I.J. Information Technology and Computer Science, **2018, 11, 58-66**
Published Online November 2018 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2018.11.06

A Comparison between Syllable, Di-Phone, and Phoneme-based Myanmar Speech Synthesis

Aye Thida
Artificial Intelligence Lab, Faculty of Computer Science, University of Computer Studies, Mandalay, Myanmar
E-mail: ayethida.royal@gmail.com

Chaw Su Hlaing
Artificial Intelligence Lab, Faculty of Computer Science, University of Computer Studies, Mandalay, Myanmar
E-mail: chawsuhlaing.ucsm.mm@gmail.com

Received: 30 August 2018; Accepted: 24 September 2018; Published: 08 November 2018

***Abstract*—**Among the speech synthesis approach, concatenative method is one of the most popular method which can produce more natural sounding speech output. The most important challenge in this method is choosing an appropriate unit for creating a database. The present used speech units are word, syllable, di-phone, tri-phone and phoneme. The speech quality may be trade-off between the selected speech units. This paper presents the three speech synthesis system of Myanmar language, respectively based on syllable, di-phone and phoneme speech units by using concatenation method. Then, we compare the speech quality of the three systems, using the subjective tests.

***Index Terms*—**Myanmar language, phoneme, concatenative speech synthesis.

I. INTRODUCTION

Text to speech conversion is a system that converts the written text into their corresponding speech. It can help the visual and speech impaired person. Nowadays, many text to speech systems have already developed for different languages and they may be available commercially from simple text to speech up to online reader by using different speech synthesis approaches. The current approaches can be classified into three groups: formant synthesis, articulatory synthesis, and concatenative synthesis approach. Although the first two approaches got the acceptable level of speech quality, they have complex computation performance and difficulties to control speech signal. Therefore, the last approach, concatenative synthesis become more popular because it can generate more natural sound and less computational efforts. In concatenative speech synthesis approach, the required speech units are pre-recorded and just concatenate them. In this case, the speech quality and the size of the system depending on the selected speech units for concatenation. The short speech unit may be degraded speech quality but does not need multiple speech units and besides, less memory required. Otherwise, the longer speech units can generate more natural speech output but required multiple speech units to cover for the specified language and large memory spaces are required. The current speech units are the word, syllable, phoneme, di-phone, tri-phone and so on [1]. Many TTS systems proposed by [2, 3, 4, 5, 6] have been implemented by using the concatenative method based on different speech units and they can generate high quality synthesized speech. A numerical TTS synthesis system for three languages: Marathi, Hindi, and English languages is proposed by [7]. They used the approach that combined rule-based approach and concatenation based approach. They used all utterances of sound units have been used for concatenation and generation of the speech signal. [8] propose algorithms and methods that point out critical subjects in developing a general Amharic text-to-speech synthesizer. They used unit selection concatenative synthesis approach and the main issue of Amharic language are described and allophone labeling scheme is introduced for easily locating target sound units. This is one of the major difficulties of the unit selection approach that selected the right unit from multiple sound units. These units are recorded and segmented with the relevant allophonic variations. According to the literature, the selection of speech units is very important in concatenative speech synthesis approaches. Therefore, later research works compare the TTS systems with different speech units in [9] that compared two Arabic text to speech systems: two screen readers, namely, Non-Visual Desktop Access (NVDA) and IBSAR. They tested the quality of two systems in terms of standard pronunciation and intelligibility tests with the visually impaired person. According to their results, the NVDA outperformed IBSAR on the pronunciation tests. However, both systems gave the competitive performance on the intelligibility tests. [10] developed the unit selection based and HMM-based speech synthesizer for the Urdu language. They also developed phonetic lexicon that contains 70597 words and 10 hours speech data is used for both synthesizers. Then, they compared these two synthesizers. According to their evaluation result, the HMM generated speech outperformed in intelligibility than the unit selection based synthesizer. Nowadays, current researchers, up to proposed the optimal character recognition (OCR) based-TTS system to help such visually challenged person by OCR [11]. The result texts from the OCR is converted into speech. They used blind deconvolution method and pre pre-processing operation to remove the effect of noise and blur so that they can achieve the efficient result of the framework for visually challenged.

For Myanmar language, there has been the considerable effort on speech processing in Myanmar natural language processing. Typically, text to speech systems in different languages have been developed by using different approaches as well as for Myanmar language. [12] designed the rule-based MTTS system in which fundamental speech units are demi-syllables with the level tone. They used a source filter model and furthermore a Log Magnitude Approximation Filter. The high intelligibility of the synthesized tone was confirmed through listening tests with correct rates of over 90%. According to their result, they have high intelligibility but the speech output is the similar robotic voice in naturalness. Therefore, di-phone based MTTS is developed by [13]. They used the concatenative synthesis method and time domain pitch synchronous overlap-add (TD-PSOLA) for smoothing concatenation points. Their diphone database which includes over 8000 diphones for 500 Myanmar sentences. The speech units are too much for the intention of resource limited devices. Moreover, if the required di-phone pair does not contain in the created database, the system results degraded. Therefore, [14] proposed a new phoneme concatenative method for MTTS system. Their system is suitable for resource limited because their phoneme speech database contained only 133 phonemes that can speak out for all Myanmar text. According to their result, they also got the acceptable level for the intelligibility but still need naturalness. Consequently, in their method, they did not consider the half sound of consonant. In Myanmar language, the most of the minor consonants stand with schwa vowel. These kinds of consonant have half-sound of original one. In the previous method [14], some rules for schwa insertion are presented but this schwa vowel sound did not consider in the speech synthesis module. Therefore, in this paper, we discuss the comparison result of three speech units: syllable, diphones and phoneme by using concatenation method. In phoneme concatenation, we also considered the half-sound consonant for which we have to prepare the text for recording. After that, segment and label the recorded sound to get half-sound of the consonant. Then, fetch the appropriate speech files from the created speech database and concatenate them.

