I.J. Information Technology and Computer Science, **2018, 12, 13-20**
Published Online December 2018 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2018.12.02

Phoneme Concatenation Method for Myanmar Speech Synthesis System

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

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

Received: 14 August 2018; Accepted: 22 September 2018; Published: 08 December 2018

***Abstract*—**This paper discusses the approach used to develop a Text-to-Speech (TTS) synthesis system for the Myanmar language. Concatenative method has been used to develop this TTS system using phoneme as the basic units for concatenation. In this proposed system, phoneme plays an important role so that Myanmar phoneme inventory is presented in detail. In Myanmar language, schwa is the only vowel that is allowed in a minor syllable or consonant that has half-sound of the original one. If these half sound can be handled, the TTS quality will be high. After analyzing the number of phoneme and half-sound consonant to be recorded, create the Myanmar phoneme speech database which contains total 157 phoneme speech sounds that can speech out for all Myanmar texts. These phonemes are fetched according to the result from the phonetic analysis modules and concatenated them by using proposed new phoneme concatenation algorithm. According to the experimental results, the system achieved the high level of intelligibility and acceptable level of naturalness. As the application area, it is intended for the resource limited device to use as language learning app and so on.

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

I. INTRODUCTION

The current progresses in speech processing domain are speech synthesis, speech recognition, speech analysis and coding. Text to speech is a system of speech synthesis that translates text into spoken speech output. Such systems were first developed for the purpose for aiding the visually and speech impaired person by allowing them to listen a computer-generated spoken voice that would read text to the user. Therefore, TTS system for different languages have been proposed by using different approaches. Later, the technological trends are focused on resource limited devices and They may be commercially available for different languages. However, there are little research works concerned with speech processing on Myanmar language especially for the resource limited devices although TTS system for Myanmar language already developed. Actually, there are some problem statements in the development of TTS system for Myanmar language. Myanmar language is written with no space or punctuation as English. Therefore, there may be some problem in text pre-processing for text analysis module. It will affect to speech synthesis part if we cannot correctly separate syllable and word. Moreover, Myanmar language is a tonal language which is different from others languages. Four tones in Myanmar pronunciation can generate more variety of sound so that the size of speech database become larger than other languages. Additionally, Myanmar language is complicated not only in the number of basic sounds but also in phonetic in nature. Some consonant sound has schwa vowel sound so that correct grapheme to phoneme conversion and control the schwa insertion and generate the half sound of the corresponding consonant is one of the challenging tasks in the developing of MTTS system. Therefore, this paper proposed the extended phoneme concatenation algorithm that considered the half sound of the consonant and the way to record and segment such sound.

The rest of this paper is organized as follow: the state of the art related with Myanmar TTS system is described in Section II. The nature of the language intended for this system is introduced in Section III. The brief explanation of Myanmar TTS system is presented in Section IV. The detail explanation of Myanmar phoneme inventory is presented in Section V. Then, the main contribution of this paper, phoneme concatenation method is described in Section VI. The error analysis of the system is discussed in Section VII. The experimental results of the proposed method are discussed in Section VIII. Finally, the paper is concluded in Section IX.

II. RELATED WORK

According to the model used for speech generation, speech synthesis systems are classified into three common types: articulatory synthesis, formant synthesis, and concatenative synthesis. Among them, concatenative synthesis is the most popular one because it can generate natural sound as a consequence of pre-recorded sound. The speech quality and the size of the system is trade off based on the different speech units for concatenation. The current speech units are word, syllable, phoneme, di-phone, tri-phone and so on. Many concatenative based TTS systems [1, 2, 3, 4, 5] have been implemented by using different speech units and they can generate the acceptable level of synthesized speech. A numerical TTS synthesis system for three languages: Marathi, Hindi and English languages is proposed by [6]. They used the approach that combined rule based approach and concatenation based approach. They used all utterances of sound units for concatenation and generation of speech signal. [7] presented TTS system by using concatenation method by using phoneme and diphone as speech unit for concatenation. They also compared based on these two speech units by using the two different methods of join costs calculation. According to their results, diphones speech units achieved more acceptable level from the subjective evaluations. In the research [8], they also used concatenative speech synthesis and part-syllable speech units was used to reduce the number of training sentence and the concatenation error for concatenation. Therefore, part-syllable transformation-based voice conversion (PST-VC) method was introduced which performed voice conversion with very limited data from a target speaker and simultaneously reduces concatenation error.

