Text-To-Speech Synthesizer for English, Hindi and Marathi Spoken Signals

G. D. Ramteke and R. J. Ramteke

1School of Computer Sciences, North Maharashtra University, Jalgaon (MS), India.

Authors’ contributions

This work was carried out in collaboration between both authors. Both authors read and approved the final manuscript.

Article Information

DOI: 10.9734/BJAST/2016/24869

Editor(s):
(1) Samir Kumar Bandyopadhyay, Department of Computer Science and Engineering, University of Calcutta, India.

Reviewers:
(1) S. K. Srivatsa, CSE, Prithyusha Engineering College, Chennai, India.
(2) Utku Kose, Usak University, Turkey.

Complete Peer review History: http://sciencedomain.org/review-history/13898

Received 5th February 2016
Accepted 10th March 2016
Published 28th March 2016

Original Research Article

ABSTRACT

The paper proposes a model of Text-To-Speech (TTS) engine for Marathi, Hindi and English languages. The characters and their representation are analyzed and synthesized with the help of TTS-engine. The engine would be spoken utterances produced from text. A concatenative approach based on linguistic rules has been applied. In order to test the artificial voice generation, an analysis of prosody and MOS tests were performed. Cepstral pitch detection algorithm has been used for extraction of fundamental frequency. Mean and standard deviation were computed on pitch reading of each spoken signals. MOS is a subjective test depends on listeners who are aware of any three of the languages. All scores are provided on the basis of two dissimilar parameters of MOS: 1. The result was achieved between fair and good for listening quality. 2. Naturalness score in between pleasant and slightly-pleasant. Ultimately, the satisfactory results were found for three different languages.

Keywords: Speech synthesizer; Phones; Dictionary; Analysis of Prosody; Mean Opinion Score.

1. INTRODUCTION

A lot of primary students have major difficulties in learning and correct hearing. Many educational organizations use state-of-art facility with several technologies. For any organization, the technology helps to improve two aspects of education: A writing style and other spoken form.
The writing style belongs to the character shape and spoken form is based on voice signals [1,2]. Voice and linguistic forms are being applied on primary education domain. The linguistics is the scientific study of languages, from the sounds, characters, words and letter-to-sound rules. Voice output is a pivotal section of such correct speech generation. A couple of these aspects include an application of text and voice interface. Text-to-speech synthesizer is one of the models of text and voice interface. With the help of voice communication, the effective text-to-speech model is tried to generate synthetic voice as close as possible to mankind of voice. Clarity of voice generation plays a vital role to convey the message positively [3-5].

For many years, text-to-speech synthesizer has been considering a rich domain than other fields of speech processing. TTS-model converts text into voice form in a specific language. Earlier the process of voice generation, it is concerned with prosody. So, supersegmental type of the prosody is used. It involves the different sections: duration of the speech, intonation level and pitch detection. The sections of prosody are fundamental needs of TTS-synthesizer. For implementation, TTS-models have number of techniques. The techniques were borrowed from the speech processing. A lot of the present techniques designed using concatenative speech synthesis. The present work is based on concatenative approach using linguistic rules. It focuses on any type of character and their representation in three unconnected languages. The common students can easily understand the expression of the speaker. Even, primary students are able to improve the basic skills like learning, writing and building their confidence level. The major contribution of TTS-engine is its application for blind students [5-7].

The major objective of TTS-system is to convert text into voice signals. The received text would be any type of character with their representation in three languages. The artificial voice signals with noiseless environment are generated in Marathi, Hindi and English languages.

The paper is organized as follows: The next section discusses on past efforts of text-to-speech (TTS) synthesis system. Third section explains the characteristic of the Marathi, Hindi and English languages. The fourth section proposes speech synthesis engine for Marathi, Hindi and English text. The experimental work and result analysis describe in fifth section. Finally, the present paper gives the conclusion in sixth section.

