Text-to-Speech Synthesis for Myanmar Language

Ei Phyu Phyu Soe, eiphyu2soe@gmail.com
Aye Thida, ayethida.royal@gmail.com

Abstract— Text-to-speech (TTS) synthesis for various languages has been discussed in the Natural Language Processing research (NLP) field. In this paper, text-to-speech synthesis for Myanmar language is presented. TTS system can be divided into two key phases such as high-level and low-level synthesis. In high-level synthesis, the input text is converted into such form that the low-level synthesizer can produce the output speech. TTS synthesis for Myanmar language consists of four components such as text analysis, phonetic analysis, prosodic analysis and speech synthesis. Syllable segmentation and number converter for Myanmar language are analyzed in the text analysis. The algorithm for number convertor for Myanmar language is proposed in this paper. In the phonetic analysis, the Myanmar syllable text is converted to phonetic sequence to analyze the best sequence of phonemes of words, numbers and symbols by applying Myanmar phonetic dictionary. In prosodic analysis, some of different Myanmar words pronounce the same pronunciation so the phonetic sequences are analyzed to produce the naturalness of synthetic speech and voice duration by applying Myanmar phonological rules. Myanmar diphone dictionary is constructed for speech synthesis and PSOLA algorithm is proposed in this paper.

Index Terms— Text-to-Speech, Syllable Segmentation, Text Analysis, Phonetic Analysis, Myanmar Phonological Rules, Prosodic Analysis, Speech Synthesis

1 INTRODUCTION

Text-To-Speech technology gives computers the ability of converting text into audible speech, with the goal of being able to deliver information via voice message. It has been utilized to provide easier means of communication and to improve accessibility for people with visual impairment to textual information. Two quality criteria are proposed for deciding the quality of a TTS synthesizer. Intelligibility – it refers to how easily the output can be understood. Naturalness – it refers to how much the output sounds like the speech of a real person. Most of the existing systems have reached a fairly satisfactory level for intelligibility, while significantly less success has been attained in producing highly natural speech [1].

The goal of text-to-speech synthesis (TTS) is the automatic conversion of unrestricted natural language sentences in text form to a spoken form that closely resembles the spoken form of the same text by a native speaker of the language. This field of speech research has witnessed significant advances over the past decade with many systems being able to generate a close to natural sounding synthetic speech. Research in the area of speech synthesis has been fueled by the growing importance of many new applications. These include information retrieval services over telephone such as banking services, public announcements at places like train stations and reading out manuscripts for collation. The purpose of this paper is to develop Myanmar Text-To-Speech system and to improve the performance of high quality synthesis by applying the diphone-concatenation speech synthesis.

2 MYANMAR LANGUAGE

The Myanmar language is the official language of Myanmar and is more than one thousand years old. Texts in the Myanmar language use the Myanmar script, which is descended from the Brahmi script of ancient South India. Other Southeast Asian descendants, known as Brahmic or Indic scripts, include Thai, Khmer and Lao. Myanmar writing is different from other language because its writing is not used white spaces between words or between syllables. Thus, the computer has to determine syllable and word boundaries by means of an algorithm such as finite-state and rule-based. Moreover, a Myanmar syllable can be composed of multiple characters. Syllable segmentation is the process of determining word boundaries in a piece of text.

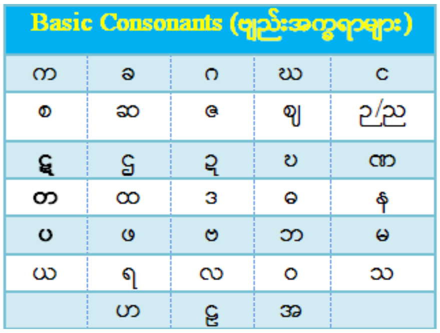
Myanmar language consists of one or more morphemes that are linked more or less tightly together. Typically, a word consists of a root or stem and zero or more affixes. Words can be combined to form phrases, clauses and sentences. A word consisting of two or more stems joined together is known as a compound word. To process text computationally, words have to be determined first [5].

2.1 Basic Consonants

A Myanmar text is a string of characters without explicit word boundary markup, written in sequence from left to right without regular inter-word spacing, although inter-phrase spacing may sometimes be used. Myanmar characters can be classified into three groups: consonants, medials and vowels. The basic consonants in Myanmar can be multiplied by medials. Syllables or words are formed by consonants combining with vowels. However, some syllables can be formed by just consonants, without any vowel. Other characters in the Myanmar script include special characters, numerals, punctuation marks and signs.

There are 34 basic consonants in the Myanmar script, as displayed in Table.1. They are known as "Byee" in the Myanmar language [5]. Consonants serve as the base characters of Myanmar words, and are similar in pronunciation to other Southeast Asian scripts such as Thai, Lao and Khmer.

TABLE 1
BASIC CONSONANTS

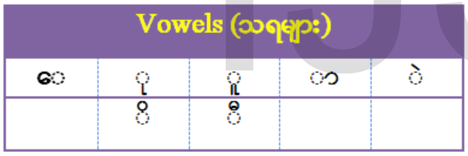


Basic Consonants (ဗျည်းအက္ခရာများ)				
က	ခ	ဂ	ဃ	င
စ	ဆ	ဇ	ဈ	ဉ/ည
ဋ	ဌ	ဍ	ဎ	ဏ
တ	ထ	ဒ	ဓ	န
ပ	ဖ	ဗ	ဘ	မ
ယ	ရ	လ	ဝ	သ
	ဟ	ဠ	အ	

2.2 Vowels

Vowels are known as “Thara”. Vowels are the basic building blocks of syllable formation in the Myanmar language, although a syllable or a word can be formed from just consonants, without a vowel as shown in Table 2. Like other languages, multiple vowel characters can exist in a single syllable.

TABLE 2
VOWLES

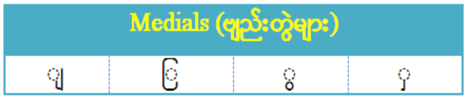


Vowels (သရများ)				
ေ◌	◌ု	◌ူ	◌ါ	◌ဲ
	◌ိ	◌ီ		

2.3 Medials

Medials are known as “Byee Twe” in Myanmar. There are 4 basic medials and 6 combined medials in the Myanmar script as shown in Table 3. The 10 medials can modify the 34 basic consonants to form 340 additional multi-clustered consonants. Therefore, a total of 374 consonants exist in the Myanmar script, although some consonants have the same pronunciation.

TABLE 3
MEDIALS

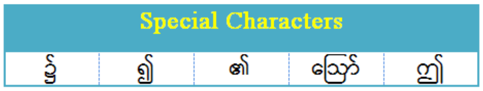


Medials (ဗျည်းတွဲများ)			
◌ျ	◌ြ	◌ွ	◌ှ

2.4 Special Characters

Special characters for Myanmar language are used as prescription noun and conjunctions words between two or more sentences.

TABLE 4
SPECIAL CHARACTERS

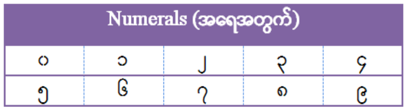


Special Characters				
၌	၍	၏	ဪ	ဤ

2.5 Numerals

Numerals for Myanmar language are known as “Counting Numbers”. Numerals are the 10 basic digits for counting.