For Myanmar language, there has been 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. [9] designed rule-based MTTS system in which fundamental speech units are demi-syllables with 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 similar robotic voice in naturalness. Therefore, di-phone based MTTS is developed by [10]. They used concatenative synthesis method and time domain pitch synchronous overlap-add (TD-PSOLA) for smoothing concatenation points. They needed 8000 diphones pairs for 500 Myanmar sentences. The speech units are too much for the intension of resource limited devices. Moreover, if the required di-phone pair does not contain in the created database, the system results degraded. Later, the statistical machine learning approaches are considered as the approach for text to speech conversion. So, the authors in [11] proposed the statistical Myanmar speech synthesizer. They intended for resource limited devices but they need 4000 sentences and about 4 hrs recording time. These training sentences may be considered for resource limited devices. Therefore, [12] proposed new phoneme concatenative method for MTTS system. Their system is suitable for resource limited device because their phoneme speech database contained only 133 phonemes that can speech out for all Myanmar texts. According to their result, they also got the acceptable level for the intelligibility test but still need naturalness. Consequently, in their method, they did not consider the half sound of consonant. In Myanmar language, most of the minor consonants stand with schwa vowel. These kinds of consonant have half-sound of original one. In the previous method [12], 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 proposed phoneme concatenation method that is extended the previous one. In our method, we also considered the half-sound consonant for which we have to prepare the texts for recording. After that, segment and label the recorded sounds to get half-sound of consonants. Then, fetch the appropriate speech files from the created speech database and concatenate them by using proposed new phoneme concatenation algorithm. This paper mainly focuses on building phoneme speech database and proposed extended phoneme concatenative method. Then, they are tested on the already developed phoneme based MTTS system [12].

III. 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 and 9 basic vowels, 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. A syllable onset is the consonant or consonant cluster that appears before the vowel of a syllable. So, the 34 consonant letters indicate as the initial consonant of a syllable and Myanmar script has four basis medials diacritics to indicate additional consonants in the onset. Like other abugidas, including the other members of the Brahmic family, vowels are indicated in Myanmar script by diacritics, 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 [13].

IV. Myanmar Text to Speech System

To develop phoneme-based MTTS system, there are four main modules. They are text analysis, phonetic analysis, prosodic analysis and speech synthesis modules respectively. In the MTTS system, firstly, Myanmar sentences are tokenized by using rule-based tokenizer [14] for the next processing. Typically, the input text can contain non-standard words (NSW) such as number, abbreviations and others than simple texts. These NSWs are transformed into readable form by using rules based on regular expressions. Then, the normalized texts are converted into their corresponding phonetic symbols in the second module that is also called grapheme-to-phoneme conversion (G2P). The quality of TTS system also depends on the right of G2P conversion. In Myanmar language, actually, the words in writing and speaking are sometime different. This problem is solved by using phonological rules to get the correct G2P conversion. After that, in the prosodic module, assigning pauses are considered to get the more natural speech output. Finally, in the speech synthesis modules, the corresponding speech files of phoneme sequences from the previous module are fetched from the created phoneme speech database for concatenation. The brief system overview of MTTS system is described in Figure 1. Besides, in this paper, the detail explanation of how to created phoneme speech database and proposed phoneme concatenation method is presented in the next section.

