Text To Speech Conversion Using Different Speech Synthesis

Hay Mar Htun, Theingi Zin, Hla Myo Tun

Abstract: Text to speech (TTS) synthesis is the automatic conversion of text into speech. Generally, TTS system consists of two phases. The first is text analysis, where the input text is transcribed into a phonetic or some other linguistic representation. The second one is the generation of speech waveforms. In this TTS system, text to phoneme conversion depends on dictionary based approach to get the exact phonetic transcription. Speech synthesis such as domain specific, phoneme based synthesis and unit selection synthesis are used for concatenating speech. For numerical text to speech system, domain specific synthesis is applied. In phoneme based synthesis, the input text is considered as word to produce sound. For input sentence , unit selection speech synthesis is applied. This TTS system is mainly used for visual impairments and handicapped people.

Keywords: Text to speech conversion, Domain specific synthesis, Phoneme based synthesis, Unit selection synthesis

1. INTRODUCTION

The text-to-speech (TTS) synthesis is to convert an arbitrary input text into intelligible and natural sounding speech. TTS system includes mainly two parts: natural language processing and digital signal processing. The general block diagram of TTS system is shown in figure 1. Natural language processing contains three steps. They are text analysis, phonetic analysis and prosodic analysis. The text analysis includes segmentation, text normalization, and part of speech (POS) tagger. Phonetic conversion is to assign phonetic transcription to each word. There are two approaches in phonetic conversion. They are rule based and dictionary based approaches. Rule based is applied for unknown words whereas dictionary based is used for known words. Prosodic analysis is to determine intonation, amplitude and duration modeling of speech. It describes speaker's emotion.

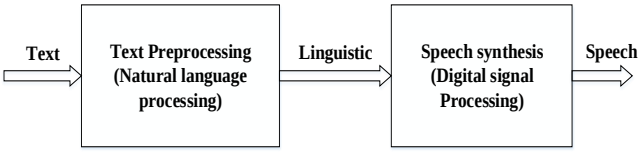

Figure1 .*General block diagram of Text to speech (TTS)*

Digital Signal Processing refers to speech synthesis. The most important qualities of a speech synthesis system are naturalness and intelligibility. Naturalness expresses how the output sounds like human speech, whereas intelligibility is the easiness with which the output is understood. The technologies for generating synthetic speech waveforms are concatenative synthesis, formant synthesis and articulartorysynthesis [1]. In formant synthesis, speech can be constantly intelligible. It does not have any database of speech samples. So the speech is artificial and robotic. Articulatory synthesis technique is based on the models of human vocal tract for synthesizing speech. Among these, Concatenative synthesis is the primary technology for speech synthesis. It is based on prerecorded natural sounds database. But it is limited to one speaker and usually require more memory capacity. This approach use a real recorded speech as the synthesis units such as: phoneme, syllable, or word and concatenate the units together to produce speech. There are three main sub-types of concatenative synthesis. They are unit selection synthesis, diphone synthesis and domain specific synthesis. In certain systems, this part includes the computation of target prosody (pitch contour, phoneme duration), which is then imposed on the output speech. In this paper, domain specific synthesis is applied to join recorded speech. Text-to-speech synthesis is an useful hardware and software tool in many application areas such as vocal monitoring system for blind people, web browser, mobile phones, personal computer and so forth. Furthermore, TTS system is currentlydeveloped inteaching aids, text reading, and talking books/toys. However, most TTS systems only focus on a limited domain of applications. [2]. In this TTS system, three types of speech synthesis such as domain specific synthesis , phoneme based speech synthesis and unit selection synthesis are applied differently depend on the input text. For input numbers and one syllable words , domain specific and phoneme based speech synthesis are applied . For input sentence , unit selection synthesis is used.

- *Nway Nway Kyaw Win, Theingi Zin, Hla Myo Tun*
  Department of Electronics Engineering
  Mandalay Technological University

2. METHODOLOGY

Text to speech system has two parts namely natural language processing and speech synthesis (digital signal processing).

2.1 Natural Language Processing (NLP)

NLP produces phonetic transcription together with prosodic feature of the input text. In this TTS system, NLPcomprises of three main components such as text analysis, phonetic conversion and prosodic phrasing.