TABLE 5
NUMERALS



Numerals (အရေအတွက်)				
၀	၁	၂	၃	၄
၅	၆	၇	၈	၉

3 NATURAL LANGUAGE PROCESSING

TTS system for Arabic langue is implemented which is based on diphone concatenation with Time Domain Pitch Synchronous Overlap-Add (TD-PSOLA) modifier synthesizer. TD-PSOLA method is based on the decomposition of the signal into overlapping frames synchronized with the pitch period. Standatrd Arabic has 34 basic phonemes, of which sex vowels and 28 are consonants and the position of the phoneme in the syllables as initial, closing, intervocalic, of suffix The main objective is to preserve the consistency and accuracy of the pitch marks after prosodic modifications of the speech signal and diphone with vowel integrated database adjustment and optimization [3].

The European Portuguese text-to-speech synthesis used the two spoken corpora. This is used the unit concatenative synthesis by applying automatic segmentation and aligiment of EUROM. For a rapid implementation and evaluation of a European Portuguese diphone-based concatenative synthesizer, the formant synthesizer module of the DIXI system is replaced by own implementation of a basic TD-PSOLA synthesis module [4].

Standard Malay TTS system is used a rule-based text- to-speech. The proposed system using sinusoidal method and some pre- recorded wave files in generating speech for the system. The use of phone database significantly decreases the amount of computer memory space used, thus making the system very light and embeddable. The overall system was comprised of two phases the Natural Language Processing (NLP) that consisted of the high-level processing of text analysis, phonetic analysis, text normalization and morphophonemic module. The module was designed specially for SM to overcome few problems in defining the rules for SM orthography system before it can be passed to the DSP module. The second phase is the Digital Signal Processing (DSP) which operated on the low-level process of the speech waveform generation. A developed an intelligible and adequately natural sounding formant-based speech synthesis system with a light and user-friendly Graphical User Interface (GUI) is introduced. A Standard Malay Language (SM) phoneme set and an inclusive set of phone database have been constructed carefully for this phone-based speech synthesizer. By applying the generative phonology, a comprehensive letter-to-sound (LTS) rules and a pronunciation lexicon have been invented for SMaTTS. As for the evaluation tests, a set of Diagnostic Rhyme Test (DRT) word list was compiled and several experiments have been performed to evaluate the quality of the synthesized speech by analyzing the Mean Opinion Score (MOS) obtained. The overall performance of the system as well as the room for improvements was thoroughly discussed [5].

3.1 Machine Translation

The term 'machine translation' (MT) refers to computerized systems responsible for the production of translations with or without human assistance. It excludes computer-based translation tools which support translators by providing access to on-line dictionaries, remote terminology databanks, transmission and reception of texts, etc. The boundaries between machine-aided human translation (MAHT) and human-aided machine translation (HAMT) are often uncertain and the term computer-aided translation (CAT) can cover both, but the central core of MT itself is the automation of the full translation process [6].

The source language model includes Part-of-Speech (POS) tagging and finding grammatical relations in Myanmar to English machine translation. The translation model includes phrase extraction, translation by using bilingual Myanmar to English corpus. The translation model also interacts with word secse disambiguation to solve ambiguities when a phrase has with more than one sense. The target languge model includes reording the translated English sentence and smoothing it by reducing grammar errors. The main goal is to construct Myanmar-English mword-aligned parallel corpus. Alignment model is central components of any statical machine translation system. The result corpus will be used in most parts of the Myanmar-English machine translation [7].

3.2 Text to Speech Synthesis

Text-to-speech (TTS) is the production of speech by machines, by way of automatic phonetization of the sentence to utter. There are two main modules in the TTS synthesizer, namely Natural Language Processing (NLP) and Digital Signal Processing (DSP). TTS synthesizer produces speech based on the corresponding text input. It has been utilized to provide easier means of communication and to improve accessibility for people with visual impairment to textual information. This language dependent system has been widely developed for various languages with continuous research to improve the quality of the produced speech. Natural language processing module is responsible for conversion of text input into phonetic transcription and prosody information. Prosody information, which includes melody (intonation) and rhythm, is necessary to make the resulting speech sounds natural (not flat/robot-like). The DSP module then transforms the resulting phonetic transcription and prosody information into corresponding speech [1].

The text-analysis module of the multilingual Bell Labs TTS system has been developed for Spanish, Italian, Romanian, French, German, Russian, Mandarin and Japanese languages [2]. They discussed the transducers are constructed using a lexical toolkit that allows declarative descriptions of lexicons, morphological rules, numeral expansion rules, and phonological rules, inter alia.

4 Text to Speech Design for Myanmar Language

The two main parts of TTS for Myanmar language are proposed in this paper. First one is a natural language processing (NLP) which reads the input text and translates it into a phonetic language and second is digital signal processing (DSP) that converts the phonetic language into spoken speech. There are two modules in NLP part such as text analysis and phonetic analysis. DSP has two modules, prosodic analysis and diphone-concatenation speech synthesis. Figure 1 shows TTS design for Myanmar language.

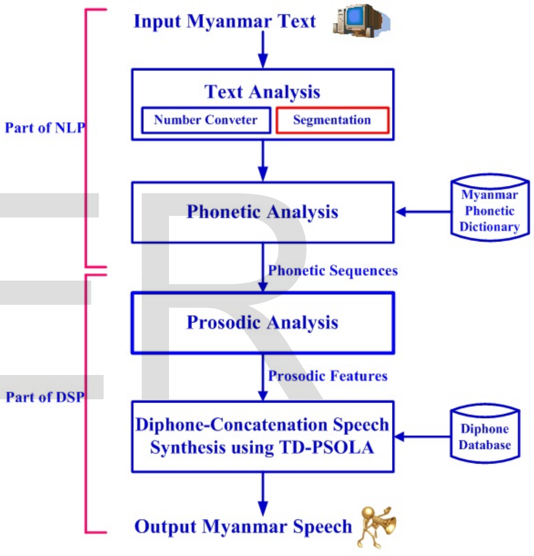
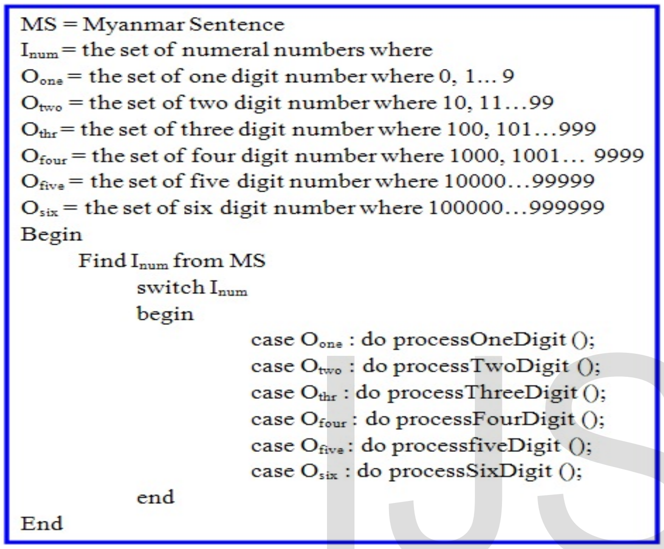


Fig.1. Text-To-Speech Design for Myanmar Langugae

5 Text Analysis

Text analysis takes input in the form of text and outputs a symbolic linguistic representation. The input text is analyzed to segmented Myanmar text like a syllable. Syllable segmentation and Number converter are analyzed in this module. Syllable segmentation is the process of identifying syllable boundaries in a text. Segmentation [16] is used in this paper to get the segmented Myanmar syllable. Algorithm for Myanmar number converter is proposed in this paper as shown in figure 2 which can convert the Myanmar number to textual versions. The non standard words are tokens like numbers, which need to be expanded into sequences of Myanmar words before they are pronounced.

