International Journal of Innovative Studies in Sciences and Engineering Technology (IJISSET)
www.ijisset.org Volume: 2 Issue: 1 | January 2016

Text-to-Speech Conversion

Mohd Bilal Ganai[1], Er Jyoti Arora[2]

[1]*M.Tech Student, Department of Computer Science, Desh Bhagat University, Punjab*
[2]*Assistant Professor, Department of Computer Science, Desh Bhagat University, Punjab*

Abstract:- *In the present time, text to speech conversions play a very important role in-order to understand and analyze the strength of the existing system and to get the appropriate and definite signature of the system. By using concatenated phoneme library techniques, we have implemented text to speech conversions. In this model we used phoneme audio library and developed a system which produces more human like speech. The system is an advantage over available text to speech conversion system in the way our system produces the sound which is very much human like. Aim of this approach to develop high-quality Text-to-Speech (TTS).*

Keywords:- *Text-to speech synthesis, unit selection, hybrid TTS, HMM, HTS, VLRs and NVLRs.*

1. INTRODUCTION

From the last decade, the quality of text-to-speech (TTS) [1] has been improved dramatically. In General, human-like clear sound can be generated from waveform-based speech synthesis but it necessarily has a huge amount of speech data with the control flexibility. At the same time, smooth sound can be generate from ANN based speech synthesis with a small amount of speech data. Further, it has flexibility in controlling speaker individuality [7]. In ANN based speech synthesis, the system is developed by creating a library of phonemes, a library of phoneme audio files and a dictionary of words with their phoneme representation. The system generates takes a sentence, analyze each word and find out their corresponding phonemes. Then it concatenates all the phonemes from the phoneme audio library and then plays the audio which sounds like a speech of sentence. It also displays a waveform and spectrum of the generated speech sound.

In both streams, intensive studies have been done to improve speech quality [6]. This smoothing might cause preferred smoothness of ANN-based speech synthesis [8], especially for the low-frequency band synthesis [8], because speech energy in the low frequency band is high and because discontinuity in the low frequency band [9] is audible. However, in terms of the high frequency band, spectrum fluctuation is more random and speech energy is lower. Therefore, smoothening is a more severe problem that discontinuity in the high frequency band [10]. Several methods are available to perform hybrid synthesis. For example, waveforms segments are selected to match the parameters generated by ANN based speech synthesis and are concatenated in the mode of waveform based speech synthesis [11]. The major differences arise mainly in the criteria of ANN parameters estimation as well as in the length of speech segments. Another approach is to mix time analysis with the speech segments [12]; ANN based speech synthesis can also generate speech waveform [12]. we compare similarity of synthetic voice with the original voice because of simple source modeling and smoothing of modeled parameters. In order to improve the synthesis quality, A ANN method gives good quality when the system covers all the phonetic contexts [4]. The Natural Language Processing unit handles phonetization and intonation along with rhythm and it outputs a phonetic transcript of the input text [5]. The Digital Signal Processing unit transforms the phonetic transcript it receives into machine speech [1]. The text analyzer comprises of four basic parts, namely: a pre-processing block, a morphological analysis block, a contextual analysis block and a syntactic-prosodic parser. The pre-processing block converts abbreviations, numbers and acronyms into full text when necessary. It also breaks input sentences into groups of words. The morphological analysis block categorizes each word in the sentence being analyzed into possible parts of speech, on the basis of the words spelling. Compound words are decomposed into their basic units in this module. The contextual analysis module streamlines the list of possible parts of speech of words in sentences, by considering the parts of speech of neighboring words. The syntactic-prosodic parser locates the text structure (in terms of clause and phrase constituents) that tends more closely to the prosodic realization of the input sentence. A Letter-To-Sound (LTS) module is used for phonetic transcription of incoming text. Worthy of note here however, is the fact that this transcription is beyond a dictionary look up operation. This is because most words have

different phonetic transcriptions depending on context. Also, pronunciation dictionaries do not account for morphological variations in words. In addition, pronunciations of words in sentences differ from pronunciation of those same words when they are isolated. Furthermore, not all words are present in a phonetic dictionary. Extractive systems tend to produce summaries with very long sentences; longer sentences score higher on metrics that rate them for importance. Abstractive approaches to single document summarization address this problem by editing the extracted sentences. They reduce a sentence by eliminating constituents which are not crucial for its understanding or salient enough to include in the summary. These approaches are based on the observation that the "importance" of a sentence constituent can often be determined based on shallow features, such as its syntactic role, the words it contains and their relation to surrounding sentences.