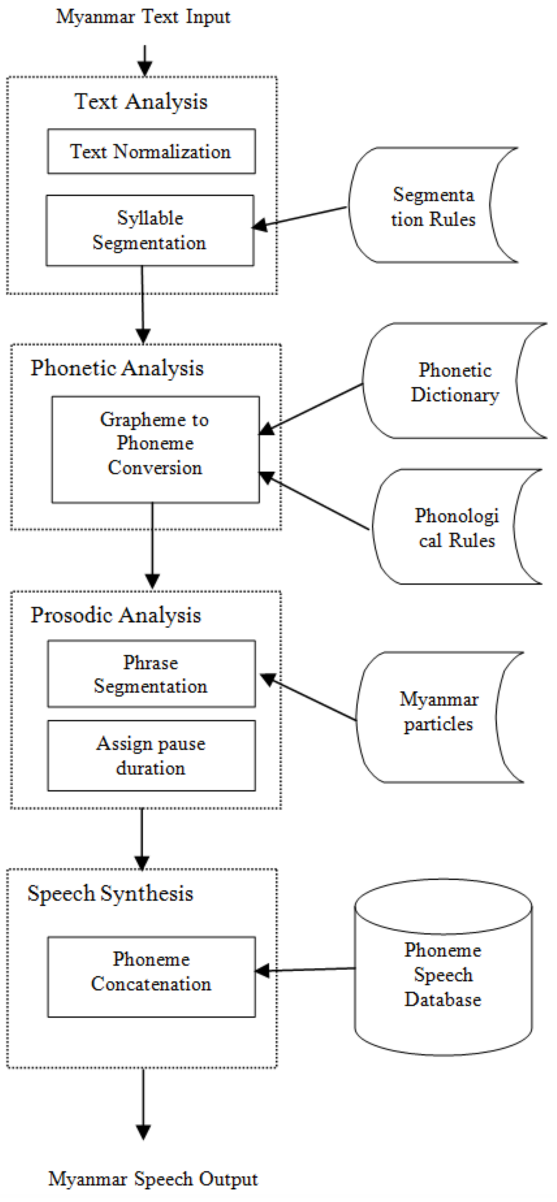


Fig.1. Myanmar Text-to-Speech System Design

V. Building Phoneme Speech Database

In the process of speech synthesis module, required speech units are fetched from speech database, concatenated and finally processed suitably to obtain high quality speech output. Hence building the speech database is one of the most important parts in concatenative synthesis-based TTS systems. The selected speech units are recorded before the system is executed. The recording may be different based on the selected speech units for concatenation. The speech units may be phoneme, syllable, word, di-phone and tri-phone. If syllables are concatenated, the necessary syllable should be recorded. In this work, the smallest speech unit, necessary phonemes are recorded and stored in the speech database. Generally, there are five steps in building a phoneme database for our system:

- Create a phoneme inventory
- Choose a speaker
- Prepare text sentences for the speaker and record each phoneme
- Segment and label the phoneme
- Store the phoneme

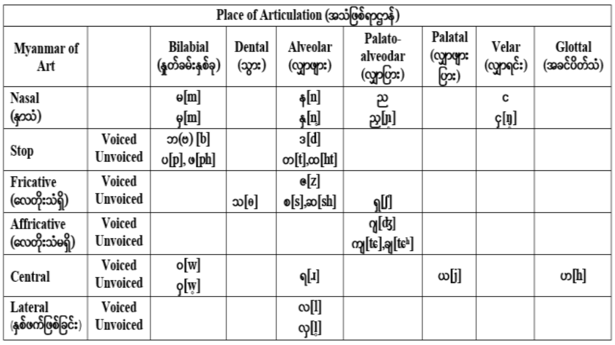
A. Myanamr Phoneme Inventory

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 [15]. 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 [16]. Myanmar phoneme that are used in this system are counted in detail in the following sub sections.

a) Consonatn Phoneme:

In Myanmar language, there are 34 basic consonants and that can be categorized as nasal, stop, fricative, affricate, central and lateral. The central /ɹ/ is occasionally used in place names that have preserved Sanskrit or Pali pronunciations. These 34 consonants are represented by 26 phonemes since some consonantal letters represents the same phonemes. For example, the consonants /ဂ/ and /ဃ/ represent the same phoneme/g/, the consonant /ဒ/ and /ဓ/ represent the same phoneme /d/. The list of Myanmar consonantal letters and their corresponding phoneme symbols in International Phonetic Alphabets (IPA) classified with in the place and manner of articulation are as shown in Table 1.