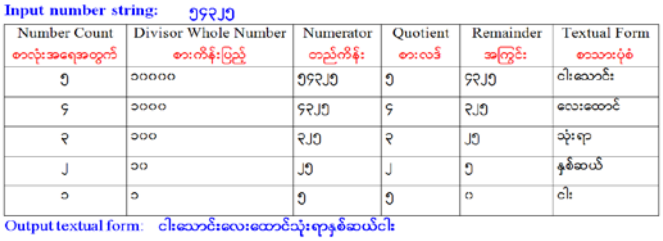
Number converter for Myanmar number is a difficult and vital task in TTS synthesis system. Myanmar number is not defined as a number like English because it has a Unicode for each number. The computer does not easily understand the Myanmar number like English number. So, the system must translate these to machine-language such as Myanmar number 0 to 9 has a Unicode for each. If a number string has 4 words, system will check with their each Unicode. So, number converter control a number string with a whole number Unicode according to the counting of number string. The number convertor algorithm is divided the input number with whole number such as 1000, 100, 10 and it marks the positions of the quotients. But this number converter cannot be transferred the decimal Myanmar number. It can be only changed the simpleMyanmar number for million counting. The transforming of Myanmar number to textual form is shown in Figure.3.



```
MS = Myanmar Sentence
$I_{num}$ = the set of numeral numbers where
$O_{one}$ = the set of one digit number where 0, 1… 9
$O_{two}$ = the set of two digit number where 10, 11…99
$O_{thr}$ = the set of three digit number where 100, 101…999
$O_{four}$ = the set of four digit number where 1000, 1001… 9999
$O_{five}$ = the set of five digit number where 10000…99999
$O_{six}$ = the set of six digit number where 100000…999999
Begin
    Find $I_{num}$ from MS
        switch $I_{num}$
        begin
                case $O_{one}$ : do processOneDigit ();
                case $O_{two}$ : do processTwoDigit ();
                case $O_{thr}$ : do processThreeDigit ();
                case $O_{four}$ : do processFourDigit ();
                case $O_{five}$ : do processfiveDigit ();
                case $O_{six}$ : do processSixDigit ();
        end
End
```

Fig. 2. Algorithm for Myanmar Numbers Converter

Fig. 3. Sample of Conversion for Myanmar Digit to Myanmar Text



Input number string: ၅၄၃၂၅

Number Count စာလုံးအရေအတွက်	Divisor Whole Number စားကိန်းပြည့်	Numerator တည်ကိန်း	Quotient စားလဒ်	Remainder အကြွင်း	Textual Form စာသားပုံစံ
၅	၁၀၀၀၀	၅၄၃၂၅	၅	၄၃၂၅	ငါးသောင်း
၄	၁၀၀၀	၄၃၂၅	၄	၃၂၅	လေးထောင်
၃	၁၀၀	၃၂၅	၃	၂၅	သုံးရာ
၂	၁၀	၂၅	၂	၅	နှစ်ဆယ်
၁	၁	၅	၅	၀	ငါး

Output textual form: ငါးသောင်းလေးထောင်သုံးရာနှစ်ဆယ်ငါး

6 PHONETIC ANALYSIS

Phonetic analysis is also called Grapheme-to-Phoneme (G2P) conversion which translates the syllable of Myanmar text to phonetic sequence. It determines the pronunciation of a syllable based on its spelling. It also analyzes the best sequence of phonemes for words, numbers and symbols and converts into phonetic sequences. The Myanmar phonetic dictionary is constructed to generate the phoneme sequence and to pronounce these phonemes.

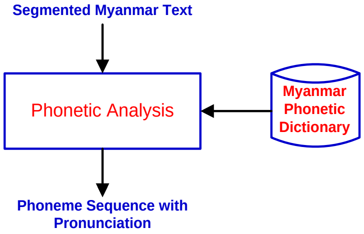

Segmented Myanmar Text
Phonetic Analysis
Myanmar Phonetic Dictionary
Phoneme Sequence with Pronunciation

Fig.4. Overview of Phonetic Analysis

6.1 Phonological Structure of Myanmar Langauge

The Myanmar language uses a rather large set of 50 vowel phonemes, including diphthongs, although its 22 to 26 consonants are close to average. Some languages, such as French, have no phonemic tone or stress, while several of the Kam-Sui languages have nine tones, and one of the Kru languages, Wobe, has been claimed to have 14, though this is disputed. The most common vowel system consists of the five vowels /i/, /e/, /a/, /o/, /u/. The most common consonants are /p/, /t/, /k/, /m/, /n/. Relatively few languages lack any of these, although it does happen: for example, Arbic lacks /p/, standard Hawaiian lacks /t/, Mohawk and Tlingit lack /p/ and /m/, Hupa lacks both /p/ and a simple /k/, colloquial Samoan lacks /t/ and /n/, while Rotokas and Quileute lack /m/ and /n/ [11]. Table4 shows the phonetic signs of 50 Myanmar vowels to pronounce the Myanmar words. These 50 phonemes show the basic symbol with four tone levels [8], [10].

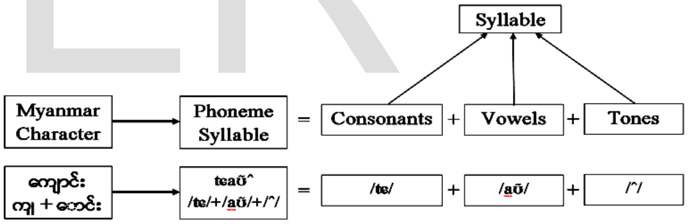


Fig.5. Combination of Phoneme Syllable

Phonology is how speech sounds are organized and affect one another in pronunciation. The combination of consonant phoneme and a vowel phoneme produces a syllable in figure2. The phonetic alphabet is usually divided in two main categories, vowels and consonants. Vowels are always voiced sounds and they are produced with the vocal cords in vibration, while consonants may be either voiced or unvoiced. Vowels have considerably higher amplitude than consonants and they are also more stable and easier to analyze and describe acoustically. Because consonants involve very rapid changes they are more difficult to synthesize properly [9].

6.2 Myanmar Phonetic Dictionary

The purpose of this paper presents about the dictionary-based approach for Grapheme-to-Phoneme (GTP) conversion. There are five steps to construct the Myanmar phonetic dictionary:

[1]. Collecting of possibility words

[2]. Separating of consonants and vowels of Myanmar words

[3]. Storing consonants and vowels of phonetic signs

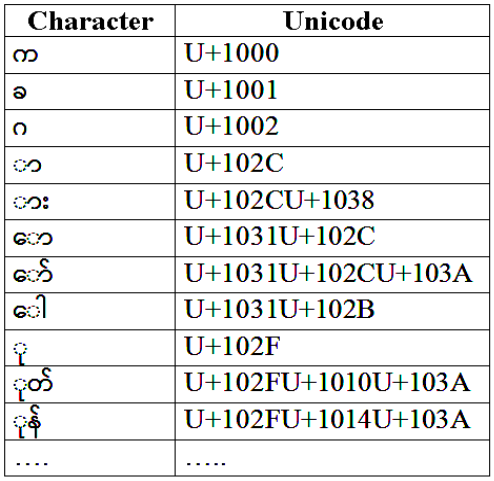
[4]. Recording of Myanmar syllables and

[5]. Segmenting of recorded speech.