The rest of this paper is organized as follow: the nature of the language intended for this system is introduced in Section II. The overview explanation of Myanmar text to speech system is presented in Section III. The method used in this paper, concatenative based speech synthesis is described in Section IV. Then, the main purpose of this paper, syllable based, diphones based and phoneme based MTTS systems are presented in Section V, VI and VII. The speech database comparison is discussed in Section VIII. The experimental results of the proposed method are discussed in Section IX. Finally, the paper is concluded in Section X.

II. Myanmar Writing System

There are approximately a hundred languages spoken in Myanmar. Among them, Myanmar language is the official language and it is spoken by two thirds of the population. Myanmar alphabet consists of 34 consonants but some has same pronunciation so that there are only 21 consonant sound as shown in Table 1 and 9 basic vowels which can be extended into 50 vowels sound according to tone level as shown in Table 2. Myanmar language is tonal language and it is written from left to right. It requires no spaces between words, although modern writing system usually contains spaces after each clause to enhance readability. Myanmar speech and letter are based on the combination of consonant and vowel phoneme so called syllable. 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 [15].

Table 1. Consonant Phonemes

No.	Phoneme Symbol	Consonant	No.	Phoneme Symbol	Consonant
1	/k/	က	12	/n/	ဏ၊န
2	/k^h/	ခ	13	/p/	ပ
3	/g/	ဂ၊ဃ	14	/p^h/	ဖ
4	/ŋ /	င	15	/b/	ဗ၊ဘ
5	/s/,	စ	16	/m/	မ
6	/s^h/	ဆ	17	/j/	ယ၊ရ
7	/z/	ဇ၊ဈ	18	/l/	လ၊ဠ
8	/ɲ/	ဉ၊ည	19	/w/	ဝ
9	/t/	ဋ၊တ	20	/θ/	သ
10	/t^h/	ဌ၊ထ	21	/h/	ဟ
11	/d/	ဍ၊ဎ၊ ဒ၊ဓ			

III. Overview of Myanmar Text to Speech

Myanmar text-to-speech system is developed by using concatenation method. In the implementation of Myanmar text to speech system, there are four main step: text analysis, phonetic analysis, prosodic analysis and speech synthesis as shown in Figure 1.

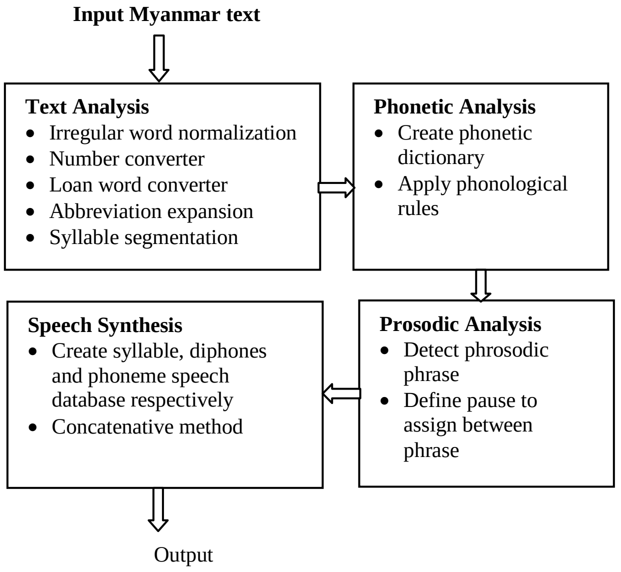

Fig.1. Block diagram of MTTS System

In the step of text analysis, there five main parts: irregular word normalization, number converter, loanword converter, abbreviation expansion and syllable segmentation. Irregular word normalization is accomplished from stack consonants to normal readable form. The determination of the number converter is to convert the number to textual styles. The non-standard words may be numbers that need to be expanded into the readable form of Myanmar words before they are pronounced. Number expands to the string of words representing cardinal, money, decimal number, fraction and so on. The input texts are segmented into Myanmar text like a syllable. Syllable segmentation is the process of identifying syllable boundaries in a text. In this paper, a rule based approach of syllable segmentation algorithm for Myanmar text is used for syllable segmentation. They created segmentation rules based on the syllable structure of Myanmar script and a syllable segmentation algorithm was considered based on these created rules [16].