Table 1. Consonant Phoneme



Place of Articulation (အသံဖြစ်ရာဌာန)								
Myanmar of Art		Bilabial (နှုတ်ခမ်းနှစ်ခု)	Dental (သွား)	Alveolar (လျှာဖျား)	Palato-alveodar (လျှာပြား)	Palatal (လျှာဖျားပြား)	Velar (လျှာရင်း)	Glottal (အခင်ပိတ်သံ)
Nasal (နှာသံ)		မ[m] မှ[m̥]		န[n] နှ[n̥]	ည ညှ[ɲ̥]		င ငှ[ŋ̊]	
Stop	Voiced Unvoiced	ဘ(ဗ) [b] ပ[p], ဖ[ph]		ဒ[d] တ[t],ထ[ht]				
Fricative (လေတိုးသံရှိ)	Voiced Unvoiced		သ[θ]	ဇ[z] စ[s],ဆ[sh]	ရှ[ʃ]			
Affricative (လေတိုးသံမရှိ)	Voiced Unvoiced				ဂျ[dʒ] ကျ[tɕ],ချ[tɕʰ]			
Central	Voiced Unvoiced	ဝ[w] ဝှ[w̥]		ရ[ɹ]		ယ[j]		ဟ[h]
Lateral (နှစ်ဖက်ဖြစ်ခြင်း)	Voiced Unvoiced			လ[l] လှ[l̥]				

Moreover, there are 4 basic medial (M) and 7 combined medials. The above 34 consonant letters (C) may be modified by one or more medial diacritics (three at most), indicating an additional consonant before the vowel. These diacritics are: Ya pin (ျ), Ya yit (ြ), Wa hswe (ွ) and Ha hto (ှ) indicated by /j/, /j/, /w/ and /h/ respectively. The first two has the same pronunciation. Therefore, the 10 medials can modify the 34 basic consonants (10 * 34 = 340). For example - မ (C) + ြ (M) = မြ(emerald), မ (C)+ ွ (M) =မွ (bleary) and မ (C)+ ှ(M) = မှ (from). However, these medials cannot combine all of the consonants. For instance: there is no combination for က(C) + ျ (M) + ွ (M) + ှ (M) = ကျွှ် that cannot be pronounced. Therefore, after final counting the syllable that can be pronounced by combining with medials are only 47 syllables in Myanmar scripts.

b) Vowel Phoneme:

There are 9 basic vowels in Myanmar language. They are *a, i, e, u, o, ai, ei, au, ou*. The seven vowels can stand itself except *ai* and *au*. They can only stand when there is a Aset (်) behind them such as အိုက်-/aiʔ/, အိုင်-/ain/,အောက်-/auʔ/, အောင်-/aun/. There are two kinds of vowel: open vowel and close vowel. The monophthongs /e/, /o/, /ə/, and /ɔ/ occur only in open vowel (those without a syllable coda); the diphthongs /ei/, /ou/, /ai/, and /au/ occur only in closed vowel (those with a syllable coda). /ə/ only occurs in a minor syllable or consonant, and is the only vowel that is permitted in a minor syllable. The vowel and its phonetic sign defined by IPA is described in Table 2.

Table 2. Vowel Phoneme

	Monophthongs			Diphthongs	
	Front	Central	Back	Front offglide	Back offglide
Close	i		u		
Close-mid	e	ə	o	ei	ou
Open-mid	ɛ		ɔ		
Open		a		ai	au

Myanmar language is a tonal language and there are four tone levels. They are described as à, aˆ, a¯, aʔ for the vowel "/a/". According to these tone levels, basic vowels can be expanded into 50 vowels sounds that can be divided into 21 nasalized vowels, 21 non-nasalized vowels and 8 glottal stop vowels. These vowels play an important role to construct syllable and to yield Myanmar speech. These 50 vowels can make any syllables and any speech sound by multiplying consonants and vowels. For instance – the consonant "က-/k/" can be expanded as "က (dance)"- /k à/ or /kə/, "ကား (car)" - /k aˆ/, "ကာ (cover)" - /ka a¯/ and "ကတ် (card)" - /k aʔ/.