In the works of possibility words collection, we find the Myanmar words by looking at Myanmar dictionary book. Myanmar word is collated based on syllables. A Myanmar syllable encoded in Unicode can be broken into 5 parts for collation: <consonant> <medial> <vowel> <final> <tone>. Only the consonant is always present, one or more of the other parts may be empty in any given syllable. In practice the vowel may be displayed before the consonant e.g. □□□□, but it is encoded as U+1000 (Myanmar letter □) U+1031U+102CU+103A (Myanmar vowel □□□).The resulting collation sequence has 5 levels, order of priority: <consonant>, <medial>, <final>, <vowel>, <tone>. Note, that the final and vowel have been switched from their encoded order. Each of these parts of the syllable may be composed of one or more characters. Collation Order read left to right and then down. The data is presented in the traditional layout of the Myanmar alphabet [6].

The step of separating of consonants and vowels is the important step in Grapheme-to-Phoneme conversion. Storing the separated consonants and separated vowels to database decrease the complexity of searching times because the total compound Myanmar syllables have about near 2000 reduce to the total consonants and vowels have about near 500. To become a Myanmar syllable is combining the possible compound vowels and consonants. So this system is not inserting the whole Myanmar syllable because one consonant has near 60 compound vowels. If the system saved 33 consonants x 60 vowels = 1980 syllables to the database, the searching time and complexity will increase. The advantage of dictionary-based approach is quick and accurate, but completely fails if it is given a word which is not in its dictionary. As dictionary size grows, so too does the memory space requirements of the synthesis system. Table 6 describes about the separated consonants and vowels of Myanmar syllables.

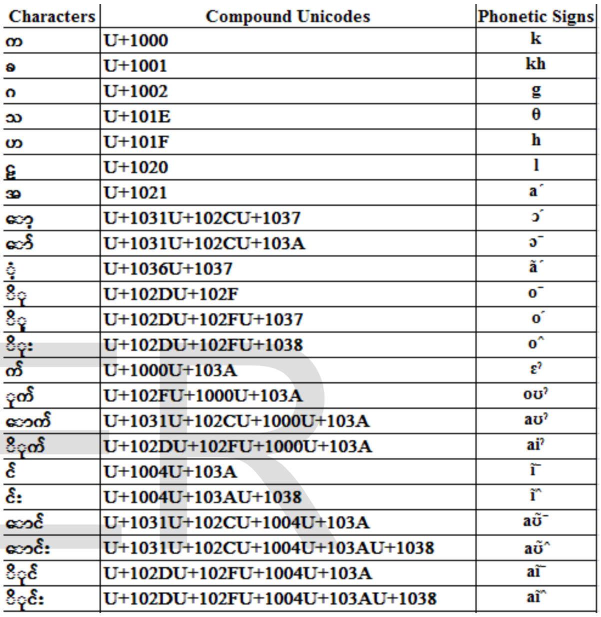
There is no problem in inserting the phonetic signs for consonants but single vowel and compound vowels have a little problem to match the pronounced of the whole syllable. In this step, we must insert the data carefully to get the correct phoneme sequences and pronounced of Myanmar syllables because of producing high quality speech synthesis depend on the performance of Grapheme-to-Phoneme conversion step. The phonetic signs are important for Myanmar language analysis and it promotes for the whole system of Grapheme-to-Phoneme (GTP) conversion [8], [10]. Table 7 shows the phonetic signs for Myanmar characters to produce the phonetic sequences.



Character	Unicode
က	U+1000
ခ	U+1001
ဂ	U+1002
ာ	U+102C
ား	U+102CU+1038
ော	U+1031U+102C
ော်	U+1031U+102CU+103A
ေါ	U+1031U+102B
ု	U+102F
ုတ်	U+102FU+1010U+103A
ုန်	U+102FU+1014U+103A
….	…..

TABLE 6
SEPARATED CONSONANTS AND VOWELS OF MYANMAR WORDS

TABLE 7
PHONETIC SIGNS OF MYANMAR CHARACTERS

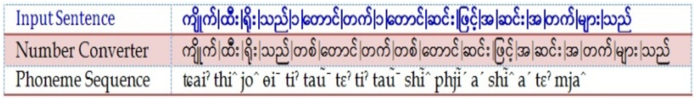


Characters	Compound Unicodes	Phonetic Signs
က	U+1000	k
ခ	U+1001	kh
ဂ	U+1002	g
သ	U+101E	θ
ဟ	U+101F	h
ဠ	U+1020	l
အ	U+1021	a´
ော့	U+1031U+102CU+1037	ɔ´
ော်	U+1031U+102CU+103A	ɔ¯
ံ့	U+1036U+1037	ã´
ို	U+102DU+102F	o¯
ို့	U+102DU+102FU+1037	o´
ိုး	U+102DU+102FU+1038	o^
က်	U+1000U+103A	ɛˀ
ုက်	U+102FU+1000U+103A	oʊˀ
ောက်	U+1031U+102CU+1000U+103A	aʊˀ
ိုက်	U+102DU+102FU+1000U+103A	aiˀ
င်	U+1004U+103A	ĩ¯
င်း	U+1004U+103AU+1038	ĩ^
ောင်	U+1031U+102CU+1004U+103A	aʊ̃¯
ောင်း	U+1031U+102CU+1004U+103AU+1038	aʊ̃^
ိုင်	U+102DU+102FU+1004U+103A	aĩ¯
ိုင်း	U+102DU+102FU+1004U+103AU+1038	aĩ^

After storing the phonetic signs to database, this system stores the segmented recorded speech according to their compound Unicodes. Firstly, we record the phone level of the whole syllable. Concerning to constituent phones and syllabic neighboring context, syllable is designed in the form of onset-nucleus-coda. Onset and coda represent single consonants. Nucleus covers short and long vowels, and, short and long diphthongs. The example of grapheme-to-phoneme conversion is shown in following table 8.