2. PAST EFFORTS OF TTS-SYSTEM

Numerous researchers have been working on text-to-speech (TTS) synthesis domain since recent decades. TTS-synthesizer is an artificial producible sound engine. TTS-models have been implemented on several languages. Katsunobu Fushikida et al. [1] developed a Japanese text to speech synthesis system for the personal computer. In their system, they have used the formant parameters based on synthesis-by-rule. Richard Sprat [2] introduced a model of multilingual text-to-speech (TTS). The model has been worked on morphological rules, numeral-expansion rules and phonological rules. Speech technologies are widely believed to have a useful contribution. Norbert Pachler [3] presented one of most useful contribution for teaching and learning processes in foreign language. The system was based on the concatenative approach. Gerry Kennedy [4] represented the benefits of speech synthesis software. The software would be beneficial to persons of any age who are visually impaired, various primary students with learning difficulties and mainstream, enjoy hearing what they have written. Bhuvana Narasimhan et al. [5] proposed schwa-deletion in Hindi text-to-speech system based on concatenative synthesis. Diemo Schwarz [6] designed concatenative technique for musical sound synthesis. The concatenative technique provided a large sound corpus. With the advanced computer device, a huge speech corpus has been used for naturalness and intelligibility. In [7] Hwai-Tsu et al. described Mandarin Text-To-Speech (TTS) synthesizer using concatenative technique. It has created 408 syllabic utterances. Usha Goswami [8] showed that synthetic phonics can be preferred the approach for young English learners. P. Zervas [9] presented statistical evaluation of a prosodic database for Greek speech synthesis. The step of Greek speech was segmented into phones. The phones recognizer based on the HTK platform. Naim R. Tyson et al. [10] implemented a new concept that the phenomenon of schwa-deletion in Hindi was handled in the pronunciation component using concatenative text-to-speech system. Pamela Chaudhur et al. [11] developed Telugu Text-To-Speech system using concatenative speech synthesis. It carried out mean opinion score test for intelligibility. Arun Soman et al. [12] proposed a corpus-driven Malayalam text-to-speech (TTS)
system based on concatenative synthesis approach for naturalness and intelligibility. [13] Kwan Min Lee et al. studied that the effect of the student choice on social responses to computer-generation speech was investigated. Matej Rojc et al. [14] proposed a corpus-based text-to-speech synthesis system based on unit selection algorithm for synthesized speech quality. Even, Lakshmi Sahu et al. [15] represented a corpus-driven text-to-speech system for Indian languages: Hindi, Telugu and Kannada. It was used for human spoken voice. Catalin Ungurean et al. [16] designed an improved preprocessing unit for a high quality Romanian TTS-system. It has a major contribution to the clearness and natural sound of a synthesized text. Madiha Jalil et al. [17] surveyed text-to-speech synthesis techniques by highlighting its digital signal processing component. Concatenative synthesis is an easy to implement than rule-based synthesis. There is a reason that no needs to establish speech production rules. Shreekanth T et al. [18] represented concatenative speech synthesis using phoneme, di-phone and allophone. The syllables were a basic unit of Hindi language. Vivek Hanumante et al. [19] aimed at providing conversion of English text into multilingual speech output using an Android application. The application assisted to people lacking the power of speech or non-native speakers. Mukta Gahlawant et al. [20] proposed natural speech synthesizer using hybrid approach. In natural synthesizer, dynamic nature of human speech was very difficult to mimic for speech factors: intelligibility understands an output easily and naturalness describes matching accuracy of output sounds with human voice. Theophile K. Dagba et al. [21] described a unit selection speech synthesis system for Fon language. It was based on phonetic rules, post lexical rules and letter-to-sound rules for automatic phonetic transcription of received text. Sunil S. Nimbhore et al. [22] developed a first level of implementation of natural sounding speech synthesizer for Marathi language using English script. It has used concatenative synthesis approach and evaluated formant frequencies which are one of part of prosodic analysis. Fiona S. Baker [23] exposed the facts of text-to-speech systems. It exposed for 24 innovative-English-speaking to college students who were experiencing reading difficulties in their freshman year of college. Soumya Priyadarshini Panda et al. [24] proposed a module for text-to-speech synthesis in Indian languages: Hindi, Odia and Bengali using the pronunciation rule. The rule was based on concatenation approach for TTS-model using some amount of the memory requirement. Smita P. Kawachale et al. [25] illustrated a new approach and methodology that helped to reduce database size of the syllabic based on concatenative speech synthesis. The model has been worked on human natural speech systematically. We [26] have attempted to design a TTS-synthesis system for Marathi numerals. It was based on rule-based approach. This system would be beneficial to a person who is able to understand the synthesized number. Saleh M. Abu-Soud [27] developed a new multilingual text-to-speech system. It was based on inductive learning. In the system has been composed of three phases: the analysis phase, learning phase and synthesis phase. It was produced the correct phonemes with high accuracy. John O. R. Aoga [28] presented the integration of Yoruba language into MaryTTS. A synthesized system carried out the unit-selection technique. For the system, it was used a speech corpus containing 2415 different sentences. German Bordel et al. [29] dealt with either very long audio tracks or acoustically inaccurate text transcripts. It revealed a probabilistic kernel based on phonetic knowledge using a large corpus of speech. Shinnosuke Takamichi et al. [30] introduced a statistical parametric speech synthesis including text-to-speech and voice conversion. This system has been worked on Hidden Markov Model (HMM), Gaussian Mixture Model (GMM)-based voice conversion for improvement of the synthetic speech quality. Generating natural voice is an uphill task of speech synthesis field.

3. CHARACTERISTIC OF THE MARATHI, HINDI AND ENGLISH LANGUAGES

For human communication, several languages in the entire world are used. In addition, human has different communication ways such as in sound signals, written form or gesture type [4]. In India, there are 22-constitutional languages. Marathi and Hindi languages are enlisted as official languages of India. In the state of Maharashtra in India, Marathi is a local language. Hindi is a national of India [5].

Marathi and Hindi languages are written in Devnagari script. And the writing style of Marathi, Hindi languages is from left to right. The characters of Marathi, Hindi and English languages are categorized into two sections: vowel (Swaras) and consonants (Vyanjanas). Marathi supports 45 characters. Hindi language has 44 alphabets. In Marathi, there are 11
vowels and 34 consonants. Similarly, 11 vowels and 33 consonants can be seen in Hindi language [10,16,18].

Marathi language is used by approximately 90 million people in Maharashtra and Goa states of India. Marathi language is officially used in most of the government and private sector in Maharashtra. In the schools of Maharashtra state, Marathi is studied as primary language [26]. In Marathi and Hindi Languages, there are five categories of vowels: Short, long, conjunct, nasal and visarg as shown in Table 1. A vowel is produced sound by the vocal cords. In short, vowels (V) in a word or syllable which makes a short sound. Long vowels are used a word generates long sound in both languages. Conjunct vowel is produced a combination of short and long vowels. Nasal vowel is generated with low tune where there is air pressure through nose and mouth as well. Visarg vowel is uttered as the voiceless sound after the vowels [18]. Marathi and Hindi languages have 33 and 34 consonants respectively. But ‘ ‘ (la) letter is absent in Hindi. Table 2 mentions that the place of articulation and manner of articulation stop air from moving out of the mouth. It refers a successively forward position of the tongue. All consonants are organized into eight groups of places of articulation: Guttural, Palatal, Cerebral, Dental, Labial, Semivowel, Sibilant and other letters. And Five categories of each group are as follows: (U)-Unaspirated Un-Voiced, (A)-Aspirated Un-Voiced, (U)-Unaspirated Voiced, (A)-Aspirated Voiced and Nasal. The first-two groups of nasal consonants can never be found it which can be written alone. Usually, it is conjuncted with another consonant in their corresponding groups [22-24].

English is the world language for communication amongst people. It uses Roman Script for written form. Nowadays, Marathi and English as well are used in school, institute or other divisions for convenience of students. All English sounds are divided into two broad categories: vowels and consonants. In the production of vowels, air comes out freely through mouth. All 26 alphabets are classified into five vowels and 21 consonants as shown in Tables 3 and 4. In phonetics, a vowel is a sound in spoken language and a consonant is articulated with complete or partial closure of the vocal tract. Consonants have friction when they are spoken mostly using the position of the tongue against the lips, teeth and roof of the mouth. Consonants may be voiced or unvoiced. The voiced consonants are pronounced with the vibration of the vocal cords. Consonants can be classified according to the place of articulation as follows: Labial is a combination of bilabial and labio-dental. Bilabial is expressed the sound by the two lips, e.g. /P, B, M, W/. Labio-dental as similar as labial is produced the voice by the lower lip against the upper teeth, e.g. /F, V/. Dental is articulated by the tip of the tongue against the upper teeth, e.g. /T, D/. Alveolar articulated by the blade of the tongue against the teeth-ridge, e.g. /S, Z, N, L, R/. Palatal articulated by the front of the tongue against the hard palate, e.g. /J/. Velar type is articulated by the back of the tongue against the soft palate, e.g. /K, G, X/. Glottal is produced by an obstruction or narrowing between the vocal cords, e.g. /H/. And other letters are /C, K, Q/. The classification of places of articulation is based on manner of articulation. All alphabets of Marathi, Hindi and English languages are used for the speech analysis and synthesis purpose. [24-27].