The Myanmar syllable structure is expressed in finite state automatic (FSA) algorithm as shown in below:

$$\text{Syllable} ::= C\{M\}\{V\}\{F\} \mid C\{M\}VA \mid C\{M\}\{V\}CA[F] \mid E[CA][F] \mid I \mid D$$

The result from the text analysis is feed into the second steps, phonetic analysis, is also called G2P (Grapheme to Phoneme). Grapheme to Phoneme conversion decodes the segmented syllable of Myanmar text into their corresponding phonetic sequence. It defines the pronunciation of a syllable based on its spelling. It also explores the most exact phonemes sequence of words, numbers and symbols and converts into phonetic sequences. We have constructed the Myanmar phonetic dictionary to generate the phoneme sequence and to pronounce these phonemes. Writing form and speaking form are different in Myanmar writing system. Therefore, the phonological rules are considered to get the more natural sound [17].

The phonetic sequences are analyzed to produce the prosodic features by applying the phonological rules in prosodic analysis step. It is the module to analyze duration and intonation such as pitch variation, syllable length to create naturalness of synthetic speech. There are three aspects of prosody, each of which is important for speech synthesis: prosodic prominence, prosodic phrasing and tune. In the proposed MTTS systems, prosodic phrasing is considered to get better quality TTS system. There is no space between words or phrase so the segmentation for this kind of state is a challenging task in Myanmar natural language processing. Therefore, detecting the phrase break and then assign pause duration between Myanmar words and phrases are the main steps in prosodic analysis.

Synthesized speech can be created by several different methods. The methods are usually classified into three groups: (i) articulatory synthesis, (ii) formant synthesis and (iii) concatenative synthesis. The aim of this paper is to improve the quality of syllable-based and diphone-based and phoneme based of Myanmar Text-To-Speech by applying the Concatenative speech synthesis. Each system and the comparison of these three speech units based MTTS system are presented in the next sections.

IV. Concatenative Based Speech Synthesis

The majority of concatenation-based TTS system employs two different approaches for the realization of natural prosody: data-driven and post-processing modification. Data-driven methods integrate prosodic features in conjunction with the segmental units so that the best prosody is attained by optimal unit selection from a large inventory of pre-recorded segments. The approach of post-processing is to modify the concatenated speech signal so as to reach the prescribed prosodic targets.

Table 2. Vowel Phonemes

Basic Symbol		**Non-nasalized (နှာသံမဲ့သရ)**							
		Tone I		**Tone II**		**Tone III**		**Tone IV**	
i¯	အီ	i¯	အီ	iˆ	အီး	i´	အိ	ɪˀ	အစ်
e¯	အေ	e¯	အေ	eˆ	အေး	e´	အေ့	eɪˀ	အိတ်
ɛ¯	အယ်	ɛ¯	အယ်	ɛˆ	အဲ	ɛ´	အယ့်	ɛˀ	အက်
								aɪˀ	အိုက်
a¯	အာ	a¯	အာ	aˆ	အား	a′	အ	ʌˀ	အတ်
ɔ¯	အော်	ɔ¯	အော်	ɔˆ	အော	ɔ´	အော့	aʊˀ	အောက်
o¯	အို	o¯	အို	oˆ	အိုး	o´	အို့	oʊˀ	အုပ်
u¯	အူ	u¯	အူ	uˆ	အူး	u´	အု	ʊˀ	အွတ်
		Nasalized (နှာသံပါသောသရ)							
		Tone I		**Tone II**		**Tone III**			
		ĩ¯	အင်	ĩˆ	အင်း	ĩ´	အင့်		
		eĩ¯	အိန်	eĩˆ	အိန်း	eĩ´	အိန့်		
		aĩ¯	အိုင်	aĩˆ	အိုင်း	aĩ´	အိုင့်		
		ʌ̃¯	အန်	ʌ̃ˆ	အန်း	ʌ̃′	အန့်		
		aʊ̃¯	အောင်	aʊ̃ˆ	အောင်း	aʊ̃´	အောင့်		
		oʊ̃¯	အုန်	oʊ̃ˆ	အုန်း	oʊ̃′	အုန့်		
		ʊ̃¯	အွန်	ʊ̃ˆ	အွန်း	ʊ̃′	အွန့်		
		ĩ¯	အင်	ĩˆ	အင်း	ĩ´	အင့်		
		eĩ¯	အိန်	eĩˆ	အိန်း	eĩ´	အိန့်		

Generally, the process of concatenative speech synthesis is very simple and it selects the corresponding speech files from a pre-recorded database and joined one after another to produce the desired utterances. Perfectly, the required speech units like syllable, diphones or phonemes, phonemes for concatenation is extracted from the created inventory such that its context in the sentence under construction is the same as that in which it was recorded. 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. Besides, the output speech from concatenative synthesis is mainly depends on the selected database. For instance, the behavior or the affective tone of the speech is hardly controllable. Despite the somewhat featureless nature, concatenative synthesis is well suited for certain limited applications [18]. As an assumption, Concatenative synthesis is based on the concatenation segments of recorded speech. Generally, it produces the most natural-sounding synthesized speech. If the selected unit is longer one, it is easier to obtain more natural sound and it can achieve a high segmental speech quality.