This data contains fluently read speech recorded by a Myanmar female student thus this reading style is general reading style. In syllable level data set, it contains approximately 2000 syllables. In phone-level data set, it contains approximately 4600 phones. The example of grapheme-to-phoneme conversion is shown in below:

TABLE 8
EXAMPLE OF GRAPHEME-TO-PHONEME CONVERSION



Input Sentence	ကျိုက်|ထီး|ရိုး|သည်|၁|တောင်|တက်|၁|တောင်|ဆင်း|ဖြင့်|အ|ဆင်း|အ|တက်|များ|သည်
Number Converter	ကျိုက်|ထီး|ရိုး|သည်|တစ်|တောင်|တက်|တစ်|တောင်|ဆင်း|ဖြင့်|အ|ဆင်း|အ|တက်|များ|သည်
Phoneme Sequence	tɕaiˀ thiˆ joˆ θiˉ tiˀ taũˉ tɛˀ tiˀ taũˉ shĩˆ phjĩ´ a´ shĩˆ a´ tɛˀ mjaˆ

7 Prosodic Analysis

In this step, the phonetic sequences are analyzed to produce the prosodic features by applying the phonological rules. It is the module to analyze duration and intonation such as pitch variation, syllable length to create naturalness of synthetic speech. The combination of consonant phoneme and a vowel phoneme produces a syllable. The phonetic alphabet is usually divided in two main categories, vowels and consonants. Vowels are always voiced sounds and they are produced with the vocal cords in vibration, while consonants may be either voiced or unvoiced. Vowels have considerably higher amplitude than consonants and they are also more stable and easier to analyze and describe acoustically. Because consonants involve very rapid changes they are more difficult to synthesize properly.

There are five tasks of phonological rules in the process of prosodic analysis for Myanmar Language as shown in overview of the system design. This system extracts the correct pronunciations by applying the rule-based phonology.Phonemic forms change to the phonetic forms by applying the phonological rules. Phonological rules link the two levels of underlying and surface of the phoneme. It describes how phonemes are realized as their allophones in the given environment. Environment in phonology typically refers to the neighboring phonemes.

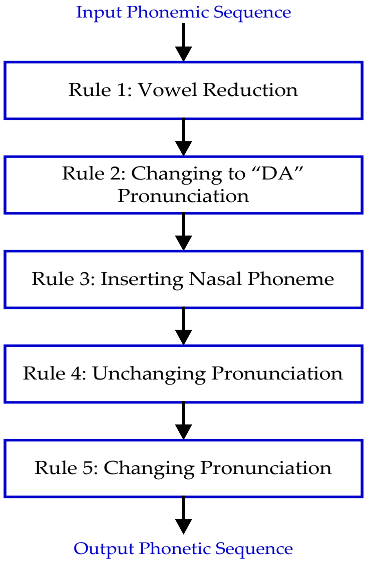

Input Phonemic Sequence
Rule 1: Vowel Reduction
Rule 2: Changing to "DA" Pronunciation
Rule 3: Inserting Nasal Phoneme
Rule 4: Unchanging Pronunciation
Rule 5: Changing Pronunciation
Output Phonetic Sequence

Fig.6. System Flow Diagrams for Five Rules

This overview system presents about five phonological rules with the summary processes in Figure6. The process of Rule 1 is the vowel reduction to reduce the tense vowel with n-gram algorithm. Rule 2 shows about the process of metathesis for counting number /tiˀ/ to /də/ pronuncia pronunciation by applying the bigram model. The manner of Rule 3 is unchanging pronunciation phoneme next to the /a´/ □ syllable until it is an unvoiced phoneme by using forward algorithm. Rule 4 is a difficult process for unvoiced phoneme to voiced phoneme depending on the defined voiced vowel, consonants and asats by applying forward backward algorithm. Rule 5 is an inserting nasal phoneme process to different types of obstruent by employing digram [12].

Phonological rules can be roughly divided into six types [12], [14]:

1. Assimilation
2. Dissimilation
3. Reduction
4. Metathesis
5. Insertion and
6. Deletion

Assimilation: When a sound changes one of its features to be more similar to an adjacent sound. This is the kind of rule that occurs in the Myanmar Unvoiced rule described as—the Myanmar syllable "□" becomes voiced or voiceless depending on whether or not the preceding consonant is voiced.

Dissimilation: When a sound changes one of its features to become *less* similar to an adjacent sound, usually to make the two sounds more distinguishable. This type of rule is often seen among people speaking a language that is not their native language, where the sound contrasts may be difficult.

Reduction: When a sound has high vowel phoneme such as /i,u/ and low vowel phoneme such as /a/ with glottalized/ˀ/ or nasal/˜/ tones, high vowel phonemes change to one level of low vowel phonemes such as [ɪ,ʊ] and low vowel phoneme changes to one level of high vowel phoneme such as [ʌ].

Metathesis: When a sound has a Myanmar countable number 'one' □□□ /tiˀ/, it changes to schwa [də] depending on the next consonant is unvoiced syllable.

Insertion: When an extra sound is added between two others. This also occurs in the Myanmar Nasal rule: when the nasal consonant *-m* is added to "bilabial obstruent".

Deletion: When a sound, such as a stress less syllable or a weak consonant, is not pronounced; for example, most American English speakers do not pronounce the [d] in "handbag". There is no syllable or a weak consonant in Myanmar language.

In this paper, we proposed the phonological rules based on five types for Myanmar language pronunciations except deletion.

8 Speech Synthesis

The aim of speech synthesis is to be able to take a word sequence and produce "human-like" speech. Linguistic analysis stage maps the input text into a standard form and determines the structure of the input, and finally decides how to pronounce it. Synthesis stage converts the symbolic representation of what to say into an actual speech waveform [13].

Speech communication relies not only on audition, but also on visual information. Facial movements, such as smiling, grinning, eye blinking, head nodding, and eyebrow rising give additional information of the speaker's emotional state. The emotional state may be even concluded from facial expression without any sound. Fluent speech is also emphasized and punctuated by facial expressions. With visual information added to synthesized speech it is also possible to increase the intelligibility significantly, especially when the auditory speech is degraded by for example noise, bandwidth filtering, or hearing impairment. The visual information is especially helpful with front phonemes whose articulation we can see, such as labiodentals and bilabials (Beskow et al. 1997). For example, intelligibility between /b/ and /d/ increases significantly with visual information (Santen et al. 1997). Synthetic face also increases the intelligibility with natural speech. However, the facial gestures and speech must be coherent. Without coherence the intelligibility of speech may be even decreased.

8.1 Text-To-Speech Synthesis

Text-to-speech (TTS) synthesizer produces speech based on the corresponding text input. It has been utilized to provide easier means of communication and to improve accessibility for people with visual impairment to textual information. This language dependent system has been widely developed for various languages with continuous research to improve the quality of the produced speech.

A TTS voice is a computer program that has two major parts: a natural language processing (NLP) which reads the input text and translates it into a phonetic language and a digital signal processing (DSP) that converts the phonetic language into spoken speech. The input text might be for example data from a word processor, standard ASCII from e-mail, a mobile text-message, or scanned text from a newspaper. The character string is then preprocessed and analyzed into phonetic representation which is usually a string of phonemes with some additional information for linguistic representation.

8.2 Diphone-Concatenative Speech Synthesis

The process of concatenative speech synthesis is cutting and pasting the short segments of speech is selected from a prerecorded database and joined one after another to pr duce the desired utterances. In theory, the use of real speech as the basis of synthetic speech brings about the potential for very high quality, but in practice there are serious limitations, mainly due to the memory capacity required by such a system. The longer the selected units are, the fewer problematic concatenation points will occur in the synthetic speech, but at the same time the memory requirements increase. Another limitation in Concatenative synthesis is the strong dependency of the output speech on the chosen database. For example, the personality or the affective tone of the speech is hardly controllable. Despite the somewhat featureless nature, Concatenative synthesis is well suited for certain limited applications [15]. Concatenative synthesis is based on the concatenation or stringing together of segments of recorded speech. Generally, Concatenative synthesis produces the most natural-sounding synthesized speech. It is easier to obtain more natural sound with longer units and it can achieve a high segmental quality. Among these techniques, this paper highlights a diphone concatenation-based synthesis technique in Myanmar text-to-speech research.