2.1.1Text Analysis

In this TTS system, the input sentence is segmented into token. After tokenization, each word is determined as part of

speech (POS) tagging. Part-of-speech is a process assigning correct POS tag to each word in a sentence from a given set of tags. Bigram Model is used for POS tagger. This method is to perform POS Tagging to determine the most likely tag for a word, given the previous and next tags [3]. This can be calculated by using equation (1).For Bigrams, the probability of a sequence is just the product of conditional probabilities of its Bigrams. So if $t_1, t_2, \ldots t_n$ are tag sequence and $w_1, w_2, \ldots\ldots w_n$ are corresponding word sequence .

$$P(t_i|w_i) = P(w_i|t_i) \cdot P(t_i|t_{i+1}) \quad (1)$$

Where t_i denotes the tag sequence and w_i denotes the word sequences. $P(w_i|t_i)$ is the probability of current word given current tag. Here, $P(t_i|t_{i+1})$ is the probability of a current tag given the previous tag. This provides the transition between the tags and helps capture the context of the sentence.

2.1.2 Phonetic Conversion

In this system, Dictionary based approach is used for phonetic transcription of input word. So, any type of input text that does not include in the dictionary cannot run.

2.1.3 Prosodic Phrasing

Prosodic Phrasing is to assign the phrase of the input text .In this part, chink 'n' chunk prosodic phrasing, shown in figure 2, is used. In this model, word classes are identified into chink and chunk group .Then, input words are compared with chink or chunk group. Prosodic phrase break is automatically set when a word belonging the chunks groups. This method basically corresponds to function and content word classed, with some minor modification.

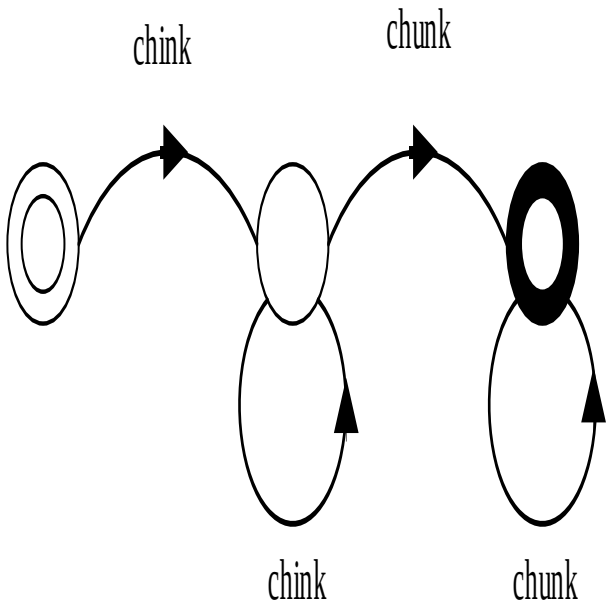

Figure 2.*Simple prosodic phrase chink 'n chunk*

2.2 Speech synthesis

The speech synthesis is to produce speech as natural and intelligible sound .There is many methods in speech synthesis. Among these, concatenative speech synthesis is natural in comparison with other methods. In this TTS system, sub-types of concatenative synthesis such as unit selection speech synthesis, phoneme based speech synthesis and domain specific synthesis are applied.

2.2.1 Unit Selection Speech synthesis

This algorithm selects an optimum set of acoustic units from the speech database to match with the given phoneme stream and target prosody. A selection mechanism using two cost functions – target cost and concatenation (join) cost is applied to find the best sequence of units [4] .The target cost function typically consists of several subcomponents of phonological features such as identity of its context, positional features and numerical features such as phrasing. The target cost, $C^t(t_i|u_i)$ can be computed by the following equation (2)

$$C^t(t_i|u_i)=\sum_{j=1}^{q} w_j^t \, C_j^t(t_t|u_i) \quad (2)$$

where p represents the number of the target cost components,w_j^t and is a feature weight of the j-th component. The concatenation or joint cost function accounts for the acoustic matching between pairs of consecutive candidate units and it can be calculated by using equation (3)

$$C^c(u_{i-1}|u_i)=\sum_{j=1}^{q} w_j^c \, C_j^c(u_{i-1}|u_i) \quad (3)$$

where q represents the number of the concatenation cost components

The unit selection module is to find the speech unit sequence which is described in equation (4)

$$C_1^n=\text{minC}(t_1^n, u_1^n) \quad (4)$$

The selection of the optimal speech unit sequence incorporates a Viterbi search.