V. Syllable Based MTTS System

A syllable is a unit that has pronunciation with one vowel sound, plus or not surrounding consonants, forming the whole or a part of a word. Typically, a syllable onset is the consonant or cluster of consonant that appears before the vowel of a syllable. So, the 34 consonant letters in Myanmar indicate as the initial consonant of a syllable and Myanmar script has four basis medials diacritics to specify additional consonants in the onset. Like other abugidas, including the other members of the Brahmic family, diacritics are used to designate as vowels in Myanmar script, which are placed above, below, before or after the consonant character. Therefore, Myanmar syllable structure has the phonemic shape of C (G) V (N/ʔ) T, where an initial consonant C is mandatory, a glide consonant G is optional, a vowel V is mandatory, a final consonant-nasal N or stopped ʔ is optional, and tone T is mandatory, respectively. The minimum syllable structure is CVT in Myanmar language. There are 6 possible syllable structures as shown in Table 3.

Table 3. Six possible syllable structure

No.	Syllable Structure	Myanmar Words
1.	CVT	မိန်း (girl)
2.	CVCT	မက် (crave)
3.	CGVT	မြေ (earth)
4.	CGVT	မျက် (eye)
5.	CVVCT	မောင် (first position of male name)
6.	CGVVCT	မြောင်း (ditch)

Different languages may have different syllable patterns so that the number of syllable may be large. Generally, the number of different syllables in each language is considerably not large as much as words. However, it is still too large for TTS system for resource limited devices. For English language, there are around

15,000 syllables [19].

In Myanmar language, the syllable can become if the consonant is attached with vowel which means the combination of consonant and vowel can make a syllable so that the 34 consonants and 47 medials consonants can be expand with 50 vowels into 4050 syllables (34*50 + 47*50 = 4050) that are all possible syllables in Myanmar language. However, some consonant has the same pronunciation and some combination cannot make syllable because they cannot be pronounced. Therefore, finally, the number of syllables is 2340 in Myanmar language. Therefore, the sentences that covered for all Myanmar syllables are selected and recorded. Then these are segmented and labeled. They all are stored in the syllable speech database. Finally, fetch the appropriate syllable speech file according to phoneme sequence the result from the prosodic analysis and generated the Myanmar speech output by using syllable concatenation method.

VI. Di-Phone Based MTTS System

Diphones are defined to extend the central point of the steady state part of the first phone to the central point of the following one, so they contain the transitions between adjacent phones. Consequently, the concatenation point will be in the steadiest state region of the signal so that it reduces the distortion from concatenation points. Another advantage of using diphones is that there is no more need to be formulated rules for the co-articulation effect. In the current speech synthesis approaches, diphones is the most widely thought speech units.

In standard, if the phones number is P in a particular language, then the number of diphones is P^2 in theoretically. Actually, all languages may have restrictions like what sounds can occur next to each other so that the number of diphones in each language is usually much smaller than p^2 because there is no need to combine all possible phonemes in the case of creating diphones speech database. For example – Spanish has about 800 diphones and German has about 2,500.

In Myanmar language, the diphones speech unit's selection is based on the one word to cut the two sound files and it is to explicitly list all possible phone-phone transition. Firstly, we list the possible words to get the enough diphones pairs for concatenation so that the size of the diphone database has around 10496 diphones in Myanmar language. The possible words for diphone pairs are calculated in following Table 4.

Table 4. Calculation of Possible Diphone Pairs

Square of words	-	Square of vowels	=	Possible Diphones
(114 x 114)	-	2500	=	10496

There are many Myanmar words for 34 consonants, 50 vowels and 47 medical consonants words. The square of words reduces the square of vowels is equal to the possible unit-based diphone pairs. The 114 words combine with 21 consonants, 47 medial and 50 vowels and 2500 vowels are the square of 50 vowels. Therefore, the Myanmar Diphone database stores around 10496 diphones pairs for every Myanmar words (114 (21 Consonants + 47 Exception Words + 50 Vowels) x 114 (21 Consonants + 47 Exception Words + 50Vowels) – 2500 (50Vowels x 50Vowles)) 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. Typically, database size is larger than syllable database, however, they are not only easier to deal with in runtime but also at ease to obtain a positive system with diphone concatenation. Moreover, in some modern systems, a combination of this unit and other methods such as unit selection are used. However, it is very large for the resource limited devices.

VII. Phoneme Based MTTS System

In speech synthesis, phonemes are perhaps the most normally used units as they are the normal linguistic representation of speech. Generally, the basic unit inventory is usually between 40 and 50, which is clearly the smallest compared to other units [19]. Using phonemes gives maximum flexibility with the rule-based systems. Nevertheless, it is difficult to synthesize the phones like plosives because they do not have a steady state target position. The articulation must also be formulated as rules. Sometimes, phonemes are used as an input for speech synthesizer to drive like diphone based synthesizer. Phoneme is the basic and smallest sound unit that can distinguish one word from another in a particular language.

