllable with fall back system has the highest score and intelligibility.

V. Conclusion

Thus from the systems developed and the observations made, we conclude that the syllable with fall back system outperforms the other systems. The synthesized speech is highly intelligible and the quality is improved to a greater extent. Although the syllable based system is intelligible, more amount of training data are required to synthesize all the out domain sentences. The voices are tested and analyzed and found from the analysis made, it is observed that many of the syllables are missing in syllable based system due to less amount of training data in the corpus, if data increased these can be neglected, which results in high memory and more time for creating a large speech database. So we conclude that the systems developed using fall back can be considered as the best synthesis technique to building speech voices.

References

[1]. Sangramsing Kayte, Dr. Bharti Gawali "Marathi Speech Synthesis: A review" International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 3 Issue: 6 3708 – 3711

[2]. Sangramsing Kayte, Monica Mundada, Dr. CharansingKayte "A Review of Unit Selection Speech Synthesis International Journal of Advanced Research in Computer Science and Software Engineering -Volume 5, Issue 10, October-2015

[3]. Sangramsing Kayte, Monica Mundada, Dr. CharansingKayte "A Marathi Hidden-Markov Model Based Speech Synthesis System" IOSR Journal of VLSI and Signal Processing (IOSR-JVSP) Volume 5, Issue 6, Ver. I (Nov -Dec. 2015), PP 34-39e-ISSN: 2319 – 4200, p-ISSN No. : 2319 –4197

[4]. Sangramsing Kayte, Monica Mundada, Dr. CharansingKayte "Di-phone-Based Concatenative Speech Synthesis Systems for Marathi Language" OSR Journal of VLSI and Signal Processing (IOSR-JVSP) Volume 5, Issue 5, Ver. I (Sep –Oct. 2015), PP 76-81e-ISSN: 2319 –4200, p-ISSN No. : 2319 –4197
[5]. Young S., G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. Woodland, "HTK Book", ver 3.2.1 Cambridge University Engineering department, 2002.
[6]. Sangramsing Kayte, Monica Mundada, Santosh Gaikwad, Bharti Gawali "PERFORMANCE EVALUATION OF SPEECH SYNTHESIS TECHNIQUES FOR ENGLISH LANGUAGE " International Congress on Information and Communication Technology 9-10 October, 2015
[7]. Sangramsing Kayte, Monica Mundada, Dr. CharansingKayte "Implementation of Marathi Language Speech Databases for Large Dictionary" IOSR Journal of VLSI and Signal Processing (IOSR-JVSP) Volume 5, Issue 6, Ver. I (Nov -Dec. 2015), PP 40-45e-ISSN: 2319 –4200, p-ISSN No. : 2319 –4197
[8]. Sangramsing Kayte, Monica Mundada, Dr. CharansingKayte "A Corpus-Based Concatenative Speech Synthesis System for Marathi" IOSR Journal of VLSI and Signal Processing (IOSR-JVSP) Volume 5, Issue 6, Ver. I (Nov -Dec. 2015), PP 20-26e-ISSN: 2319 –4200, p-ISSN No. : 2319 –4197
[9]. Sangramsing Kayte, Monica Mundada "Study of Marathi Phones for Synthesis of Marathi Speech from Text" International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 (Volume-4, Issue-10) October 2015
[10]. Sangramsing Kayte, Monica Mundada, Dr. CharansingKayte "Di-phone-Based Concatenative Speech Synthesis System for Hindi" International Journal of Advanced Research in Computer Science and Software Engineering -Volume 5, Issue 10, October-2015
[11]. Sangramsing Kayte, Monica Mundada, Dr. CharansingKayte " Performance Calculation of Speech Synthesis Methods for Hindi language IOSR Journal of VLSI and Signal Processing (IOSR-JVSP) Volume 5, Issue 6, Ver. I (Nov -Dec. 2015), PP 13-19e-ISSN: 2319 –4200, p-ISSN No. : 2319 –4197
[12]. Sangramsing Kayte, Monica Mundada,Dr. Charansing Kayte" Speech Synthesis System for Marathi Accent using FESTVOX" International Journal of Computer Applications (0975 – 8887) Volume 130 – No.6, November2015
[13]. Sangramsing Kayte, Monica Mundada,Dr. Charansing Kayte "Screen Readers for Linux and Windows – Concatenation Methods and Unit Selection based Marathi Text to Speech System" International Journal of Computer Applications (0975 – 8887) Volume 130 – No.14, November 2015
[14]. Sangramsing Kayte, Monica Mundada,Dr. Charansing Kayte " Performance Evaluation of Speech Synthesis Techniques for Marathi Language " International Journal of Computer Applications (0975 – 8887) Volume 130 – No.3, November 2015
[15]. Sangramsing Kayte, Monica Mundada, Jayesh Gujrathi, " Hidden Markov Model based Speech Synthesis: A Review" International Journal of Computer Applications (0975 – 8887) Volume 130 – No.3, November 2015
[16]. Sangramsing N. Kayte ,Monica Mundada,Dr. Charansing N. Kayte, Dr.Bharti Gawali "Approach To Build A Marathi Text-To-Speech System Using Concatenative Synthesis Method With The Syllable" Sangramsing Kayte et al.Int. Journal of Engineering Research and Applications ISSN: 2248-9622, Vol. 5, Issue 11, (Part-4) November 2015, pp.93-97
[17]. Sangramsing N. Kayte, Dr. Charansing N. Kayte,Dr.Bharti Gawali* "Grapheme-To-Phoneme Tools for the Marathi Speech Synthesis" Sangramsing Kayte et al.Int. Journal of Engineering Research and Applications ISSN: 2248-9622, Vol. 5, Issue 11, (Part -4) November 2015, pp.86-92
[18]. Sangramsing Kayte "Duration for Classification and Regression Tree for Marathi Text-to-Speech Synthesis System" Sangramsing Kayte Int. Journal of Engineering Research and Applications ISSN: 2248-9622, Vol. 5, Issue 11, (Part-4)November2015
[19]. Sangramsing Kayte "Transformation of feelings using pitch parameter for Marathi speech" Sangramsing Kayte Int. Journal of Engineering Research and Applications ISSN: 2248-9622, Vol. 5, Issue 11, (Part -4) November 2015, pp.120-124
[20]. Sangramsing N. Kayte, Monica Mundada, Dr. Charansing N. Kayte, Dr.BhartiGawali "Automatic Generation of Compound Word Lexicon for Marathi Speech Synthesis" IOSR Journal of VLSI and Signal Processing (IOSR-JVSP)Volume 5, Issue 6, Ver. II (Nov -Dec. 2015), PP 25-30e-ISSN: 2319 – 4200, p-ISSN No. : 2319 – 4197