2.2.2 Phoneme based speech synthesis

Phonemes are the small pieces of speech unit. English language has about 44 phonemes of which 22 sounds are vowels and 22 sounds are consonants [5]. Phoneme based speech synthesis is the concatenation of phonetic units to form word. . Using phonemes as the synthesis unit requires a small storage, but it causes little discontinuity between adjacent units.

2.2.3 Domain-specific Synthesis

Domain-specific synthesis concatenates pre-recorded words and phrases to create complete utterances. The technology is very simple to implement .The level of naturalness of these systems can be very high because the variety of sentence types is limited, and they closely match the prosody and intonation of the original recordings.

3. IMPLEMENTATION

Different types of speech synthesis systems such as domain specific synthesis, phoneme based synthesis and unit selection synthesis is implemented in this TTS system.

3.1 Domain Specific Synthesis

The flowchart of text to speech synthesis based on domain specific synthesis for numbers is illustrated in figure.3. In this

part, only numbers are considered as input text.Firstly, speech is recorded for each number and converted to wav file format. Database (lexicon) for digit 0 to 9 is also constructed to compare with the input number .Respective speech are chosen digit by digit based on the numbers. These sounds are then concatenated to generate the wav file. If the input is one digit, the speech can be produced directly. When the input is two or more digits, it is necessary to concatenate each digit.

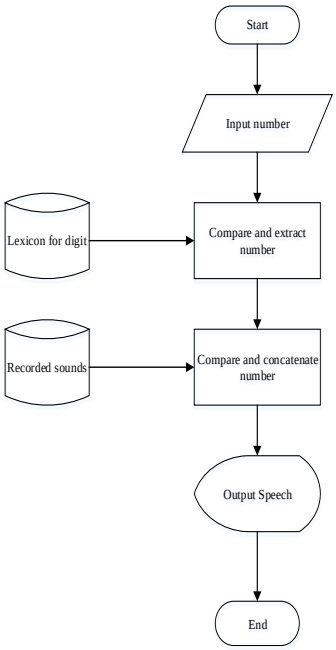

Figure .3 *Flowchart of text to speech synthesis based on domain specific synthesis for numbers*

3.2 Phoneme based speech synthesis

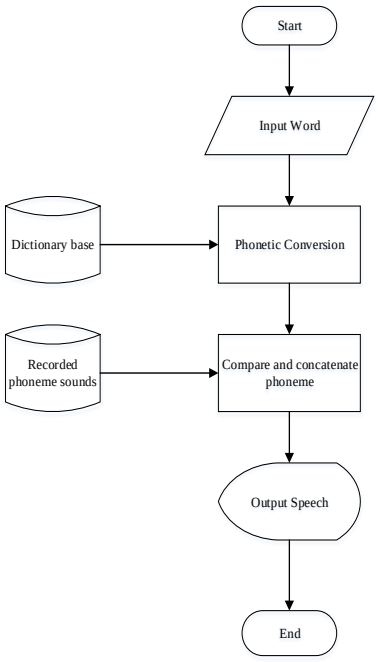

Figure 4. *Flowchart of Phoneme based text to speech synthesis for words*

The flowchart of phoneme based text to speech synthesis for words is shown in figure .4.In this part; the input text is considered only syllable word to produce speech as natural. Firstly, the input word is given from the keyboard of computer .In the next step, it is necessary to convert from word to phonetic transcription which is also called “grapheme to phoneme” conversion. Dictionary based approach, more exact than rule based approach, is applied in this step. Then, phoneme sounds are concatenated by depending on the phonetic transcriptions of word to produce speech.

3.3 Unit Selection Speech Synthesis

In this system, input text of abbreviations, numbers and symbols are not considered. Figure .5 shows the block diagram of TTS using unit selection synthesis. In natural language processing, firstly the input sentence is spitted into words and Part of speech (POS) tagger is assigned to each word by using bigram method. Then, each word is converted to phoneme transcription with the use of dictionary based approach. Simple prosodic phrase chink ‘n’ chunk are applied in this system to get speech naturally. These linguistic features are taken as an input of unit selection.