In English, the two words *pit* and *bit*, they have different sound in the only first position such as start with /p/ in pit and /b/ for bit. These two phoneme /p/ and /b/ can distinguish two different sound and meaning of word as they are the smallest unit of sound and that cannot be separated again. In the word "*pet* and *pit*", they are only different in vowel sound /e/ and /i/. Therefore, /p/, /b/, /e/, /i/ can be defined as phoneme in English language [11]. Phonemes are conventionally placed between slashes in transcription. In this case, the word which has only one different phoneme is called minimal pair. The Myanmar word ပန်း-/pan/ (flower) and ဗန်း-/ban/ (tray) are different in the first position of sound as /p/ and /b/. Likewise, for the word ပန်း-/pan/ (flower) and ပုန်း-/pon/ (hide), they have only different vowel sound /a/ and /o/. Therefore, /p/, /b/, /a/ and /o/ are phoneme in Myanmar language and the word ပန်း, ဗန်း and ပုန်း are minimal pair because they have only one differ phone [20].

Phonology is a branch of linguistics concerned with the systematic organization of sounds in languages [21]. It has traditionally focused largely on the study of the systems of phonemes in particular languages. Myanmar phonology is generally constructed by means of combining the consonant phoneme, vowel phoneme and tone as shown in Table 5.

Table 5. Myanmar Phonological Structure

Consonant	Vowel	Tone	Myanmar Syllable
က	◌ား	/ˆ/	ကား
ကျ	◌ောင်း	/ˊ/	ကျောင်း

Although there are 34 consonants, some consonant has the same pronunciation so that there are only 21 consonant phonemes. The above 21 consonant letters (C) may be modified by one or more medial diacritics (three at most) before the vowel indicating an additional consonant. These diacritics are: Ya pin (◌ျ), Ya yit (ြ◌), Wa hswe (◌ွ) and Ha hto (◌ှ) indicated by /j/, /j/, /w/ and /h/ respectively. The basic medials can be combined to become totally 11 medials. These 11 medials can modify the 34 basic consonants (11*21=231). For example –မ (C) + ြ◌ (M)=မြ (emerald), မ (C)+ ◌ှ(M) =မှ (from). However, these medials cannot combine all of the consonants. For example – က(C) + ◌ျ (M) + ◌ှ (M) = ကျှ. Therefore, after final counting the syllable that can be pronounced by combining with medials are only 47 syllables in Myanmar scripts. The vowel plays in vital roles in the construction of syllable. Any syllable can be achieved by combining 50 vowels sound with 21 consonants and 47 medials consonant.

Moreover, in Myanmar language, when basic consonants (က-/k/, ခ-/kh/, မ-/m/) is combined with other syllables (လေး, ရေ, နက်ဖြန်) to become word (ကလေး (baby), ခရေ (star flower), မနက်ဖြန် (tomorrow), the speech sound of these basic consonant is not fully pronounced. It turns into half-vowel or schwa as [kə leˆ], [kʰə ɹe¯], [mə n ɛˀ pʰj ʌ̃¯] instead of [kà leˆ], [kʰà ɹ e¯], [mà nɛˀ pʰjʌ̃¯]. Therefore, mostly, if the basic consonant is situated in the first position of the word, it is pronounced as schwa. This kind of half sound consonant may have every consonant and some medials word such as (ကျ ချ ဂျ). Therefore, the extra 25 (22+3) consonant sound file which has half sound nature. Figure 2 shows that how to get the half sound of the consonant "ကလေး (baby)" is got.

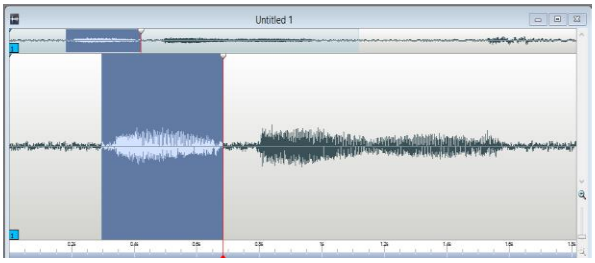

Fig.2. Cutting the half sound of consonant "k" from the word "ကလေး"

In the Myanmar phoneme speech database, at first, there are only 133 phoneme speech sound in [14] that did not consider half-sound. Now, after considering this kind of sound, the speech units increase up to 157 phoneme speech files by adding extra 24 minor consonants as shown in Table 6. These phonemes can make sound for all kind of Myanmar texts. Therefore, this amount of phoneme is significantly reduced than the di-phone based and syllable based MTTS system so that it is very suitable for resourced limited devices.

Table 6. Number of phonemes used in Phoneme based MTTS

Phoneme Type	Number of Phonemes	Cannot pronounce	Can pronounce
Consonants (C)	21	0	21
21(C)* 6 Medial	126	79	47
Vowels(V)	50	0	50
Half sound consonant	24	0	24
Special Character	5	0	5
Number	10	0	10
Total Phonemes			**157**

VIII. Speech Database Comparison

Myanmar speech synthesis has been developed based on three speech units: syllable, diphones and phoneme by using concatenative synthesis approach. The number of speech files of corresponding speech units for MTTS system is shown in the Table 7.