8.3 Myanmar Diphone Database Construction

The basic idea behind building Myanmar diphone databases is to explicitly list all possible phone-phone transitions in a language. One technique is to use target words embedded carrier sentences to ensure that the diphones are pronounced with acceptable duration and prosody. Speech synthesis unit finds the corresponding pre-recorded sounds from its database and tries to concatenate them smoothly. It uses an algorithm like TD-PSOLA (Time-Domain Pitch Synchronous Overlap and Add) to make a smooth pass in diphone. PSOLA method takes two speech signals. One of these signals ends with a voiced part and the other starts with a voiced part.PSOLA changes the pitch values of these two signals so that pitch values at both sides become equal. The advantage of this technique is to obtain a better output speech when compared to other techniques [1].

The structure of diphone database constructs with Arpabet signs to understand the retrieving phonemes. After retrieving the phonemes, we can then retrieve each individual phoneme from a diphone database and concatenate them together with only 50 phonemes; this would be the most economical choice to save space on embedded devices. Diphones are just pairs of partial phonemes. This might be recovered from the pronouncing dictionary by taking into account the 1 or 0 designation applied to vowels concerning stress instead of representing a single phoneme; a diphone represents the end of one phoneme and the beginning of another. This is significant because there is less difference in the middle of a phoneme than there is at the beginning and ending edges. The problem is that it greatly increases the size of the diphone database from around 10496 diphones (114 (22Consonants + 42ExceptionWords + 50Vowels) x 114 (22Consonants + 42ExceptionWords + 50Vowels) – 2500 (50Vowels x50Vowles)) in Myanmar Language. The pair of vowel and vowel is not in phoneme sequence for diphone database. So the number of double vowels subtracts from the total diphone database.

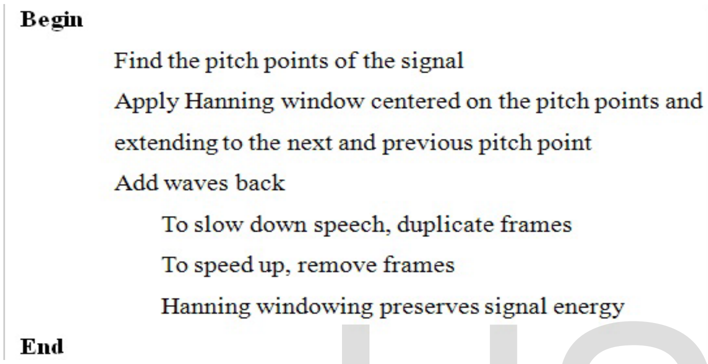
The diphone list will be categorized in different categories [11]: Consonants-Consonants, Consonants-Exception Words, Consonants-Vowels, Exception Words-Consonants, Exception Words-Exception Words, Exception Words Vowels, Vowels-Consonants, Vowels-Exception Words, Consonants-Silence, Exception Words-Silence, and Vowels-Silence, Silence-Consonants, Silence-Exception Words and Silence-Vowels pairs.

8.4 TD-PSOLA Method for Myanmar Language

France Telecom (CNET) develops Pitch Synchronous Overlap and Add method. It allows prerecorded speech samples smoothly concatenated and provides good controlling for pitch and duration. Time-domain version, TD-PSOLA, is the most commonly used due to its computational efficiency. The basic algorithm consists of three steps:

1. original speech signal is divided into separate short analysis signal
2. the modification of each analysis signal to synthesis signal and
3. the synthesis step where these segments are recombined by means of overlap-adding [15].

The purpose of TD-PSOLA (Time-Domain) is to modify the pitch or timing of a signal as shown in figure7. The process of the TD-PSOLA algorithm is to find the pitch points of the signal and then apply the hamming window centered of the pitch points and extending to the next and previous pitch point. If the speeches want to slow down, the system defines the frame to double. If the speeches want to speed up, the system removes the frames in the signal.

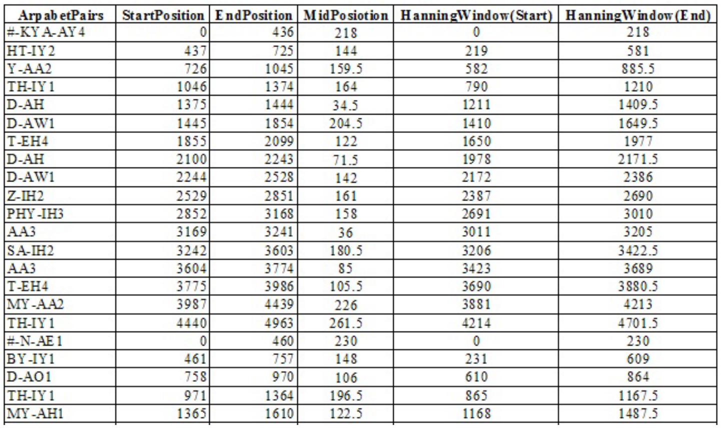


```
Begin
        Find the pitch points of the signal
        Apply Hanning window centered on the pitch points and
        extending to the next and previous pitch point
        Add waves back
            To slow down speech, duplicate frames
            To speed up, remove frames
            Hanning windowing preserves signal energy
End
```

Fig.7. TD-PSOLA Algorithm

TD-PSOLA requires an exact marking of pitch points in a time domain signal. Pitch marking any part within a pitch period is okay as long as the algorithm marks the same point for every frame. The most common marking point is the instant of glottal closure, which identifies a quick time domain descent. The algorithm creates an array of sample numbers comprise an analysis epoch sequence P = {p_1, p_2… p_n} and it estimates pitch period distance = (p_k - p_{k+1})/2 to get the midpoint of pitch marking as shown in Table 9.

TABLE 9
DEFINING THE PITCH MARKS



ArpabetPairs	StartPosition	EndPosition	MidPosiotion	HanningWindow(Start)	HanningWindow(End)
#-KYA-AY4	0	436	218	0	218
HT-IY2	437	725	144	219	581
Y-AA2	726	1045	159.5	582	885.5
TH-IY1	1046	1374	164	790	1210
D-AH	1375	1444	34.5	1211	1409.5
D-AW1	1445	1854	204.5	1410	1649.5
T-EH4	1855	2099	122	1650	1977
D-AH	2100	2243	71.5	1978	2171.5
D-AW1	2244	2528	142	2172	2386
Z-IH2	2529	2851	161	2387	2690
PHY-IH3	2852	3168	158	2691	3010
AA3	3169	3241	36	3011	3205
SA-IH2	3242	3603	180.5	3206	3422.5
AA3	3604	3774	85	3423	3689
T-EH4	3775	3986	105.5	3690	3880.5
MY-AA2	3987	4439	226	3881	4213
TH-IY1	4440	4963	261.5	4214	4701.5
#-N-AE1	0	460	230	0	230
BY-IY1	461	757	148	231	609
D-AO1	758	970	106	610	864
TH-IY1	971	1364	196.5	865	1167.5
MY-AH1	1365	1610	122.5	1168	1487.5

9. EXPERIMENTAL RESULTS

This paper gives the results for the diphone-concatenation with TD-PSOLA method. This system is tested with the 200 Myanmar sentences and this sentence structure is very complex. The Myanmar diphone database stores over 5000 diphones for these sentences. Firstly, this system accepts the segmented Myanmar sentence and then it can produce the phonetic sequence with the pairs of consonants and vowels by using Myanmar phonetic dictionary in grapheme-to-phoneme stage [4]. And then this system checks the phonetic sequence to get the prosodic features with phonological rules. Finally, it produces the high quality speech by applying the Myanmar diphone database with concatenation method that uses TD-PSOLA algorithm.