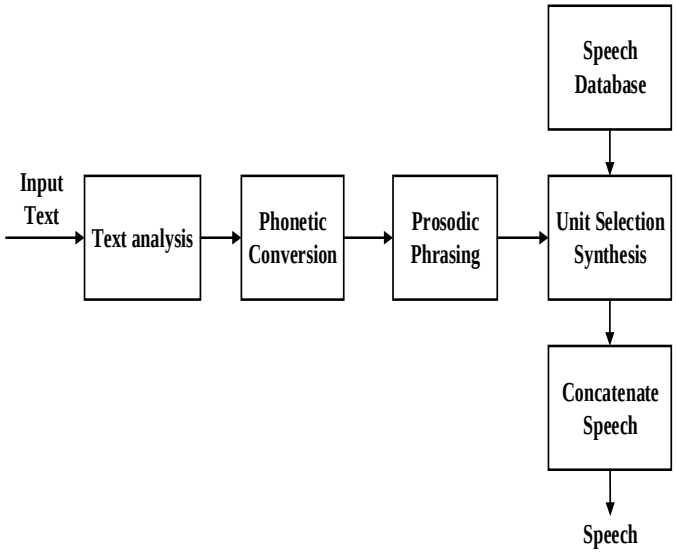

Figure 5. *Block diagram of Text to Speech using Unit Selection Synthesis*

In speech synthesis, speech is recorded at sampling rate of 16kHz. Then these recorded speeches are segmented phonetically in discrete time domain. So the segmented units are stored according to their sample values. There are two approaches in speech segmentation. They are handlabeling and automatic speech segmentation. In this TTS system, hand-labeling is applied because database is small. But it is time consuming and has little errors in this method. Unit selection algorithm selects the best acoustic units which match the target linguistic features. Then these units are concatenated to produce speech.

4. SIMULATION RESULT

Software implementation is based on MATLAB programming language. In domain specific synthesis, the input number (one or more digits) can be pronounced speech easily and quickly. The output speech is natural and intelligible like human speech. But the domain specific synthesis is not general-purpose .For the combination of words; it must be pre-programmed to synthesize speech.