Table 7. Speech Database of three speech units

Speech Unit	Number of speech files	Recording Time (hours)
Syllable	2305	4
Diphones	10496	16
Phoneme	157	2

The structure of speech database for three units is created with Arpabet sign as shown in Table 9 and 10 to be easy in the retrieving of the respective speech files. Firstly, all possible syllables (2305 syllables) are analyzed and recorded for syllable speech database. Then they are segmented and labeled with pair of Arpabet signs according segmental sound file. It took over 8 hours for recoding. For the diphones database, the training Myanmar sentences are recorded to get the diphones pairs. It took about 8 hours in two days to complete the desired diphones speech database. Then, the segmented diphone pairs (10496) are also assigned with their corresponding Arperbet sign. For the phoneme speech database, the selected consonant phoneme, medials phonemes, vowel phoneme and the half sound consonant speech files, totally 157 phonemes are recoded separately. Then, they are also segmented and labeled. It takes only around 2 hours. Sample labeling with the corresponding with arperbet signs for three speech units is presented in the Table 8 with example sentence "မင်းနေကောင်းလား" *(How are you?).*

Table 8. Labeling with Arperbet sign of three speech units

Input Sentence (Phonetic Symbol)	မင်းနေကောင်းလား။ *(How are you?)* [mĩˆ ne¯ kaʊ̃ˆ laˆ]
Syllable	MIH2+NAE1+KAW2+LAA2
Diphone	PAU-M+M-IH2+IH2-N+N-AE1+AE1-K+K-AW2+AW2-L+L-AA2+AA2-PAU
Phoneme	PAU+M-IH2+N-AE2+K-AW2+L-AA2

These three units have advantage and disadvantages. The speech units are selected based on the system demands. According to the Table 7, the diphones speech database is the largest and the syllable speech database is more than the phoneme speech database. The more the speech units the database has, the more take time for recording, segmenting and labeling speech files. It is time consuming tasks and need the human effort for it. In the application point of view, if the MTTS system is intended for resourced limited devices, the size of the app is comparable based on the created speech database size. Therefore, the phoneme speech databased size is the most suitable for the resource limited device.

Table 9. Sample Arpabet symbol for consonant

IPA	Arpabet
k	K
k^h	KH
g	G
….	….
a^h	AH

Table 10. Sample Arpabet symbol for vowel

IPA	Arpabet
ɪ̃ ¯	IH1
…	…
ɛˀ	IH4
aʊ̃ ¯	AW1
…	…
aʊ̃ ´	AW4
a ¯	AA1
…	…
a′	AA3

IX. Evaluation Method

The two-quality measures are considered for testing the quality of three MTTS system. They are intelligibility test which measures how much the user understand what the speech output is and the naturalness test which measure how much the system output is similar to the real human speech.

The most popular methods for evaluation: mean opinion score (MOS) is used for two quality tests. The MOS method is the most useful and simplest method to test the quality of MTTS systems. It has five level scales in MOS: bad (1), poor (2), fair (3), good (4) and excellent (5) [22]. The formula for MOS calculation is described as the following equation (1).

$$MOS = \frac{\sum_{j=1}^{M}\left(\frac{\sum_{i=1}^{N} S_{i\,j}}{N}\right)}{M} \qquad (1)$$

Moreover, the word error rate (WER) is another evaluation method for the intelligibility test. For this method, the speech files are played for the evaluators. Then, they have to write whatever they heard, even if they don't understand the meaning. According to their results, we calculated WER by using the following equation (2).

$$WER = \frac{substitutions + insertions + deletions}{reference\_length} \qquad (2)$$

Generally, the evaluation can also be made at several levels, such as phoneme, word, or sentence level, depending on what kind of information is needed. In this paper, MTTS systems which has been developed by using concatenative speech synthesis approach is evaluated based on three speech units (syllable, diphones and phonemes) The evaluation process is usually done by subjective listening tests. For MTTS testing, the 8 evaluators help and test the speech quality of MTTS system for the 100 sentences that have average words length is 15.

A. Naturalness Test

For the syllable based MTTS system, the evaluators are asked the question of the sound quality how much the listeners feel the voice is similar to the real person. Regarding their answers, for the syllable based MTTS system, 0.5% of listeners thought about the output speech is very natural, 15.16% considered the speech are natural and 60.45% of listener identified the voice are acceptable. Around 22.82% assumed the speech output is needed to get more naturalness and only 0.84% though the worst. For the diphones based MTTS system, 0.5% of listeners supposed that the output speech is very natural, 15.16% well-thought-out the speech are natural and 60.45% of listener recognized the voice are acceptable. Round 22.82% assumed the speech output is needed to get more naturalness and only 0.84% though the worst condition. For the phonem based MTTS system, 0.5% of listeners thought about the output speech is right natural, 15.16% considered the speech are natural and 60.45% of listener identified the voice are acceptable. Around 22.82% assumed the speech output is needed to get more naturalness and only 0.84% though the worst. The average percent for three speech units are shown in Figure 3.

B. Intelligibility Test

The question for the intelligibility of the speech quality is how much the subjective understood the voice or how much of what the voice said the subjective understood. In these case, for the syllable based MTTS system, 19.75% of the subjective understood very well. 15.5% did understand the voice very much and 37.47% neither much nor little and another 23.60% understood a little and noone did understand very well. For the diphone based MTTS system, 84.34% of the subjective understood very well. 13.66% did understand the voice very much and 1.86% neither much nor little. For the phoneme based MTTS system, 19.87% of the subjective understood very well. 39.13% did understand the voice very much and 32.29% neither much nor little and another 8.07% understood a little and only 0.62% did not understand very well. The WER for three speech are 10,4 and 7 respectively. The percnentage of intelligibility for three units are shown in the following Figure 4.