2. RELATED WORK

[1] In this paper 2014.Inoue, Takuma et.al. [1] has been worked on hybrid text-to-speech based on sub-band approach. This paper proposes a sub-band speech synthesis approach to develop high-quality Text-to-Speech (TTS). For the low-frequency band and high-frequency band, Hidden Markov Model (HMM)-based speech synthesis and waveform-based speech synthesis are used, respectively. Both speech synthesis methods are widely known to show good performance and to have benefits and shortcomings from different points of view. One motivation is to apply the right speech synthesis method in the right frequency band. Experiment results show that in terms of the smoothness the proposed approach shows better performance than waveform-based speech synthesis, and in terms of the clarity it shows better than HMM-based speech synthesis.

[2] 2013. Schultz, Tanja et. al. [3] has been implemented the Global Phone. The global phone is a multilingual text & speech database in 20 languages. This paper describes the advances in the multilingual text and speech database Global Phone, a multilingual database of high quality read speech with corresponding transcriptions and pronunciation dictionaries in 20 languages. Global Phone was designed to be uniform across languages with respect to the amount of data, speech quality, the collection scenario, the transcription and phone set conventions. With more than 400 hours of transcribed audio data from more than 2000 native speakers Global Phone supplies an excellent basis for research in the areas of multilingual speech recognition, rapid deployment of speech processing systems to yet unsupported languages, language identification tasks, speaker recognition in multiple languages, multilingual speech synthesis, as well as monolingual speech recognition in a large variety of languages.

3. HYBRID TTS SYSTEM BASED ON A SUB-BASED APPROACH

3.1. Outline of the Proposed Work

To verify the effectiveness of the sub-band approach, we simply combined a waveform-based TTS system and an HMM based TTS system. In terms of high-frequency band, speech synthesis methods are implemented: speech synthesis waveform concatenation and Straight vocoder [13]. Figure 1 presents an overview of the proposed system. Details are explained in the following sections.

3.2. Waveform Concatenation Methods in High Frequency Band

As depicted in fig 2, output of a waveform based TTS system and HMM based TTS system are superposed in a time domain after filtering and a low pass filter, respectively. In this implementation, we resume the following two points:

1) **Raw signals** in the high frequency band should be used because waveform based speech synthesis generates clear and human like sound from using non processed raw signal.
2) **The** influence of FO harmonics in high frequency band is small. Therefore, fine tunning by PSOLA is not necessary in the high frequency band.