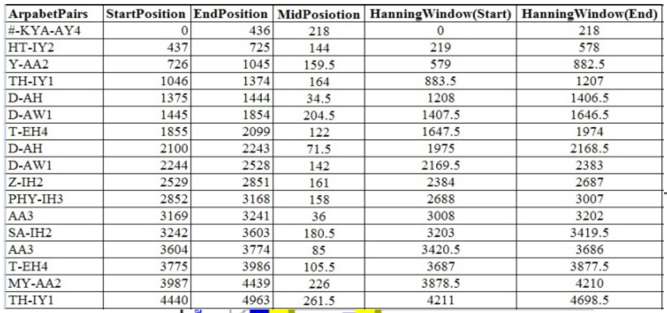
The experimental results of diphone-concatenation speech synthesis can be calculated with precision, recall and f-measure. The results for 14 types of diphone pairs according to the total number of 7837 diphone pairs for 350 sentences is shown in Table10.

TABLE 10
EXPERIMENTAL RESULTS FOR DIPHONE-CONCATENATION

Diphones-Pairs	Recall	Precision	F-Measure
C-C	99%	99%	99%
C-Ex	97.5%	97.5%	97.5%
C-V	97%	97%	97%
Ex-C	100%	100%	100%
Ex-Ex	90%	90%	90%
Ex-V	94%	94%	94%
V-C	100%	100%	100%
V-Ex	92%	92%	92%
C-S	100%	100%	100%
Ex-S	98%	98%	98%
V-S	95%	95%	95%
S-C	100%	100%	100%
S-Ex	90%	90%	90%
S-V	100%	100%	100%

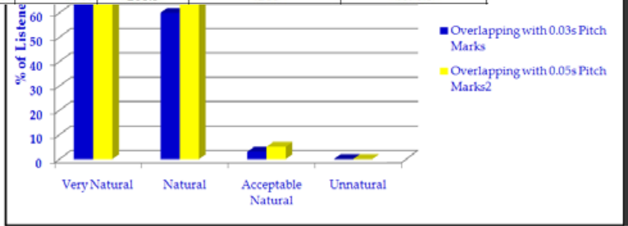
The original waveforms without TD-PSOLA method and the length is 1.206s for first 4 words, "#-KYA-AY4-HT-IY2-Y-AA2-TH-IY1". Table11 gives the data for the overlapping time according to the pitch marks with 0.03s of the waveforms of the voices. This table shows the 22 diphone pairs for 20 words Myanmar sentence. The diphone-concatenation pairs for this sentence have the 20 pairs for speech synthesis.

TABLE 11
DEFINING THE PITCH MARKS WITH 0.03S



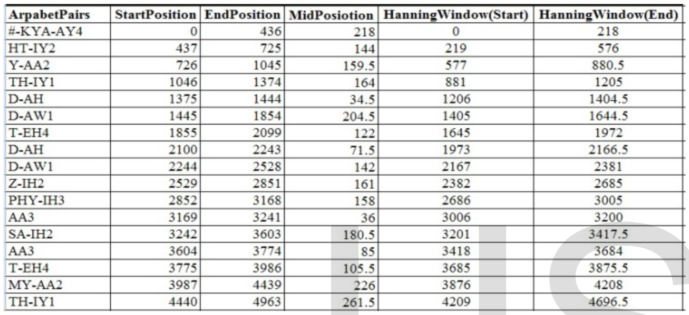
ArpabetPairs	StartPosition	EndPosition	MidPosiotion	HanningWindow(Start)	HanningWindow(End)
#-KYA-AY4	0	436	218	0	218
HT-IY2	437	725	144	219	578
Y-AA2	726	1045	159.5	579	882.5
TH-IY1	1046	1374	164	883.5	1207
D-AH	1375	1444	34.5	1208	1406.5
D-AW1	1445	1854	204.5	1407.5	1646.5
T-EH4	1855	2099	122	1647.5	1974
D-AH	2100	2243	71.5	1975	2168.5
D-AW1	2244	2528	142	2169.5	2383
Z-IH2	2529	2851	161	2384	2687
PHY-IH3	2852	3168	158	2688	3007
AA3	3169	3241	36	3008	3202
SA-IH2	3242	3603	180.5	3203	3419.5
AA3	3604	3774	85	3420.5	3686
T-EH4	3775	3986	105.5	3687	3877.5
MY-AA2	3987	4439	226	3878.5	4210
TH-IY1	4440	4963	261.5	4211	4698.5

The waveforms



smooth between one joint of waveform and another by using TD-PSOLA method for Myanmar language. The quality of speech is more speed and smooth by are overlapping each to each with 0.03s than original speech waveforms. The total length of overlapping speech waveforms is shorter than original waveforms without any method. The length of overlapping waveforms reduces 0.05s from 1.206s. The next table 12 shows the overlapping of pitch marks with 0.05s between one joint and other joints with 0.05s of waveforms of the voice for 20 words of Myanmar sentence. The hanning window calculates with the overlap of 0.05s pitch marks in each diphone label. The values of the start of hanning window and the end of hanning windows are changed according to the overlapping pitch marks values. The sound quality is better than the overlapping pitch marks 0.03ms.

TABLE 12
DEFINING THE PITCH MARKS WITH 0.05S



ArpabetPairs	StartPosition	EndPosition	MidPosiotion	HanningWindow(Start)	HanningWindow(End)
#-KYA-AY4	0	436	218	0	218
HT-IY2	437	725	144	219	576
Y-AA2	726	1045	159.5	577	880.5
TH-IY1	1046	1374	164	881	1205
D-AH	1375	1444	34.5	1206	1404.5
D-AW1	1445	1854	204.5	1405	1644.5
T-EH4	1855	2099	122	1645	1972
D-AH	2100	2243	71.5	1973	2166.5
D-AW1	2244	2528	142	2167	2381
Z-IH2	2529	2851	161	2382	2685
PHY-IH3	2852	3168	158	2686	3005
AA3	3169	3241	36	3006	3200
SA-IH2	3242	3603	180.5	3201	3417.5
AA3	3604	3774	85	3418	3684
T-EH4	3775	3986	105.5	3685	3875.5
MY-AA2	3987	4439	226	3876	4208
TH-IY1	4440	4963	261.5	4209	4696.5

The quality of speech is more speed and smooth by are overlapping each to each with 0.05s than original speech waveforms and 0.03s pitch marks overlapping. The total length of overlapping speech waveforms is shorter than original waveforms without any method. The length of overlapping waveforms reduces 0.12s from 1.206s.