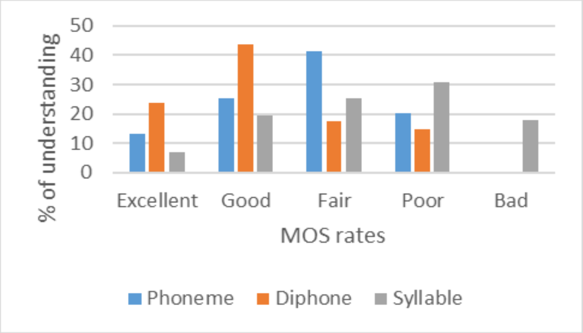

Fig.3. MOS value of naturalness test of three speech units

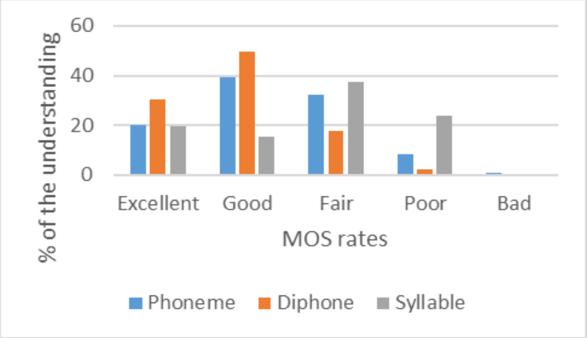

Fig.4. MOS value of Intelligibility test of three speech units

X. Comparison Result For Three Speech Units

In the syllable based system, there are 725 syllable speech files for the selected 100 testing sentences. On if the system can retrieve the corresponding speech file of the input sentence, the speech quality may be increased. Therefore, the quality depends on the stored syllable speech files. In the diphone based MTTS system, the quality of the system depends on number of diphone pairs which covered the meaning for each word. The 1382 diphone pairs are needed for the 100 sentences in the diphone-concatenation. The drawback of their system is that the system cannot produce speech output if the diphone pairs do not exist in the created speech database. However, the defined 157 phonmes speech files can speech out for all Myanmar words. The comparison results for average MOS values and number of speech files for 100 sentences are shown in the Table 11.

Table 11. Comparison results for three speech units

Criteria	Syllable	Diphones	Phoneme
No. of sentence	100	100	100
Required speech files	628 (syllable)	1541 (diphone pairs)	157 (phoneme)
Allow any Myanmar Text	No	No	Yes
Speech unit size	31MB	53MB	19MB
MOS for Intelligibility	3.2	4.81	3.69
MOS for Naturalness	2.52	4.56	2.97

In the application point of view, the proposed concatenative based MTTS systems are intended for the resourced limited device so that the size of the app is comparable based on the created speech database size. After developing MTTS systems, phoneme based MTTS has smallest speech inventory size but its MOS values are less than the diphones based and higher than the syllables based. Diphones based MTTS systems has the highest MOS score for speech quality. The syllable based is weaker than the last two units. As conclusion, according to the required demand, there may be different in the case of choosing the speech units for concatenation. If the system only considers speech quality and not for storage performance, the diphones speech units is the best to choose. Otherwise, the system considers the storage requirement and accepts the fair state in speech quality, the phoneme is the most suitable for resource limited devices. The final results for the three speech units are shown in the Figure 5.

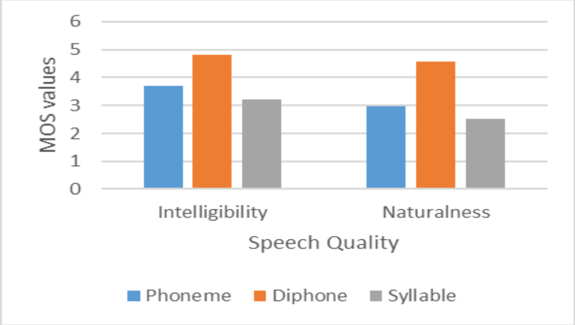

Fig.5. The MOS value comparison result for three units

XI. Conclusion

This paper presented the three Myanmar speech system based on three speech units: syllable, diphones and phonemes. In the case of developing MTTS system on resource limited devices, while there are 2305 syllable speech units are needed for the syllable based MTTS system, 10496 diphones pairs are required for the diphone based MTTS system. However, 157 phoneme units can speech out for any Myanmar text in phoneme based MTTS system. These three MTTS systems can support offline support translation. Finally, the comparison of these three units are presented. According to the experimental result, the selected speech units are depended on the user demand. If more naturalness is needed, diphones is the best and if the storage power is considered, phoneme is the most suitable to choose.

ACKNOWLEDGMENT

The authors wish to thank all members in the NLP lab at the University of Computer Studies, Mandalay.

REFERENCES

[1] S. Lemmetty, 1999. Review of speech synthesis technology. Helsinki University of Technology, 320, pp.79-90.

[2] A. Black and N. Campbell, “Optimizing selection of units from speech database for concatenative synthesis,” Proceeding of EUROSPEECH’95, vol. 1, pp. 581-584, Sept. 1995.