B. Choose a speaker

For any speech-based system, it is important to record the best sound quality possible since each minor distortion can often occur in complex speech. Therefore, choice of the right voice talent for recording is a crucial aspect. The voice talent should be made familiar with the texts in advance. Consistent and steady recording has to be ensured. The speech quality depends upon the quality of the recorded sound so a professional Myanmar female speaker is selected for recording.

C. Preparing text for recording

Then, prepare the texts of the phonemes (consonant, vowel , combined medial and so on) to be recorded. 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 [k**ə** leˆ], [kʰ**ə** ɹe ¯], [m**ə** n ɛˀ pʰj ʌ̃¯] instead of [k **à** leˆ], [kʰ **à** ɹ e ¯], [m **à** nɛˀ pʰjʌ̃ ¯]. Therefore, mostly, if the basic consonant is situated in the first position of the word, it is pronounced as schwa. Consequently, select the words that contains the minor consonant combined with schwa vowel such as "ခရေပန်း - [kʰə ɹe¯ bʌ̃ˆ] (star flower)" in which "ခ-/ kʰə /" is minor syllalbe with schwa. In this case, the right choice of text is also important because the half sound of voice and unvoiced consonants are different.

The selected phonemes and words are recoded with 44100HZ sampling rate, 16-bit, mono quality format by the selected person. The recording has been done in the LA Studio, Mandalay and it took only two hours. If it is compared with other speech recording time, we found that it is significantly reduced than others.

D. Segment and label phoneme

Finally, the recorded sound files are segmented and labeled for the next processing. For this purpose, the sound editing software "*WavePad Sound Editor*" has been used. The phoneme sounds have been labeled manually one by one, after carefully listening and analyzing the recorded sounds and then these are stored by using the transliterated the phoneme name itself. For example, the sound file of "က" /k/ is named "K.wav" and AA3.wav for အား-/aˆ/.

For the half-sound consonant, from the recorded word (ခရေ *(star flower))* according to the above example word, only the very first syllable ("ခ") is segmented and label as shown in the Figure 2. Then, it is named as KHH.wav which mean the original consonant name (KH) plus H (half sound). Example segmentation to get half sound consonant is shown in the following Figure. 1. Therefore, the recorded word sounds are segmented in this way to get extra 24 minor consonants sounds.

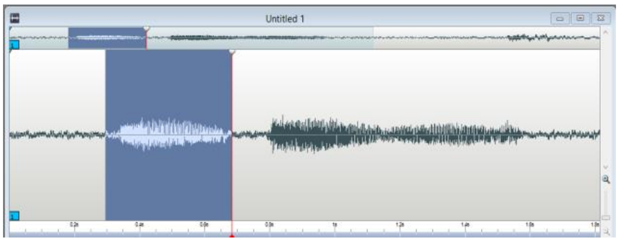


Fig.2. Half-sound of (□-k^h) cutting from the word "ခရေ"

E. Store the phoneme

In this step, the segmeted speech wave files are stored in the Myanmar phoneme speech database. At first, there are only 133 phoneme speech sounds in [12] that did not consider half-sounds. Now, we consider this kind of sound so that in this database, there are 157 phoneme speech files by adding extra 24 minor consonants as shown in Table 3. These phonemes can make sound for all kind of Myanmar texts. Therefore, this amount of phoneme is significally reduced than the previous di-phone based MTTS system so that it is very suitable for resourced limitted devices.