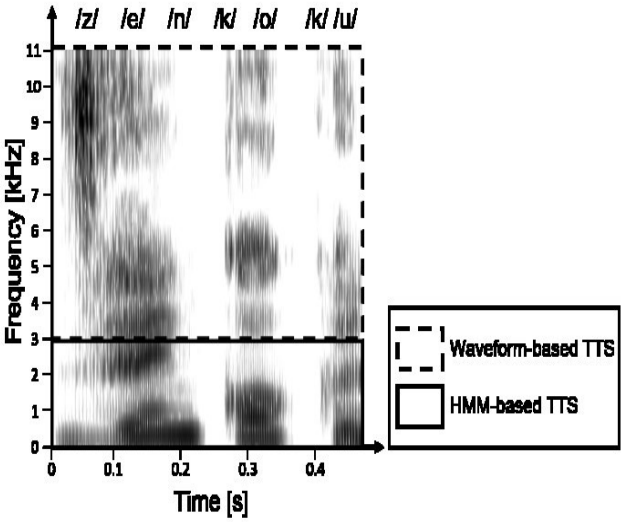


Fig 1: *Two sub-band in the proposed system*

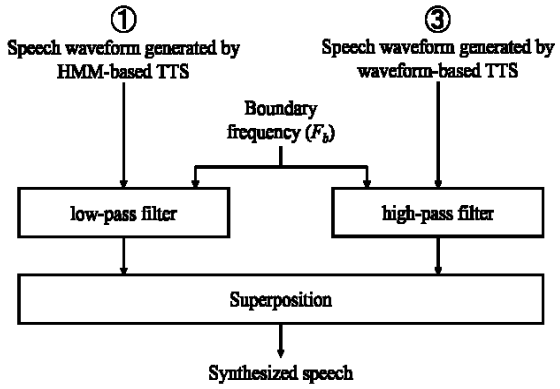


Fig 2: *(Method 1) Waveform concatenation method in high frequency*

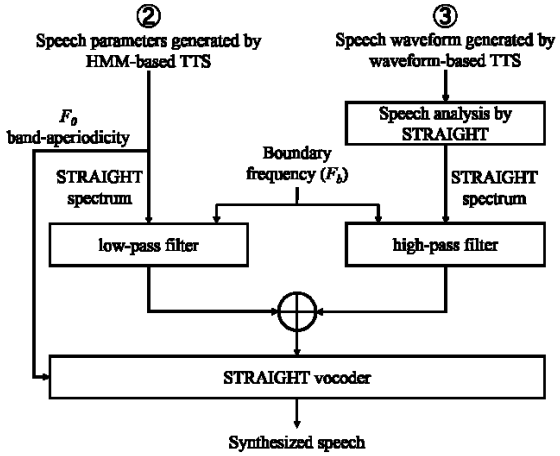


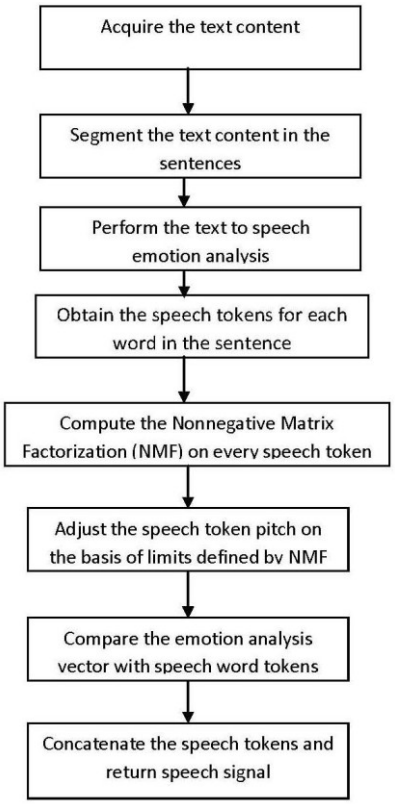
Fig 3: *(Method 2) STRAIGHT vocoder in High frequency band*

A. Straight Vocoder in High Frequency Band

A difference method is speech synthesis by straight vocoder for the high frequency band. The selected speech segments and prosodic parameter are identical to these described for method. In other words, signals in the high frequency band are regarded as synthesized by a conventional non uniform unit vocoder based TTS, not a waveform based TTS [6]. As portrayed in fig 3, the spectra of low frequency and high frequency band, are combined in the spectrum domain, and the full band speech is synthesized by a straight vocoder using a mixed excitation, source generated by an HMM based TTS system. Because of this is the phase spectra between low and high frequency bands are matched in method D.

4. PROPOSED WORK

The literature would be studied in detail on the text to speech conversion techniques, speech wave generation or concatenation techniques and speech recognition and enhancement techniques in order to know their workflow, advantages and disadvantages. The literature survey will be carefully conducted to find the research gaps in the existing speech waveform techniques. Then the proposed model will be designed and improved to remove the shortcomings and research gaps of the existing schemes. The proposed model will be then implemented using the MATLAB. The proposed model will be designed with all essential input and output parameters. Then the proposed model will be tested and debugging would be done if any errors or functional abnormalities would be found. Afterwards, the final results would be obtained and analyzed, and compared with the existing model results in order to form the final conclusion



Flow diagram of proposed work

The text-to-speech conversion system is coded in MATLAB. The system is developed by creating a library of phonemes, a library of phoneme audio files and a dictionary of words with their phoneme representation. The system generates takes a sentence, analyze each word and find out their corresponding phonemes. Then it concatenates all the phonemes from the phoneme audio library and then plays the audio which sounds like a speech of sentence. It also displays a waveform and spectrum of the generated speech sound.

5. EXPERIMENTS AND RESULTS

The text-to-speech conversion system is coded in MATLAB. The system is developed by creating a library of phonemes, a library of phoneme audio files and a dictionary of words with their phoneme representation.

The system generates takes a sentence, analyze each word and find out their corresponding phonemes. Then it concatenates all the phonemes from the phoneme audio library and then plays the audio which sounds like a speech of sentence. It also displays a waveform and spectrum of the generated speech sound.