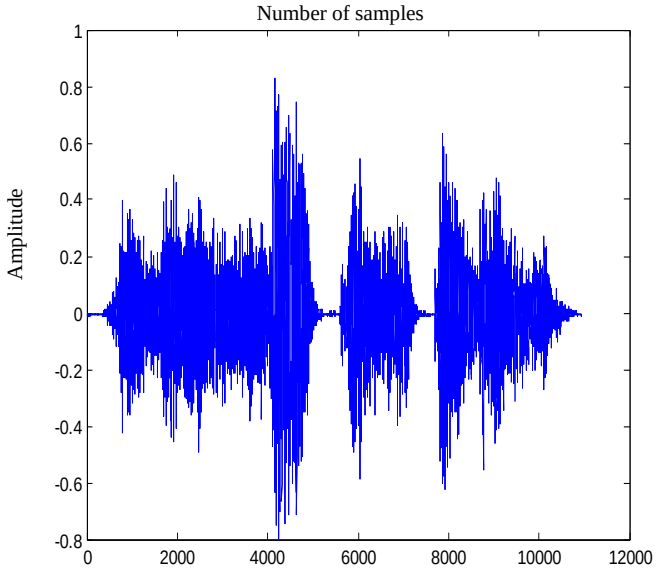

Figure .6 *Output speech for number sixty*

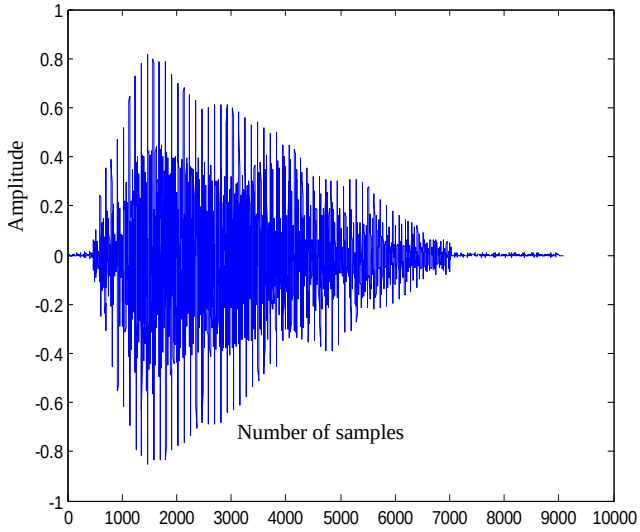

Figure .7 *Output speech for number five*

The output speech signals for number five and sixty are as shown in figure .6 and 7. This figure is only for one digit and thus there is no effect whether it contains delay or not in speech. The output speech quality is excellent.

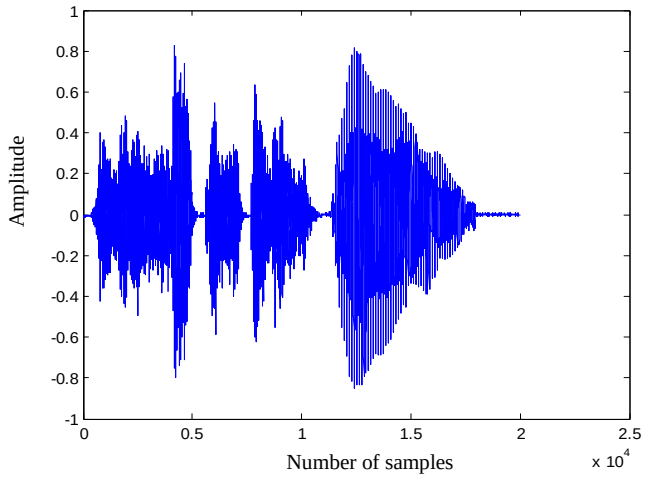

Figure 8. *Output Speech for Number Sixty- Five after Concatenation of Sixty and Five Waves*

This figure .8 illustrates the concatenation of sixty and five waves to get sixty-five as output speech. In this figure, the sampling rate of output speech is equal for two digits. Before concatenation, speech waves are already removed delay to get continuous speech. Finally, the two speech waves are concatenated digit by digit. In phoneme based speech synthesis, the output speech is produced by concatenation of phoneme sounds.

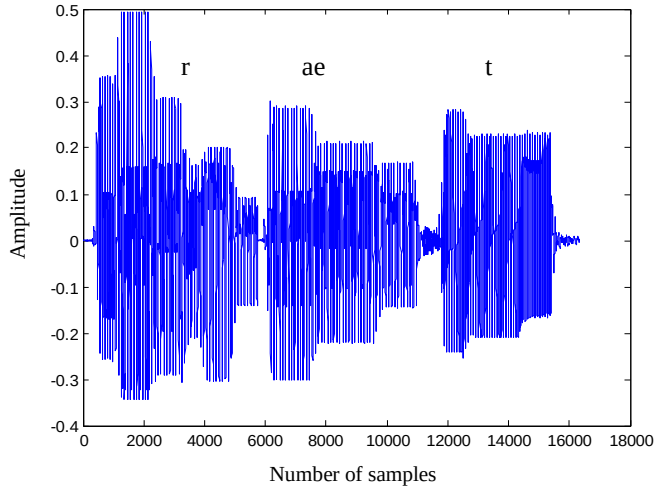

Figure 9. *Output speech for word 'rat' after concatenation of phonemes r, ae and t waves*

Figure.9 shows the output speech of 'rat' based on the phoneme concatenation. The sampling rate of speech signal is 16 kHz. This is the integration of each phoneme where delay is removed between the units. The word is only for one syllable and concatenate similar phoneme sound. This is easy to implement. If two or more syllable words are created, syllabification method will be more appropriate.