Table 3. Number of sound file used in this system

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**

VI. Proposed Phoneme Concatenation Method

The concatenative speech synthesis is the most usable method in the field of TTS system and it can generate high quality natural speech because of selecting the pre - recorded speech segments [17]. In our proposed system, we used concatenative method and phoneme is selected as basic units for concatenation. Phoneme can be defined as a minimum sound unit of a language by which the meaning may be differentiated. It is a unit of speech made up of vowels and consonants. The inventory of basic units is the smallest compared to other. Therefore, using phoneme gives maximum flexibility for TTS system for resource limited devices.

Typically, Myanmar language is syllabic language and thus its spoken sentence form is based on the syllable that is combination of consonant phoneme (CP) and vowel phonemes (VP). For example –the sound of syllable "ကျောင်း(school)-/tɕaũˆ/" come from the combination of consonant phoneme as "ကျ-/tɕ/" (KYA) and vowel phoneme "အောင်း-/aũˆ/" (AUNG:). Therefore, every sound in Myanmar language is made up of these phonemes combinations.

Phoneme based Myanmar TTS system has been developed in [12]. In their paper, the syllable that has half sound did not consider. Therefore, we extend the previous phoneme concatenation method as shown in Algorithm 1.

Algorithm 1: Phoneme Concatenation Method

```
Input: consonant phoneme and vowel phoneme ($CP_i$, $VP_i$)
Output: concatenated voice file $Voice = \{Voice_1, Voice_2, \dots\}$
Begin

CP: a set of consonant phoneme
VP: a set of vowel phoneme
$WF_1$: modified consonant file
$WF_2$: modified vowel file
C_SP: start position of consonant wave file
C_EP: end position of consonant wave file
V_SP: start position of vowel wave file
V_EP: end position of vowel wave file
S: a set of concatenated wave files
procedure phoneme_ concatenation ($CP_i$, $VP_i$) {
$C_i$_SP ← 0
$C_i$_EP ← duration of $C_i$ * threshold value
if ($V_i$. equals ("ə")) then
    $WF_1$ ← fetch the half sound file of CPi
else if ( $V_i$. equals (" ə̆")) then
    $WF_1$ ← fetch the sound file of CPi
else
{
   if ($C_i$ is aspirated consonant) then
      $WF_1$ ← trim ($C_i$_SP, $C_i$_EP)
  else
      $WF_1$ ← trim ($C_i$_SP, $C_i$_EP)
}
end if
$V_i$_SP ← duration of $V_i$ * threshold value;
$V_i$_EP ← ($V_i$ duration -  $V_i$_SP) - 1
if (duration of $VP_i$ is zero) then
    $WF_2$ ← fetch the sound file of $V_i$
else
$WF_2$ ← trim ($V_i$_SP, $V_i$_EP)
$Voice_i$ ← create voice file by concatenating $WF_1$ and $WF_2$
predicting pause for each .mp3 file by using duration
        modeling
Voice ← Voice ∪ $Voice_i$
End
```

According to the algorithm, firstly, the syllable phoneme is separated into consonant and vowel respectively such as the syllable phoneme "tɕaũˆ" into consonant and vowel phoneme "/tɕ/(KYA) and /aũˆ/(AW2)" respectively. In this case, if these two phonemes are directly concatenated, we get "KYA-AUNG" instead of our desired sound "KYAUNG". Therefore, we proposed new phoneme concatenation algorithm for Myanmar language to get our desired syllable speech output.

For the input phoneme sequence, the corresponding speech *.wav* files, according to the example, the KYA.wav and AW2.wav are fetched from the created phoneme speech database. Then, the consonant phoneme (*CP*) and vowel phoneme (*VP*) are modified based on their duration by defining threshold values for start and end position.

For *CP*, firstly, set the start position into “0” (*Ci_SP* = 0) and then generate the duration of this *CP* and multiply with the threshold value for setting the end position (*Ci_EP* = duration of *VP* * threshold). The threshold may be different based on the consonant. If the consonant is aspirated consonant such as “ဖ, ဆ, ထ, ရ, ယ”, the threshold value is set to 0.27, otherwise, it is set to 0.50. However, if the consonant phoneme contains half-sound vowel sign ("ə"), fetch the related half-sound voice file *(VF_H$_i$)* and do not modify anything as well as the vowel is “ ä̀” than fetch their corresponding vowel file *(VF$_i$)*.