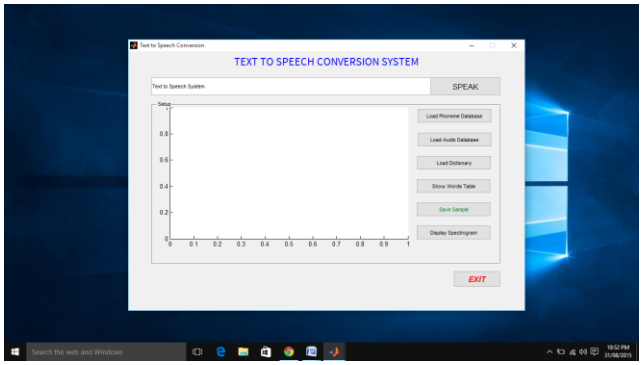


Figure 1: *The Front Screen of the Text to Speech Conversion System Developed in MATLAB*

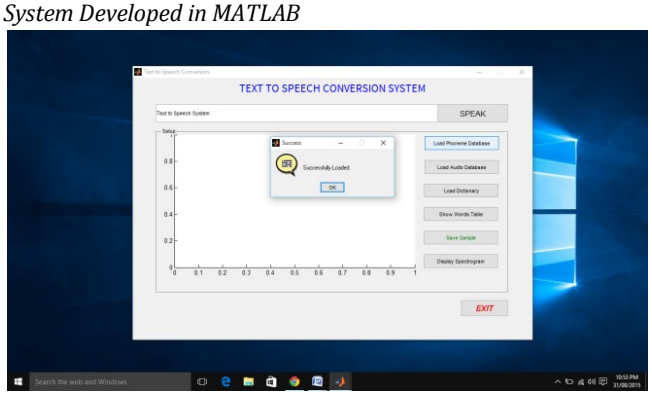


Figure 2: *Phoneme Table is Loaded Successfully*

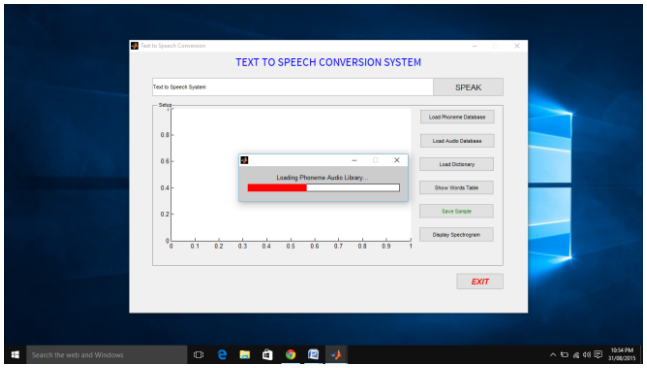


Figure 3: *Phoneme Audio Library is loading*

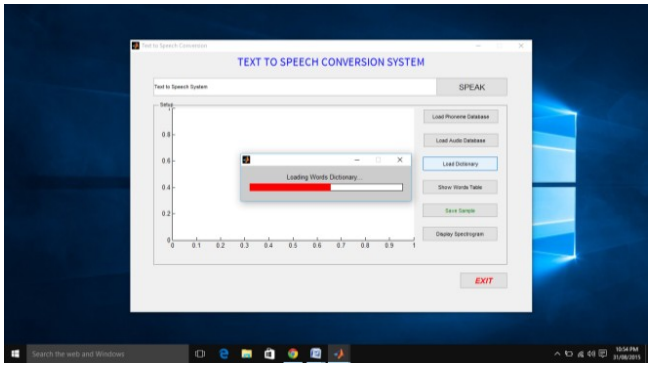


Figure 4: *Words Library is loading*

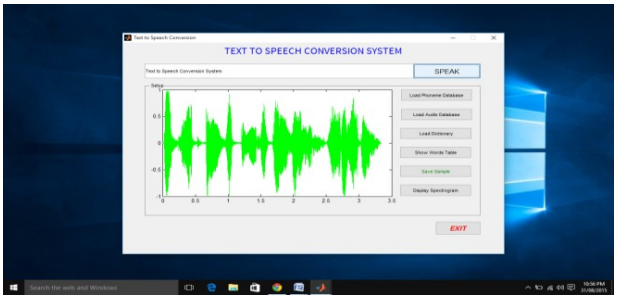


Figure 5: *The Audio Waveform of the Spoken Text: "Text to Speech Conversion System"*