Table 1. Clusters of Swaras (Vowels) in Marathi and Hindi Languages

| Five Categories of the Vowels | Short | Long | Conjunct | Nasal | Visarg |
|------------------------------|-------|------|----------|-------|--------|
| अ   | आ    | अ + ई = ए | अ + उ = ऑ | अ   | अ   |
| (a) | (/a/) | (/aa/) | (/e/) | (/am/ or /an/) | (/ah/) |
| ई   | (/i/) | (/ee/) | (/ai/) | - | - |
| उ   | (/u/) | (/oo/) | (/o/) | - | - |
| -   | -     | -     | -       | -     | -     |
Table 2. Set of Vyanjanas (Consonants) for Marathi and Hindi Languages

| Cases          | Alphabets |
|---------------|-----------|
| **Un-voiced** | A E I O U |
| **Voiced**    | A E I O U |
| **Nasal**     | A E I O U |

Table 3. Set of English vowels in Roman script

| Case | Alphabets |
|------|-----------|
| Upper| A E I O U |
| Lower| a e i o u |

Table 4. Chart of English Consonants in Roman script

| Manner of articulation | Places of articulation | 1 | 2 | 3 | 4 | 5 | 6 |
|------------------------|------------------------|---|---|---|---|---|---|
| **Stops**              |                        |   |   |   |   |   |   |
| Guttural               |                        | P | B |   |   |   |   |
| Palatal                |                        | T | D |   |   |   |   |
| Cerebral               |                        |   |   | S | Z |   |   |
| Dental                 |                        |   |   |   |   | J |   |
| Labial                 |                        | K | G |   |   | H |   |
| Glottal                |                        |   |   |   |   |   |   |
| Other letters          |                        | C | K | Q |   |   |   |

4. PROPOSED SPEECH SYNTHESIS SYSTEM FOR MARATHI, HINDI AND ENGLISH TEXT

TTS-synthesizer means a system electronically produces the human sound. Text in three distinct languages would be recognized into a couple of scripts. Marathi and Hindi are identified into Devnagari script. Roman script supports English language. In order to process the generation of audio form, it is analyzed by the prosody as shown in Fig. 1. Concatenative-based synthesis technique is used for concatenation the sound units.
4.1 Procedure of Text and Sound Selection Based on Concatenative Approach

For the procedure of text and sound selection engine, the personal computer has picked up enough processing speech and memory capacity to serve as TTS platforms. Concatenative approach plays a key role to implement the text-to-speech model. The model is used for any type of character with their representation in three unrelated languages based on the using linguistic process in Fig. 2. The procedure of text and sound selection involves two broad processing: linguistic and producible audio form. Linguistic processing consists of syllables. Syllables are being constructed various forms such as C, V, CV, VC, CCV, VVC or so on. C constructs a form of consonant and V is a Vowel. Producible audio form [2-4,25] is an integral part and sub-component of speech synthesis system. Using concatenative technique, producible audio form has performed a role in well manner.

Concatenative approach is divided into tokens or units. This approach provides the extreme intelligibility and to close natural sound. All vowels or consonants with their representation in Roman or Devnagari scripts are stored into the information. The information is classified into two dissimilar sections: Text and voice. The received text can be available in Unicode and phonetic form. But phonetic form is used for implementation process. Text section is classified into the natural and non-natural tags for each language. In order to analyze the natural tag, a lot of characters in three languages are normalized [6,18]. All characters are retrieved from a text dictionary. The set of syllables is matched the text in each language.

The text and sound selection uses context sensitive rewrite rule, which is shown in the equation form as follows:

\[
A \rightarrow B/d/C
\]  
(1)

Where, A is translated d when preceded by B and followed by C. In implemented systems, the context B and C can be any length for character and their representation.

The clean voice is a fundamental need for listening it. But the common voice is recorded with noisy atmosphere. Reducing the noise from original voice signals is an uphill task for speech processing. The process of noise reduction is damaged on common sound signals, but spectral subtraction method doesn’t affect the voice signals. The method is used of the acoustic terms for quality speech. The acoustic terms of voice signals are required such as Frame length – 30 ms; Acoustic Parameter - FFT Spectrum (256 pts.); Frame shift – 10 ms; Sampling frequency – 22 (22050 Hz) KHz; Analysis window – Hamming window [26].

Text unit is matched with a quality sound unit which is retrieved from a speech dictionary. All selected voice units are merged into one speech unit using concatenating process. The complete unit is posted to audio system. Audio system is performing a main task of generating the voice.
The nature of audio system converts digital signals into analog signals. Human beings can understand only analog signals [22,26,30].

4.2 Enriched Prosody

Prosody is a heart of TTS-model. With the help of prosody, speech corpus can be created and extracted prosodic features of speech signals. Speech synthesis requires a speaker. There is no limitation of speaker who is either an unprofessional or a professional. But the speaker should be aware of Marathi, Hindi and English language. So, single unprofessional speaker is sufficient for recording sound signals. The sound information would be in three different languages. All sounds with noisy environment are recorded. In order to enhance the speech dictionary, speech signals have to go the process of analysis of prosody as shown in Fig. 3. Enriched prosody includes three parts: Period of the sound signals, pitch detection of spoken form and intensity of spoken signals. All spoken forms in three untouched languages are stored into speech dictionary. But the section focuses on different issues of superasegmental prosody such as period of sound signals and pitch detection [5,10].

Fig. 3. Process of Enriched Prosody in Three Languages

4.2.1 Period of Sound Signals

The following equation is to find out period of sound signals as follows:

\[ \text{Period} = \frac{1}{F} \]  

(2)

Where, period is total duration of sound signals, 1 is the sampled information of sound, F is an actual sampling frequency of sound signals [26].

4.2.2 Pitch Detection Technique

In speech processing, pitch detection or fundamental frequency (F0) relates to the note of the voice. It is determined by the frequency of vibration of the vocal cords. For pitch detection, cepstral technique is implemented. Cepstral technique has the problems. Potential problems can be avoided by electing frames from the phones if the phones are unexpected signals. [5] The cepstral coefficients are found by the following equation:

\[ c(\tau) = F\{\log(|F\{x[n]\}|^2)\} \]  

(3)

Where F indicates Fourier Transform and \( x[n] \) denotes the signal which would be in the time domain, \(|F\{x[n]\}|^2\) is estimation of the power spectrum of the signal.