For the *VP*, set the start position by multiply the duration of *VP* and threshold value (0.50) and then *V_SP* is subtracted from *VP*’s duration for the end position. These phonemes are modified according to the respective start and end position. Then, these two wave files are concatenated so that we can get our desired speech output for syllable that can make word, phrase and up to sentences. The extended phoneme based MTTS system outperform than the previous one especially for the half sound consonant so that it achieves the high level of intelligibility but acceptable level of naturalness TTS.

VII. Evaluation Method

The proposed new phoneme concatenation method is tested in the already developed phoneme based MTTS system [12]. The two-quality measures are considered for testing the quality of MTTS system. They are intelligibility test which measures how much the user understand what the speech out is and the naturalness test which measure how much the system output is similar to the real human speech. The evaluation can also be made at several levels, such as phoneme, word, or sentence level, depending on what kind of information is needed. The evaluation process is usually done by subjective listening tests. For MTTS testing, the 15 evaluators who are from the students from the University of Computer Studies, Mandalay (UCSM) help and test the speech quality of MTTS system for the 100 sentences that are from any fields such as greeting sentences, asking time and so on.

A. Naturalness Test

For the naturalness test, we used the most popular methods for evaluation: mean opinion score (MOS). The MOS method is the most useful and simplest method to test quality of MTTS system. It has five level scales in MOS: bad (1), poor (2), fair (3), good (4) and excellent (5) [18].

Formula for calculation of MOS is as follow equation (1) and the symbol in the equation are, S_1 = score of i^{th} evaluator, N = number of evaluators, M = number of Sentences, j = sentence index.

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

a) MOS for Naturalness Test –

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, 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. The average MOS score for testing quality of naturalness is 2.91 as shown in Figure 3.

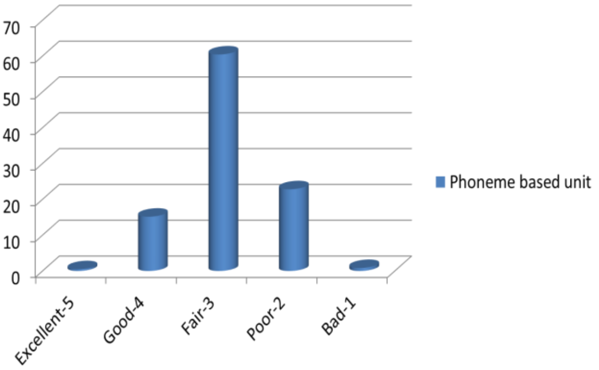


Fig.3. MOS value of naturalness test

B. Intelligibility Test

Intelligibility is one of the important factors affecting speech quality. We can calculate intelligibility either by MOS as in Equation 1 and word error rate (WER).

a) MOS for 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, `18% of the subjective understood very well. 34.87% did understand the voice very much 38.23% neither much nor little and another 6.16% understood a little and only 0.28 did not understand very well as shown in the Figure 4. The average MOS score for intelligibility test is 3.58.

b) WER for 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. Acording to their results, we calculated WER by using the equaion (2) and the WER value is only 7% in intelligibility test.

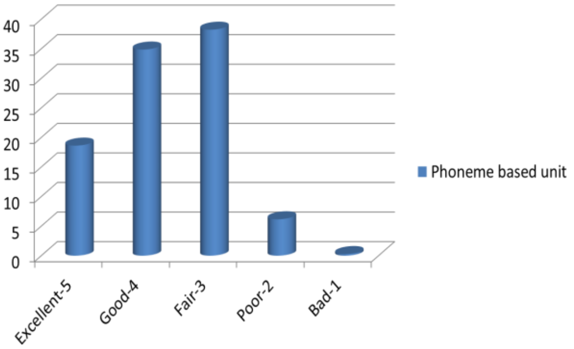


Fig.4. MOS value of Intelligibility test

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

In the application point of view, our proposed phoneme concatenative is 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 with this extended method, the size is little increase than the previous one like from 7.5 MB to 9.79MB. However, it is not significant and it is still suitable for the resource limited devices.