[3] A. Conkie, “A robust unit selection system for speech synthesis,” Proc. of 137th meet. ASA/Forum Acusticum, 1999.

[4] A. J. Hunt and A. W. Black, “Unit selection in a concatenative speech synthesis system using a large speech database,” Proc. of ICASSP, vol. 1, pp. 373-376, 1996.

[5] T. Toda, H. Kawai, M. Tsuzaki and K. Shikano, “Unit selection algorithm for Japanese speech synthesis based on both phoneme unit and diphone unit,” Proc. of ICASSP, vol. 1, pp. 465-468, May 2002.

[6] M. Douke, M. Hayashi, and E. Makino, “A study of automatic program production using TVML,” Short Papers and Demos, Eurographics, pp. 42-45, 1999

[7] G. D. Ramteke, R. J. Ramteke, “Efficient Model for Numerical Text-To-Speech Synthesis System in Marathi, Hindi and English Languages”, International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.9, No.3, pp.1-13, 2017.DOI: 10.5815/ijigsp.2017.03.01

[8] E.B. Kasie and Y. Assabie, October. Concatenative speech synthesis for Amharic using unit selection method. In Proceedings of the International Conference on Management of Emergent Digital Eco Systems, pp. 27-31. ACM, 2012.

[9] N. K. Bakhsh, S. Alshomrani, Imtiaz Khan, “A Comparative Study of Arabic Text-to-Speech Synthesis Systems”, IJIEEB, vol.6, no.4, pp.27-31, 2014. DOI: 10.5815/ijieeb.2014.04.04

[10] F. Adeeba, T. Habib, S. Hussain, and K.S. Shahid, October. Comparison of Urdu text to speech synthesis using unit selection and HMM based techniques. In Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA), 2016 Conference of The Oriental Chapter of International Committee for (pp. 79-83). IEEE, 2016.

[11] A. Verma, D. K. Singh, “Robust Assistive Reading Framework for Visually Challenged”, International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.9, No.10, pp. 29-37, 2017.DOI: 10.5815/ijigsp.2017.10.04

[12] K. Y. Win and T. Takara, “Myanmar text-to-speech system with rule-based tone synthesis,” Acoustical Science and Technology, vol. 32, no. 5, pp. 174–181, 2011.

[13] E. P. P. Soe and A. Thida, “Text-to-speech synthesis for Myanmar language”, International Journal of Scientific & Engineering Research, vol. 4, no. 6, pp. 1509–1518, 2013.

[14] C.S. Hlaing and A. Thida, “Phoneme based Myanmar text to speech system”, International Journal of Advanced Computer Research, *8*(34), pp.47-58, 2018

[15] Myanmar Language Commission, Myanmar Grammar, 30th Year Special Edition, University Press, Yangon, Myanmar, 2005.

[16] Z. M. Maung and Y. Mikami, “A rule-based syllable segmentation of Myanmar text”. In Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, 2008.

[17] T. Tun, “Acoustic phonetics and phonology of the Myanmar language”, School of Human Communication Sciences, La Trobe University, Melbourne, Australia, 2007.

[18] Schwarz, Diemo. “Current research in concatenative sound synthesis.” In International Computer Music Conference (ICMC), pp. 1-1. 2005.

[19] M. Karjalainen Review of Speech Synthesis Technology. Helsinki University of Technology, Department of Electrical and Communications Engineering. 1999 Mar 30.

[20] M. T. Noe, “Study of Myanmar Phonology”, 3rd Edition, University Press, Yangon, Myanmar, 2007.

[21] Clark, John; Yallop, Colin; Fletcher, Janet, “An Introduction to Phonetics and Phonology (3rd ed.)”. Massachusetts, USA; Oxford, UK; Victoria, Australia: Blackwell Publishing. ISBN 978-1-4051-3083-7, 2007.

[22] “Text to speech testing strategy, Version 2.1”, Technology Development for Indian Languages Programme DeitY, 07 July, 2014

Authors’ Profiles

Aye Thida is a Professor, at Faculty of Computer Science, Artificial Intelligence Lab, University of Computer Studies, Mandalay and Myanmar. Her research interests include in Machine Translation, Text-to-Speech System and Big Data management. She is currently working NLP researches. She received B.Sc. (Hons:), Math degree from the Mandalay University, Myanmar and her M.I.Sc. and Ph.D. degrees in Computer Science from the University of Computer Studies, Yangon (UCSY), Myanmar.

Chaw Su Hlaing is a Ph.D. student at Artificial Intelligence Lab, Faculty of Computer Science, University of Computer Studies, Mandalay and Myanmar. She had received her B.C.Sc. and M.C.Sc. from Faculty of Computer Science, University of Computer Studies, Mandalay and Myanmar. Her current research interests are Web Data Mining, Digital Signal Processing, Natural Language Processing and Linguistic Research. She is currently working in the research of Speech Synthesis for Myanmar language.

How to cite this paper: Aye Thida, Chaw Su Hlaing, "A Comparison between Syllable, Di-Phone, and Phoneme-based Myanmar Speech Synthesis", International Journal of Information Technology and Computer Science(IJITCS), Vol.10, No.11, pp.58-66, 2018. DOI: 10.5815/ijitcs.2018.11.06