Fig. 4 shows that pitch detection using cepstral method is designed. Cepstral technique is applied on various units of speech. The units are generated sound signals. The sound information is any type of characters with their representation in three languages. The complete sound information or a time-domain waveform is called as a window. Segmentation of a window is segmented into frames. When the cepstral pitch technique is detected the pitch, some acoustic terms are necessary for pitch detection. The acoustic terms are such as Sampling frequency – 22 KHz (22050 Hz); Frame length – 30 ms; Frame overlap – 20ms; Analysis window – Hamming window; Acoustic Parameter - FFT Spectrum. A frame is passed through hamming window, which is an inbuilt function of MATLAB. The cepstral domain represents the frequency, which would be in the logarithmic magnitude spectrum of a signal. The cepstrum is formed by taking the FFT of log magnitude spectrum of a signal. The pitch can be estimated by picking the peak value of the resulting signal within a certain range. A peak factor in the cepstrum indicates that the signal is a linear combination of a couple of pitch or fundamental frequency (F0). The pitch period is able to find out many coefficients where the peak occurs. Cepstral method enables to detect the pitch whether a signal contains periodic, aperiodic or silence region. The detection of pitch in a specific range is between 140-340 Hz for a male or a female [8-10].

4.2.3 Recording and Storing of Information

Recording and storing of the information is a part of data acquisition. The data acquisition of vowels or consonants is used. It is divided into two areas. The primary area is a text dictionary and another area is a speech dictionary. As no standardized text corpus for Marathi, Hindi and
English languages is available, the text dictionary has been created for primary education domain. A special document is designed for collecting the information. Manually, the information is collected from the primary books. The overall size of text dictionary for three unconnected languages is 4,910 syllables. Another area, the speech dictionary is acquired through standard PRAAT tool. The sound information is any type of vowels or consonants with their representation in three unrelated languages. The sampling frequency of each sound sample is 20 KHz (22,050Hz). The additional two acoustic things are one mono channel and another file format (wav) for analyzed and synthesized. Speech synthesizer is essential only one male or female speaker for implementation. For the system, single male speaker is used. The speaker should be familiar with Marathi, Hindi and English languages for implementation, All characters with their representation are produced the sound signals by a male speaker with noisy environment, age group is 25 to 35 in 10x12 rooms at School of Computer Sciences in North Maharashtra University, Jalgaon. [1,7,12]

Speech signals are divided into three different categories of the sound: Normal, Noise and Clean signals in Table 5. Speech synthesizer requires simply quality speech. And also, listeners demand the clear sound for listening. For this system, phones material has been classified into three parts: vowels, consonants and their representations or words. The phones were collected. The overall size of phoneme for Marathi, Hindi and English languages are 291, 312 and 165 respectively. A set of phoneme is 768. All phonemes are useful for pronouncing from text. A well-designed speech corpus based on text dictionary is a huge impact on the quality of the synthesized speech [19, 22, 24, 27].

### 4.3 Mean Opinion Score Value

In voice communication, the quality of output voice is subjective. Measuring the quality of voice is a challenging issue. A numerical method of expressing voice quality is called as MOS (Mean Opinion Score). Therefore, MOS value is proposed using two parameters. It is expressed in the range from 0 to 5 as per Table 6.

#### Table 5. Category of Sounds for Three Languages

| Three categories of sound | Language | Types of phones | Number of text | Total phonemes |
|--------------------------|----------|-----------------|----------------|----------------|
| Normal                   | Marathi  | Vowels          | 11             | 291            |
| Noise                    | Marathi  | Consonants      | 34             |                |
| Clean                    | Marathi  | Words           | 52             |                |
| Normal                   | Hindi    | Vowels          | 11             |                |
| Noise                    | Hindi    | Consonants      | 33             | 312            |
| Clean                    | Hindi    | Words           | 60             |                |
| Normal                   | English  | Vowels          | 5              |                |
| Noise                    | English  | Consonants      | 21             | 165            |
| Clean                    | English  | Words           | 29             |                |

**Overall size of phonemes**: 768
Table 6. Scores of Mean Opinion Score (MOS)

| MOS value | Listening quality  | Criteria of pleasant |
|-----------|--------------------|----------------------|
| 5         | Excellent          | Well-pleasant        |
| 4         | Good               | Slightly pleasant    |
| 3         | Fair               | Pleasant             |
| 2         | Poor               | Slightly pleasant    |
| 1         | Bad                | Unpleasant           |
| 0         | Very poor          | Unfit for pleasant   |

Table 7. Listeners and its Opinions for Marathi, Hindi and English Languages

| Language | Type of character | Number of characters with their representation | Listener (Male-M/ Female-F) | Total opinions |
|----------|-------------------|-----------------------------------------------|----------------------------|----------------|
| Marathi  | Vowels            | 11*2+11=33                                    | M – 11                      | 121            |
|          |                   |                                               | F – 08                      | 88             |
|          |                   |                                               | M – 11                      | 363            |
|          |                   |                                               | F – 08                      | 264            |
|          | Consonants        | 33*2+33=99                                    | M – 05                      | 55             |
|          |                   |                                               | F – 06                      | 66             |
|          |                   |                                               | M – 05                      | 170            |
|          |                   |                                               | F – 06                      | 204            |
| Hindi    | Vowels            | 11*2+11=33                                    | M – 05                      | 55             |
|          |                   |                                               | F – 06                      | 66             |
|          |                   |                                               | M – 05                      | 170            |
|          |                   |                                               | F – 06                      | 204            |
|          | Consonants        | 34*2+34=102                                   | M – 07                      | 35             |
|          |                   |                                               | F -06                       | 30             |
|          |                   |                                               | M – 07                      | 147            |
|          |                   |                                               | F - 06                      | 126            |
| English  | Vowels            | 5*2 + 5=15                                    | M – 07                      | 35             |
|          |                   |                                               | F -06                       | 30             |
|          |                   |                                               | M – 07                      | 147            |
|          |                   |                                               | F - 06                      | 126            |
|          | Consonants        | 21*2+21=63                                    | M – 07                      | 35             |
|          |                   |                                               | F - 06                      | 30             |
|          |                   |                                               | M – 07                      | 147            |
|          |                   |                                               | F - 06                      | 126            |