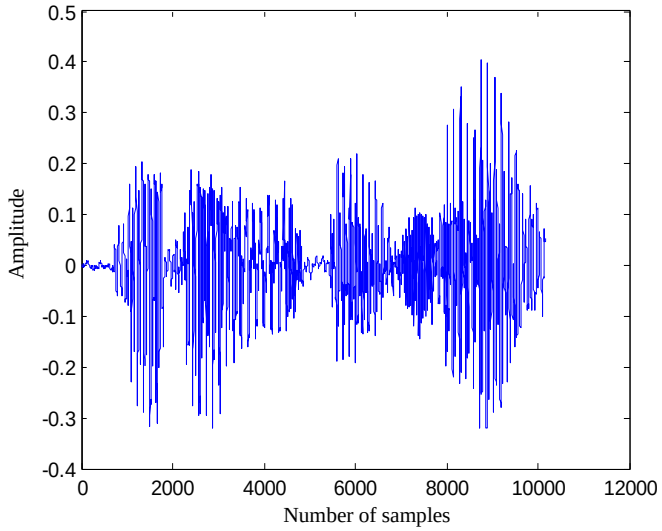

Figure .10 *Output speeches for "It is easy"*

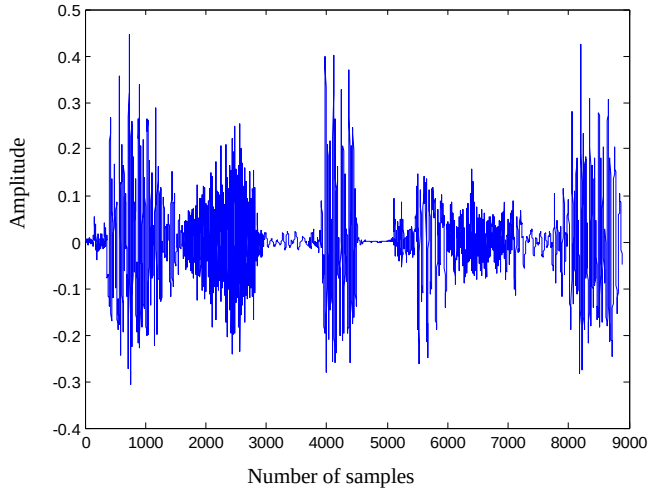

Figure .11 *Output speech for "This book is good"*

Figure .10 and.11 show the output speech for sentence. In this part, output speech is integelliable and natural .But there is a little glitch in speech output.

5. CONCLUSION

In this paper, text to speech system is developed for numbers, words and sentence. For numbers (one or more digits), the output speech is natural and pleasant to listen. It is necessary to remove delay in speech when speech waveforms are done concatenation. So, domain specific synthesis can smoothly produce speech .But the output speech of words are discontinuities between transitions of phoneme. For two or more syllable words, syllabification method is more appropriate. The sound quality is intelligibility. Thus, domain specific and phoneme based synthesis are very easy and efficient to implement unlike other methods which involve many complex algorithms. But in unit selection synthesis, the implementation is not easy as these two methods. The output sentence of speech has little glitch. However, the speech output is better than phoneme based synthesis.

ACKNOWLEDGMENTS

The author would like to thank to Dr. Hla Myo Tun, Associate Professor and Head of Department of Electronic Engineering, Mandalay Technological University for his help. And thanks to the supervisor, DawTheingiZin, AssistantLecturer, Department of Electronic Engineering, Mandalay Technological University for her guidance, support and encouragement.

REFERENCES

[1] D.Sasirekhaand E.Chandra, "Text to speech: A simple tutorial", ISSN: 2231-2307, Volume-2, Issue-1, March 2012.

[2] Prof. Dr. Hilal M. Yousif, Dr. Mouyad A. Fadhiland Yahya M. Hadi, "Text-to-Speech Synthesis State-Of-Art", 2004.

[3] Sumeer Mittal , MrNavdeep Singh Sethi and Sanjeev KumarSharma,"Partof Speech Tagging of Punjabi Language using N Gram Model",International Journal of Computer Applications (0975 – 8887) ,Volume 100– No.19, August 2014.

[4] Ing. Milan Legat, "Concatenation Cost in Unit Selection Speech synthesis",Faculty of Applied Science, University of West Bohemia in Pilsen ,2012.

[5] N.Swetha and K. Anuradha, "Text-to-speech conversion" ,International Journal of Advanced Trends in Computer Science and Engineering, Vol.2, No.6, Pages: 269-278(2013).