9.1 TEXTING TEXT-TO-SPEECH QUALITY

Testing the naturalness and speed of the Myanmar speech contains 12 female people between the ages 16 to 40. The test can be divided into two parts with 20 pairs of words of confusability. The first part contains naturalness of the diphone-concatenative speech synthesis. The last part tests the speed of the synthesis system. The participants heard one word at a time and marked on the answering sheet which one of the two words they think is correct.

9.1.1Naturalness

The results of the overlapping with 0.03s pitch marks of listening compared to overlapping with 0.05s pitch marks of listening are shown in figure8 below. The system tested with 20 words complex sentence structure. The listeners or users are regarding the question whether the voice is nice to listen to or not, 90% considered the voice natural, 60% thought that the naturalness of the voice was acceptable and 3 % considered the voice unnatural for pitch marks 0.03s. The users regard 98% considered the voice natural, 70% thought that the naturalness of the voice was acceptable and 5 % considered the voice unnatural for 0.05s pitch marks. The results changed slightly after the second time of listening.

Fig.8. Comparison of Naturalness of the Speech Synthesis

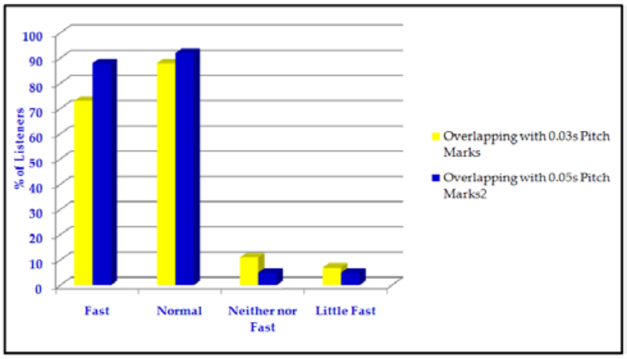
9.1.2Speed

The results of speed of the voices for 0.03s overlapping pitch marks and the speed of 0.05s pitch marks overlapping are shown in figure9. The system tested with 20 words complex sentence structure. The questions of speed for the listeners or users are asked understood or not the voice or how much of what the voice said the participants understood, 88% of the participants are normal, 73% of the participants fast. 11% neither much nor fast and another 7% understood little fast. The results of 0.05s pitch marks overlapping are 92% of the participants regard normal, 88% of the participants fast. 5% neither much nor fast and another 5% fast a little. The comparison of this speed is shown in figure9.

Fig.9. Comparison of Speed of the Speech Synthesis

10. CONSLUSION

This paper gives the results of speech quality for two types of pitch marks overlapping with TD-PSOLA method. The Myanmar diphone database stores over 7000 diphones for 350 sentences of complex structure. This system accepts the segmented Myanmar sentence and then it can produce the phonetic sequence with the pairs of consonants and vowels by using Myanmar phonetic dictionary in grapheme-to-phoneme stage. And then this system checks the phonetic sequence to get the prosodic features with phonological rules. Finally, it produces the high quality speech by applying the Myanmar diphone database with concatenation method that uses TD-PSOLA algorithm.



This paper shows the speech synthesis results with variety of time domain such as 0.03s and 0.05s pitch marks of hanning windows. The quality of speech with 0.03s is not good when compare with overlapping pitch marks 0.05s. The speed and naturalness of speech for 0.05s is better then the overlapping pitches mark 0.03s in this paper.

This system can be processed by five phonological rules for changing to unvoiced to voiced pronunciations. Emotional state of speech can be extended as one of the research of text-to-speech synthesis.

ACKNOWLEDGMENT

I wish to express my deep gratitude and sincere appreciation to all persons who contributed towards the success of my research. It was a great chance to opportunity to study Myanmar text-to-speech research in one of the most famous research in the world. I would like to respectfully thank and appreciate Dr. Mie Mie Thet Thwin, Rector of the University of Computer Studies, Mandalay (UCSM), for her precious advice, patience and encouragement during the preparation of my research. I am grateful to my supervisor, Dr. Aye Thida, an Associate Professor of Research and Development Department (1) at University of Computer Studies, Mandalay (UCSM), and Myanmar, one of the leaders of Natural Language Processing Project for having me helped during the preparation of my research. She was, I am also deeply thankful to Dr. Aye Thida, for her advising from the point of view of natural language processing. I also take this opportunity to thank all our teachers of the University of Computer Studies, Mandalay (UCSM), for their teaching and guidance during my research life.

I especially thank my parents, my sisters and all my friends for their encouragement, help, kindness, providing many useful suggestions and giving me their precious time give to me during the preparation of my research.

REFERENCES

[1] A. Nur Aziza, H.Rose Maulidiyatul, T.Teresa Vania and N.Anto Satriyo," Evaluation of Text-to-Speech Synthesizer for Indonesian Language Using Semantically Unpredictable Sentences Test: IndoTTS, eSpeak, and Google Translate TTS", Proc. of International Conference on Advanced Computer Science & Information Systems, 2011.

[2] Richard Sproa," Multilingual Text Analysis for Text-to-Speech Synthesis".

[3] Abdelkader Chabchoub and Adnan Cherif," High Quality Arabic Concatenative Speech Synthesis", Signal & Image Processing: An International Journal (SIPIJ) Vol.2, No.4, and December 2011.

[4] Pedro M. Carvalho, Luís C. Oliveira, Isabel M. Trancoso, M. Céu Viana, "CONCATENATIVE SPEECH SYNTHESIS FOR EUROPEAN PORTUGUESE",

[5] Othman O. Khalifa, Zakiah Hanim Ahmad, and Teddy Surya Gunawan, "SMaTTS: Standard Malay Text to Speech System", International Journal of Electrical and Computer Engineering 2:4 2007.

[6] W.John Hutchins, "MACHINE TRANSLATION: A BRIEF HISTORY", Concise history of the language sciences: from the Sumerians to the cognitivists. Edited by E.F.K.Koerner and R.E.Asher. Oxford: Pergamon Press, 1995. Pages 431-445.

[7] Khin Thandar Nwet, Khin Mar Soe and Nilar Thein, "Word Alignment System Based on Hybried Approach for Myanmar-English Machine Translation",

[8] Dr. Thein Tun, "Acoustic Phonetics and the Phonology of the Myanmar Language", School of Human Communication Sciences, La Trobe University, Melbourne, Australia, 2007.

[9] D.J. RAVI Research Scholar, "Kannada Text to Speech Synthesis Systems: Emotion Analysis", JSS Research Foundation, S.J College of Engg, Mysore-06, 2010.

[10] Ei Phyu Phyu Soe, "Grapheme-to-Phoneme Conversion for Myanmar Language", The 11th International Conference on Computer Applications (*ICCA 2013*).

[11] "*Phoneme*", http://en.wikipedia.org/w/index.php, April 2012.

[12] Ei Phyu Phyu Soe, "Prosodic Analysis with Phonological Rules for Myanmar Text-to-Speech System", AICT 2013.

[13] Tractament Digital de la Parla, "Introduction to Speech Processing".

[14] Hayes, Bruce (2009). "Introductory Phonology." Blackwell Textbooks in Linguistics. Wiley-Blackwell. ISBN 978-1-4051-8411-3.

[15] S. Lemmetty, "Review of Speech Synthesis Technology", Master's Thesis, Helsinki University of Technology, 1999.

[16] Zin Maung Maung, Yoshiki Mikami, "A Rule-based Segmentation for Myanmar Text", Nagaoka University Technology, 2007.