The listening quality of opinion score values are divided into 6 categories. Similarly, the criteria of pleasant are divided into 6 categories. MOS is subjectively measured for quality sounds, when a panel of various kinds of listeners is involved [29]. But the listener should be aware the Marathi, Hindi and English languages. For scoring, MOS values were used. For the experiment, each listener was asked to provide the score under each parameter with 0 (low) to 5 (high). 0 means very poor for listening quality and the score of 5 is excellent rate for quality of listening. Each listener has been given the individual score. But few listeners can understand Hindi and Marathi. A small group of listeners are able to be aware Hindi and English. Other listeners are well-known to three languages as shown in Table 7. There were 43 listeners (23-Male and 20-Female) who can give the MOS score. The overall characters with their representation are 345. And its total opinions for Marathi, Hindi and English sound are 1669. The declared opinions were used to predict the awareness of text. [14,23-26,30]

5. EXPERIMENTAL WORK AND RESULTS ANALYSIS

TTS-model designed to convert character with their representation into sound form. Speech synthesizer has been disposed of two disciplines tests: Analysis of Prosody test and MOS test. First test expresses in auditory terms as pitch detection and duration of speech signals. Second test is MOS (Mean Opinion Score) which depends on subjective analysis.

5.1 Analysis of Prosody Test

In the field of speech synthesis, there are a number of open questions to be solved. One of the most open questions is that how to generate artificial sound. When a system feeds the input text data, the machine can generate the synthesized sound. Due to it, the analysis of prosody test aims at evaluating the prosodic analysis of individual synthesized voice signals. All synthesized sounds have been passed through the present test. The panel of a test is involved two bizarre tasks: Pitch detection and period of spoken form. The period of spoken form assists to the prosodic feature. In pitch detection, the synthesized clean sound in Marathi, Hindi and English languages is extracted the numerical appearance. Pitch detection based on cepstral technique is demonstrated in the shape of its graphical representation as depicted in Fig. 5 and Fig. 6. The Fig. 5 (A) and Fig. 6 (A) are shown in the time domain waveform. The pattern of the
waveforms is plotted time against amplitude. Pitch estimation of voice signals without noise is depicted tracking waveforms of pitch in Fig. 5 (B) and Fig. 6 (B). It is difficult to show all figures. But single English word and a Marathi word can be seen. The results of the test are explained in following Tables 8, 9, 10, 11, 12 and 13. The exactness of pitch frequency is executed by cepstral pitch detection algorithm. All readings of pitch frequency have been exposed in mean and standard deviation.

These tracking values of pitch are relied on a specific range. The range of the pitch detection is measured between 240-300 Hz for mean and 100-120 Hz for standard deviation. Some pitch reading of each speech has to be neglected because various unexpected signals are closest to zero hertz. The computed pitch frequency of three separate languages can be achieved. The evaluation of the test is influenced on the prosodic features.

5.2 Mean Opinion Score Test

Second test of TTS-model is MOS (Mean Opinion Score). MOS is subjective listening test on the basis of input received from listeners. A subjective type question is asked to all listeners based on six rates for listening quality as similar as criteria of pleasant. The process of MOS was performed on any type of characters and their representation in Marathi, Hindi and English languages.

Those listeners follow the instructions on the basis of MOS section to judge the listening quality. Two parameters for MOS test are examined on generating a mankind of sound. Individual feedback was taken from 43 listeners. The MOS test was figured out the average for each language. The results were declared the scores in Tables 14 and 15 for Marathi language.

Table 8. Cepstral Pitch Detection of 5 isolated Marathi Consonants out of 34 Consonants using a Male Voice

| 5 Isolated Marathi Consonants with their representation | Spoken Format in English | Period of Sound Signals in Sec. | Pitch Detection Technique in Hz using Mean as MN and Std. Dev. as SD |
|--------------------------------------------------------|--------------------------|--------------------------------|---------------------------------------------------------------|
| ग म्हणूने गाव      | Ga Mhanje Gavat          | 2.99                           | 249.07, 115.13                                                  |
| झ म्हणूने झाले     | Jha Mhanje Jhabale       | 3.06                           | 258.64, 113.49                                                 |
| न म्हणूने नक     | Na Mhanje Nal            | 2.94                           | 262.84, 114.31                                                 |
| श म्हणूने शहर   | Sha Mhanje Shahar        | 2.86                           | 250.43, 113.73                                                 |
| ट्र म्हणूने त्रिकोण | Tra Mhanje Trikon        | 3.14                           | 251.82, 108.41                                                 |

Fig. 5. English Word “Apple” Uttered by MALE Voice A] Speech Waveform with Noise-Free B] Pitch Tracking

Fig. 6. Marathi Word “पायल” (Pineapple) uttered by MALE Voice A] Speech Signals with Noise-Free B] Pitch Tracking
Table 9. Pitch Reading in mean as MN and standard deviation as SD for Marathi Vowels using a Male Voice

| Marathi Vowel with their representation | Spoken Format in English | Period of Sound Signals in Sec. | Cepstral Pitch Tracking Technique in Hz |
|----------------------------------------|--------------------------|---------------------------------|----------------------------------------|
| आ महाजन अनाझा                        | A Mhanje Ananas          | 2.62                            | 269.47                                 |
| आ महाजन आरासार                        | Aa Mhanje Aarsaa         | 2.52                            | 270.59                                 |
| इ महाजन क्रामार                        | I Mhanje Ijar            | 2.52                            | 289.73                                 |
| ई महाजन स्खार                         | Ee Mhanje Eedhan         | 2.46                            | 274.10                                 |
| उ महाजन उसल                         | U Mhanje Usal           | 2.54                            | 277.55                                 |
| ऊ महाजन ऊस                        | Oo Mhanje Oos            | 2.30                            | 283.30                                 |
| ए महाजन एडाका                        | E Mhanje Edaka           | 2.83                            | 281.58                                 |
| ऐ महाजन आईवाज                            | Ai Mhanje Aivaj          | 2.68                            | 263.44                                 |
| ओ महाजन ओडाहान                         | O Mhanje Odhani          | 2.66                            | 276.20                                 |
| औ महाजन ऑषधार                        | Au Mhanje Aushad         | 2.49                            | 272.62                                 |