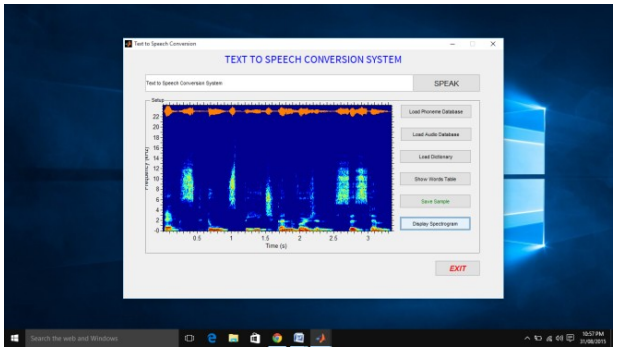


Figure 6: *Spectrogram of the Spoken Text*

6. CONCLUSION

As described in this paper, we proposed a sub-band speech synthesis approach to develop high-quality Text-to-Speech (TTS). Our motivation is that speech synthesis methods are not necessarily identical for all spectrum bands, but they should be adaptive. As the first step, we developed a TTS system by combining the respective benefits of the two methods; HMM-based speech synthesis and waveform-based speech synthesis. According to experimental results, we can say that preliminary benefits of the approach were confirmed. However, the current system is not good enough as overall performance. As future works, we would like to improve the system to perform detail controls.

REFERENCES

[1] Barnett, M.P; Ruhsam, W.M.." A Natural Language programming system for text processing"2007. Engineering writing and speech, IEEE Transactions.

[2] Adiga, Nagaraj, and S. R. MahadevaPrasanna. "A hybrid Text-to-Speech synthesis using vowel and non vowel like regions." In India Conference (INDICON), 2014 Annual IEEE, pp. 1-5. IEEE, 2014.

[3] Schultz, Tanja, Ngoc Thang Vu, and Tim Schlippe. "GlobalPhone: A multilingual text & speech database in 20 languages." In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pp. 8126-8130. IEEE, 2013.

[4] Neumann, Lukáš, and JiříMatas. "A real-time scene text to speech system." In Computer Vision–ECCV 2012.Workshopsand Demonstrations, pp. 619-622. Springer Berlin Heidelberg, 2012.

[5] Zue, v; Glass J; Goodline,D; Lueng; H; Phillips,M; Polifroni,J; Seneffs,S" Integration of speech recognition and natural language processing in the Mit VOYAGER system".1991. Acoustics, Speech, and signal processing, 1991.Acoustics, speech, and signal processing, 1991;. ICASSP -91, 1991 International conference.

[6] Ze, Heiga, Andrew Senior, and Mike Schuster. "Statistical parametric speech synthesis using deep neural networks." In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pp. 7962-7966. IEEE, 2013.

[7] Fan, Hao-Teng, Jeih-weih Hung, Xugang Lu, Syu-Siang Wang, and Yu Tsao. "Speech enhancement using segmental nonnegative matrix factorization." In Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, pp. 4483-4487. IEEE, 2014.

[8] Murase, Yoshikazu, Hironobu Chiba, Nobutaka Ono, Shigeki Miyabe, Takeshi Yamada, and Shoji Makino. "On microphone arrangement for multichannel speech enhancement based on nonnegative matrix factorization in time-channel domain." In Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA), pp. 1-5. IEEE, 2014.

[9] Gerkmann, Timo. "MMSE-optimal enhancement of complex speech coefficients with uncertain prior knowledge of the clean speech phase." In Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, pp. 4478-4482. IEEE, 2014.

[10] Gonzalez, Sira, and Mike Brookes. "Mask-based enhancement for very low quality speech." In Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, pp. 7029-7033. IEEE, 2014.

[11] Sprechmann, Pablo, Alex M. Bronstein, and Guillermo Sapiro. "Supervised non-euclidean sparse NMF via bilevel optimization with applications to speech enhancement." In Hands-free Speech Communication and Microphone Arrays (HSCMA), 2014 4th Joint Workshop on, pp. 11-15. IEEE, 2014.

[12] Zou, Y. X., Y. Q. Wang, Peng Wang, C. H. Ritz, and Jiangtao Xi. "An effective target speech enhancement with single acoustic vector sensor based on the speech time-frequency sparsity." In Digital Signal Processing (DSP), 2014 19th International Conference on, pp. 547-551. IEEE, 2014.

[13] Toda, K. Tokuda, "Speech parameter generation algorithm considering global variance for HMM-based speech synthesis," Proc. of Eurospeech'05, pp. 2801–2804, 2005.