VIII. Discussion

Actually, the Myanmar speech could be synthesized by using the defined 133 phoneme speech units with proposed phoneme concatenation method. In this case, the half sound did not consider. So, this paper extended the previous method to get this kind of half sound. Therefore, final MOS score for intelligibility test is increased up to 3.58 and The MOS score of naturalness test is also increased up to 2.91. According to the evaluation results, the system got high MOS score for intelligibility and acceptable level of naturalness. Moreover, the vowel plays important role in the case of generation of Myanmar speech. When the generated speech output files are analyzed, some vowels combination cannot generate the clear voice, especially, the non-nasalized vowel such as "အေ၊အေး၊အေ့၊အို၊အိုး၊အို့၊အယ်၊အဲ့၊အယ်". If a consonant is combined with such vowels, the vowel sound is influence over consonant sound so that it cannot get the desired speech output. For instance, the sound for the syllable "မေး၊ပေး၊တို့၊ကယ် " is generated as "AYE, PAY, OOE, AEL" as the vowel sound instead of "MAY, PAY, TOE, KAL". Therefore, in the future work, we will consider to get the more quality speech file for such vowel by considering the extra speech units.

IX. Conclusion

This paper presented the proposed extended phoneme concatenation method for Myanmar text to speech system by considering half sound of the consonant. Phoneme is used as a basic unit for concatenation. Therefore, detail explanation of phoneme inventory is presented. Then, this paper described phoneme speech database and it contains only 157 phoneme speech units that can speech out for all Myanmar text. Therefore, it is support MTTS system for resource limited devices. According to the experimental result, we achieved the high intelligibility and naturalness result than the previous one. To get more natural output speech, we are considering to used signal processing techniques.

Acknowledgment

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

References

[1] 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.

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

[3] 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.

[4] 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.

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

[6] 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

[7] P. Kasparaitis. and K. KANČYS., "Phoneme vs. Diphone in Unit Selection TTS of Lithuanian", BALTIC JOURNAL OF MODERN COMPUTING, 6(2), pp.162-172, 2018.

[8] M.J. Jannati, and A. Sayadiyan, A, "Part-Syllable Transformation-Based Voice Conversion with Very Limited Training Data", Circuits, Systems, and Signal Processing, 37(5), pp.1935-1957, 2018.

[9] 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.

[10] 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.

[11] Y. K. Thu, W.P. Pa, Ni, J., Shiga, Y., Finch, A., Hori, C., Kawai, H. and Sumita, E. HMM based Myanmar text to speech system. In Sixteenth Annual Conference of the International Speech Communication Association. 2015.

[12] 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

[13] Myanmar language commission, Myanmar grammar. 30th year special edition. University Press, Yangon, Myanmar;

2005
[14] 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.
[15] R. Lass. English Phonology and Phonological Theory: Sincronic and Diacronic Studies. Cambridge: Cambridge University Press. 1976.
[16] T. Tun, “Acoustic phonetics and phonology of the Myanmar language”, School of Human Communication Sciences, La Trobe University, Melbourne, Australia, 2007.
[17] M. Karjalainen. “Review of Speech Synthesis Technology”. Helsinki University of Technology, Department of Electrical and Communications Engineering. 1999 Mar 30.
[18] Text to speech testing strategy, Version 2.1”, Technology Development for Indian Languages Programme DeitY, 07 July, 2014

Authors’ Profiles

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 Bachelor of Computer Science and Master of Computer Science from 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.

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.

How to cite this paper: Chaw Su Hlaing, Aye Thida, "Phoneme Concatenation Method for Myanmar Speech Synthesis System", International Journal of Information Technology and Computer Science(IJITCS), Vol.10, No.12, pp.13-20, 2018. DOI: 10.5815/ijitcs.2018.12.02