Table 10. Pitch Tracking in mean as MN and standard deviation as SD for Hindi Vowels using a Male Sound

| Hindi Vowel with their representation | Utterances in English Format | Length of Sound Signals in Sec. | Cepstral Pitch Detection Technique in Hz |
|---------------------------------------|-----------------------------|---------------------------------|----------------------------------------|
| आ माने अनार                           | A Maane Anar                | 2.53                            | 293.12                                 |
| आ माने आम                            | Aa Maane Aam                | 2.53                            | 316.88                                 |
| इ माने इमाल                           | I Maane Imali               | 2.47                            | 305.30                                 |
| ई माने इद्हान                           | Ee Maane Edhan              | 2.65                            | 307.74                                 |
| उ माने उजाला                           | U Maane Ujala               | 2.50                            | 310.43                                 |
| ऊ माने ऑंट                            | Oo Maane Oont               | 2.23                            | 293.37                                 |
| ए माने एक                             | E Maane Ek                  | 2.28                            | 280.01                                 |
| ऐ माने ऐगक                            | Ai Maane Ainak              | 2.24                            | 272.49                                 |
| ओ माने ओडाहान                         | O Maane Odhani              | 2.38                            | 314.28                                 |
| औ माने ऑरात                            | Au Maane Aurat              | 2.19                            | 289.02                                 |

Table 11. Pitch detection in mean as MN and standard deviation as SD for 5 Isolated Hindi Consonants out of 33 Consonants using a Male Voice

| 5 Isolated Hindi Consonants with their representation | Uttered in English Format | Length of Sound Signals in Sec. | Cepstral Pitch Detection Technique in Hz |
|-------------------------------------------------------|---------------------------|---------------------------------|----------------------------------------|
| ग माने गम्ला                                            | Ga Maane Gamla            | 3.12                            | 289.99                                 |
| ज भ माने झांदा                                         | Jha Maane Jhanda          | 2.85                            | 273.09                                 |
| न भ माने नाल                                          | Na Maane Nal              | 2.61                            | 259.08                                 |
| क भ माने शक्कार                                         | Sha Maane Shakkar         | 2.89                            | 267.73                                 |
| त भ माने त्रिकोन                                         | Tra Maane Trikon           | 2.93                            | 274.33                                 |

For 11 Marathi vowels, the average of MOS score for listening rate and pleasant rate was 3.86 (to close a good rate of listening quality) and 3.58 (between pleasant and
slightly-pleasant) respectively by 19 listeners. Similarly, MOS test for Marathi consonants is 3.01 for listening quality and 3.61 for feeling pleasant.

In Hindi, MOS score test for the vowels were given the score 3.57 for listening rate and 3.51 for pleasant rate by 11 listeners (ML-05 and FL-06) as shown in Table 16. And for the consonants of Hindi, the mean of MOS score for listening rate and pleasant rate was 3.92 and 3.52 respectively in Table 17.

For English vowels, the mean of MOS test was 4.11 for listening rate and 2.99 for pleasant rate by 13 listeners (ML-07 and FL-06) in Table 18. And, the mean of MOS score for listening rate and pleasant rate is 3.99 (to close a good listening quality) and 3.69 (between pleasant and slightly-pleasant) for the English consonants respectively in Table 19.

**Table 12. Pitch detection in mean as MN and standard deviation as SD for English Vowels using a Male Speaker**

| English vowel in roman script | Duration of speech signals | Cepstral pitch tracking technique in Hz |
|------------------------------|----------------------------|----------------------------------------|
| A for Apple                  | 2.51 sec.                  | 298.47                                 |
| E for Elephant               | 2.63 sec.                  | 294.14                                 |
| I for Ice-cream              | 2.76 sec.                  | 276.75                                 |
| O for Onion                  | 2.47 sec.                  | 272.60                                 |
| U for Umbrella               | 2.64 sec.                  | 268.75                                 |

**Table 13. Pitch Tracking in mean as MN and standard deviation as SD for 5 Isolated English Spoken Consonants out of 21 Consonants in Voice of Male**

| English consonants in roman script | Length of sound signals in seconds | Cepstral pitch detection method in Hz |
|-----------------------------------|-----------------------------------|--------------------------------------|
| C for Cat                         | 2.27                              | 293.17                               |
| J for Joker                       | 2.57                              | 290.71                               |
| Q for Queen                       | 2.55                              | 284.34                               |
| T for Tiger                       | 2.42                              | 284.43                               |
| Z for Zebra                       | 2.28                              | 292.12                               |

**Table 14. Mean Opinion Score (MOS) Values with Listening Rate (LR) and Pleasant Rate (PR) for 11 Marathi Vowels using a Male Generated Voice**

| Marathi Vowel with their representation | Pronounced in English Format | Duration of Sound Signals | Average MOS Score by 19 Listeners |
|-----------------------------------------|-----------------------------|--------------------------|----------------------------------|
| अ मानणे अननत                          | A Mhanje Ananas             | 2.62 sec.                | 4.27                             |
| आ एकलाती मानत                          | Aa Mhanje Aarsaa           | 2.52 sec.                | 3.72                             |
| इ इतराती िजार                          | I Mhanje Ijar              | 2.52 sec.                | 4.36                             |
| इई मानणे इदान                       | Ee Mhanje Eedhan           | 2.46 sec.                | 4.09                             |
| उ मानणे उसार                       | U Mhanje Usal              | 2.54 sec.                | 4.09                             |
| ओ उन्शने उन्म                         | Oo Mhanje Oos              | 2.30 sec.                | 3.63                             |
| ए मानणे एडाक                         | E Mhanje Edaka             | 2.83 sec.                | 3.63                             |
| ऐ मानणे ईव                        | Ai Mhanje Aivaj            | 2.68 sec.                | 3.72                             |
| ओआ मानणे ओहात                       | O Mhanje Odhani            | 2.66 sec.                | 3.27                             |
| ऑ मानणे ऑधान                       | Au Mhanje Aushad           | 2.49 sec.                | 3.81                             |
| ओं माणणे ओषधार                       | An Mhanje Angan            | 2.66 sec.                | 3.90                             |

Mean of MOS score

3.86 3.58
Table 15. Mean Opinion Score (MOS) Values with Listening Rate (LR) and Pleasant Rate (PR) using Male Voice for 5 Isolated Marathi Consonants out of 34 Consonants

| 5 Isolated Marathi Consonants with their representation | Pronounced in English Format | Duration of Sound Signals (sec.) | Average MOS Score by 19 Listeners |
|--------------------------------------------------------|------------------------------|---------------------------------|----------------------------------|
| ग म्हणजे गवत | Ga Mhanje Gavat | 2.99 | 3.69 | 2.92 |
| झ म्हणजे झवधे | Jha Mhanje Jhabale | 3.06 | 3.38 | 3.07 |
| न म्हणजे नल | Na Mhanje Nal | 2.94 | 2.30 | 4.38 |
| श म्हणजे शहर | Sha Mhanje Shahar | 2.86 | 2.61 | 3.69 |
| ट्र म्हणजे तिकिंठ | Tra Mhanje Trikon | 3.14 | 3.07 | 4.00 |

Mean of MOS score: 3.01

Table 16. Mean Opinion Score (MOS) Values with Listening Rate (LR) and Pleasant Rate (PR) for Hindi Vowel using a Male Generated Voice

| Hindi Vowels with their representation | Uttered in English Format | Period of Sound Signals (sec.) | Average MOS Score by 11 Listeners |
|----------------------------------------|---------------------------|--------------------------------|----------------------------------|
| अ माने अनार | A Maane Anar | 2.53 | 3.72 | 3.18 |
| आ माने आँम | Aa Maane Aam | 2.53 | 3.36 | 2.63 |
| इ माने इमाली | I Maane Imali | 2.47 | 4.09 | 4.63 |
| ई माने ईदना | Ee Maane Edhan | 2.65 | 4.45 | 3.81 |
| उ माने उजला | U Maane Ujala | 2.50 | 3.54 | 3.45 |
| ऊ माने ऊंदन | Oo Maane Oont | 3.00 | 3.90 | 4.81 |
| ए माने एक | E Maane Ek | 2.28 | 3.54 | 3.54 |
| ऐ माने ऐंक | Ai Maane Ainak | 2.74 | 3.09 | 2.94 |
| ऑ माने ऑंधनी | O Maane Oodhni | 2.38 | 4.54 | 2.00 |
| औ माने ओंदता | Au Maane Aurat | 2.19 | 3.27 | 4.63 |
| एल माने एंगर | An Maane Angur | 2.19 | 2.72 | 3.36 |

Mean of MOS score: 3.57

Table 17. Mean Opinion Score (MOS) Values with Listening Rate (LR) and Pleasant Rate (PR) using Male Voice for 5 Isolated Hindi Consonants out of 33 Consonants

| Hindi Consonants with their representation | Produced in English Format | Length of Sound Signals (sec.) | Average MOS Score by 11 Listeners |
|------------------------------------------|---------------------------|--------------------------------|----------------------------------|
| ग माने गमल | Ga Maane Gamla | 3.12 | 4.81 | 3.64 |
| झ माने झंरा | Jha Maane Jhanda | 2.85 | 3.36 | 4.54 |
| न माने नल | Na Maane Nal | 2.61 | 3.45 | 3.36 |
| श माने शक्क | Sha Maane Shakkar | 2.89 | 4.00 | 3.36 |
| ट्र माने तिकिंठ | Tra Maane Trikon | 2.93 | 4.00 | 2.81 |

Mean of MOS score: 3.92

Table 18. Mean Opinion Score (MOS) Values for English Vowel using a Male Generated Voice

| English vowel with their representation | Period of sound signals (sec.) | Average MOS score by 13 listeners |
|----------------------------------------|-------------------------------|----------------------------------|
| A for Apple | 2.51 | 3.85 | 2.94 |
| E for Elephant | 2.63 | 4.28 | 2.68 |
| I for Ice-cream | 2.76 | 4.28 | 3.26 |
| O for Onion | 2.47 | 3.78 | 3.52 |
| U for Umbrella | 2.64 | 4.35 | 2.87 |

Mean of MOS score: 4.11
Table 19. Mean Opinion Score (MOS) Values using Male Voice for 5 Isolated English Consonants out of 21 Consonants

| English vowel with their representation | Period of sound signals | Average MOS score by 13 listeners |
|----------------------------------------|-------------------------|----------------------------------|
|                                        |                         | Listening rate | Pleasant rate |
| C for Cat                              | 2.27 sec.               | 4.52            | 3.94           |
| J for Joker                            | 2.57 sec.               | 4.26            | 4.15           |
| Q for Queen                            | 2.55 sec.               | 3.00            | 3.31           |
| T for Tiger                            | 2.42 sec.               | 4.21            | 3.47           |
| Z for Zebra                            | 2.28 sec.               | 3.94            | 3.57           |
| Mean of MOS score                      |                         | 3.99            | 3.69           |

The speech form is reliable to human voice for listening rate and criteria of pleasant. The listening rate of voice generation was very clear for intelligibility. Intelligibility means that how many characters and their representations are recognized correctly. The identification of speech fully depends on listeners what exactly produced the sound by a male speaker. The pleasant rate of voice generation was clear for naturalness. Naturalness means that the generated sound is closed to mankind of sound. The overall observation of MOS test is satisfactory.

6. CONCLUSION

The proposed model has performed a vital role to convert Marathi, Hindi or English text into audible form. The text would be any character with their representation among three languages. The size of text dictionary is 4,910 syllables. All speech samples in three languages have been recorded by a male speaker. And the overall size of speech corpus in Marathi, Hindi and English languages was 768 samples. In order to verify the synthetic voice generation, two bizarre tests have been applied. First test, the analysis of prosody was detected pitch frequency by cepstral pitch detection technique and to assist period of spoken signals. Pitch tracking technique has been applied on voice signals. Mean and standard deviation were calculated on pitch frequency. And the range of mean and standard deviation was [240 Hz-320 Hz] and [100 Hz-120 Hz] respectively. Another test, mean opinion score (MOS), which is a score received by various listeners. So, two distinct parameters of MOS have been used. First parameter is listening rate and another one is criteria of pleasant. Initially, the result of listening rate was considered between fair and good. Second MOS parameter, the result has been positively received the score between pleasant and slightly-pleasant for natural sound. For MOS test, the overall size of opinions was 1669 using two parameters. The achievement of the present model enables to recognize the character with their representation in Marathi, Hindi and English languages and near to mankind sound.

FUTURE SCOPE

In future work, speech synthesizer will be attempted on various numbers system in Marathi, Hindi, English languages. And the spoken samples will be recorded by speakers in noisy environment. The prosody analysis of voice signals will have to be evaluated by a machine.

ACKNOWLEDGEMENTS

The authors would like to thank the Rajiv Gandhi Science and Technology Commission, North Maharashtra University Centre, Govt. of Maharashtra, India for funding the project (Code No. 7-II-DP/2014) and G. H. Raisoni Doctoral fellowship, North Maharashtra University, Jalgoan (MH-India).

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Katsanobu Fushikida, Yukio Mitome, Yuji Inoue. A text to speech synthesizer for the personal computer. IEEE Transactions on Consumer Electronics. 1982;CE-28(3): 250-256.
2. Richard Sproat. Multilingual text analysis for text-to-speech synthesis. Journal on Natural Language Engineering. 1996;2(4):369-380
3. Norbert Pachler. Speech technologies and foreign language teaching and learning. The Language Learning Journal. 2002; 26(1):54-61.
4. Gerry Kennedy. Benefits of text to speech software. Australian Journal of Learning Disabilities. 2003;8(3):31-34.

5. Bhuvana Narasimhan, Richard Sproat, George Kiraz. Schwa-deletion in Hindi text-to-speech synthesis. International Journal of Speech Technology. 2004;7:319-333.

6. Diemo Schwarz. Concatenative sound synthesis: The early years. Journal of New Music Research. 2006;35(1):3-22.

7. Hwai-Tsu Hu, Hsin-Min Wang. Integrating coding techniques into LP-based Mandarin text-to-speech synthesis. International Journal Speech Technology. 2007;10:31-44.

8. Dominic Wyse, Usha Goswami. Synthetic phonics and the teaching of reading. British Educational Research Journal. 2008;34(6):691-710.

9. Zervas P, Fakotakis N, Kokkina Kis G. Development and evaluation of a prosodic database for Greek speech synthesis and research. Journal of Quantitative Linguistics. 2008;15(2):154-184.

10. Naim R. Tyson, Ila Nagar. Prosodic rules for schwa-deletion in Hindi text-to-speech synthesis. International Journal Speech Technology. 2009;12:15-25.

11. Pamela Chaudhur, Vinod Kumar K. Vowel classification based approach for Telugu text-to-speech system using symbol concatenation. International Conference [ACCTA-2010]. 2010;1(2):183-187.

12. Arun Soman, Sachin Kumar S, Hemanth VK, Sabarimalai Manikandan M, Soman KP. Corpus driven malayalam text-to-speech synthesis for interactive voice response system. International Journal of Computer Applications (0975-8887). 2011;29(4):41-46.

13. Kwan Min Lee, Younbo Jung, Clifford Nass. Can user choice alter experimental findings in human-computer interaction? Similarity attraction versus cognitive dissonance in responses to synthetic speech. International Journal of Human-Computer Interaction. 2011;27(4):307-322.

14. Matej Rojc and Zdravok Kacic. Gradient-descent based unit-selection optimization algorithm used for corpus-based text-to-speech synthesis. Applied Artificial Intelligence: An International Journal. 2011;25(7):635-668.

15. Lakshmi Sahu, Avinash Dhole. Hindi & Telugu Text-to-Speech (TTS) and inter-language text conversion. International Journal of Scientific and Research Publications. 2012;2(4):1-5.

16. Catalin Ungurean, Dragoș Burileanu, Mihai Surmei. Statistically augmented preprocessing/normalization module for a romanian text-to-speech system. IEEE 7th Conference on Speech Technology and Human-Computer Dialogue (SpeD). 2013;1-6.

17. Madiha Jalil, Faran Awais Butt, Ahmed Malik. A survey of different speech synthesis techniques. IEEE International Conference on Technological Advances in Electrical, Electronics and Computer Engineering (TAEECE). 2013;204-207.

18. Shreekanth T, Udayashankara V, Arun Kumar. An unit selection based hindi text to speech synthesis system using syllable as a basic unit. IOSR Journal of VLSI and Signal Processing (IOSR-JVSP). 2014;4(4):49-57.

19. Vivek Hanumante, Rubi Debnath, Disha Bhattacharjee, Deepi Tripathi, Sahadev Roy. English text to multilingual speech translator using android. International Journal of Inventive Engineering and Sciences (IJIES). 2014;2(5):4-9.

20. Mukta Gahlawat, Amita Malik, Poonam Bansal. Natural speech synthesizer for blind persons using hybrid approach. 5th Annual International Conference on Biologically Inspired Cognitive Architectures (BICA). 2014;4:83-88.

21. Theophile K. Dagba, Charbel Boco. A text to speech system for fon language using Multisyn algorithm. 18th International Conference on Knowledge-Based and Intelligent Information & Engineering – KES2014. 2014;35:447-455.

22. Sunil S. Nimbhore, Ghanshyam D. Ramteke, Rakesh J. Ramteke. Implementation of English-Text to Marathi-Speech (ETMS) synthesizer. IOSR Journal of Computer Engineering (IOSR-JCE). 2015;17(1):34-43.

23. Fiona S. Baker. Emerging realities of text-to-speech software for nonnative-english-speaking community college students in the freshman year. Community College Journal of Research and Practice. 2015;39(5):423-441.

24. Soumya Priyadarsini Panda, Ajit Kumar Nayak. An efficient model for text-to-speech synthesis in Indian languages. International Journal Speech Technology. 2015;18(3):305-315.
25. Smita P. Kawachale, Chitode JS. Position based syllabification and objective spectral analysis in Marathi text to speech for naturalness. International Journal Speech Technology. 2015;18(3):367-386.
26. Ramteke GD, Ramteke RJ. Text-to-speech synthesis of Marathi numerals. International Journal of Engineering and Technical Research (IJETR). 2015;3(7):360-367.
27. Saleh M. Abu-Soud. ILA talk: A new multilingual text-to-speech synthesizer with machine learning. International Journal Speech Technology. 2016;19(1):55-64.
28. John OR. Aoga, Theophile K. Dagba, Codja C. Fanou. Integration of Yoruba language into MaryTTS. International Journal Speech Technology. 2016;19(1):151-158.
29. German Bordel, Mikel Penagarikano, Luis Javier Rodriguez-Fuentes. Probabilistic Kernels for improved text-to-speech alignment in long audio tracks. IEEE Signal Processing Letters. 2016;23(1):126-129.
30. Shinnosuke Takamichi, Tomoki Toda, Alan W. Black, Graham Neubig, Sahriani Sakti, Satoshi Nakamura. Post-filters to modify the modulation spectrum for statistical parametric speech synthesis. IEEE/ACM Transactions on Audio, Speech and Language Processing. 2016;(99):1-13.