Research Article

International Journal of Advanced Computer Research, Vol 8(34)
ISSN (Print): 2249-7277 ISSN (Online): 2277-7970
http://dx.doi.org/10.19101/IJACR.2017.733036

Phoneme based Myanmar text to speech system

Chaw Su Hlaing* **and Aye Thida**
Artificial Intelligence Lab, Faculty of Computer Science, University of Computer Studies, Mandalay, Myanmar

Received: 28-September-2017; Revised: 27-December-2017; Accepted: 03-January-2018

Abstract
The text to speech (TTS) is one of the recommended research level topics in the domain of natural language processing and speech processing. In this day and age, the usage of mobile phones is extremely increasing so the researchers focus on speech processing on mobile devices. TTS system for mobile phones is difficult to implement as they have limited storage capacity and computing performance. Therefore, phoneme based Myanmar TTS (MTTS) system is proposed for resource limited devices. In this paper, rule based Myanmar number conversion and new phonological rules are proposed. For speech generation, firstly, phoneme speech database in which there are only 133 phoneme units is created and then the new phoneme concatenation algorithm is applied. Moreover, each module of MTTS system is presented in detail with their respective experimental results and the system achieved the acceptable level of intelligibility although naturalness is still needed to achieve the satisfactory level according to these results.

Keywords
Text to speech, Myanmar language, Phoneme, Concatenative speech synthesis.

1.Introduction
The Language is the most important attribute that distinguish humans from other living in which speech is the main key of the language. For the past decades, speech technology has taken the largest part of the research trend. Text-to-speech (TTS) system is one of the research level topics in the domain of natural language processing and speech processing. TTS is simply defined as just written text input is converted into speech output and it is to be used mainly by language learners, disabled or impaired users and finally students. As the state of the art, challenges to control the quality of voice of synthesized speech have stayed alive for more than a decade now. Significant research progress in this field has contributed in conversion of text to speech for languages like English, Japanese, and Chinese by using different speech synthesis techniques but not much work has been done in TTS for Myanmar languages. Nowadays, the research trend of speech processing focus on resourced limited devices. The high quality TTS system can be achieved by using concatenative method. However, they are not always suited for implementation on resource limited devices because they require rather large recorded speech databases.

The former Myanmar TTS systems have been developed and that can generate reasonably well enough synthesized output quality on personal computer. However, there has still been a great effort in the development of text to speech in Myanmar language and nobody has implemented Myanmar TTS system for mobile phones. TTS system for resource limited device is difficult to implement as they have limited storage capacity and computing performance. Therefore, the goal of this study is to develop Myanmar speech synthesizer that can produce an acceptable quality of synthesized output in almost real time on a mobile device by using concatenative speech synthesis technique. In our proposed system, the phoneme is used as basic speech unit for concatenation and the phoneme segmentation and concatenation for speech generation is proposed. Our proposed method achieved the acceptable level of intelligibility, but still need to get very natural sound that are two quality measures for any TTS systems.

The rest of this paper is organized as follows: the previous research work of TTS system is discussed in section 2. Section 3 presented the proposed phoneme based Myanmar TTS system and explained in detail the proposed Myanmar number converter and phonological rules and how to build a phoneme speech database. The phoneme concatenation method

*Author for correspondence

is also discussed in this section. Then the evaluation for each module is also presented. Finally the paper concluded in section 4.

2.Related work

At the state of the research in speech synthesis techniques, the concatenative speech synthesis is one of the most popular methods and it can generate the best natural speech output. Many TTS systems proposed by [1-5] have been implemented by using concatenative method based on different speech units and they can generate high quality synthesized speech. 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. Rule based Myanmar TTS was proposed [6]. They used semi-syllable is used as the basic unit. They conducted their intelligibility tests and found that their system had a syllable correctness rate of over 90%. Diphone based Myanmar TTS system was proposed by [7]. They used concatenative synthesis method and time domain pitch synchronous overlap-add (TD-PSOLA) for smoothing concatenation points. They achieved an acceptable level of intelligibility and but still needed naturalness. Moreover, their approach is not suitable for resource limited devices because of their diphone database which includes over 8000 diphones for 500 Myanmar sentences. They evaluated their system with different pitch marks of hanning windows for quality of Myanmar TTS.

Moreover, in this day and age, the researchers focus on speech synthesis for mobile devices. However, there is still a challenge for researchers to integrate the TTS system on mobile devices because of its storage capacity and processing power. However, android based TTS systems for some languages have been developed by using different methods in concatinative synthesis. They have an level of intelligibility and naturalness dimension. In 2009, Wongpatikaseree et al. [8] developed a Thai speech synthesizer for a mobile device that can be used in real time. It was designed based on Flite, an open source synthesis library designed by a small and fast runtime TTS engine. The authors modified the Flite to work for complicated Thai language. They used di-phone speech unit that was transformed from a unit selection database to be convenient for resource limited devices. Their system is much faster than pTalk, however may not be good in speech quality as the output from pTalk. In 2010, Wongpatikaseree et al. [9] presented Thai text-to-speech system for mobile devices. Their main focus is selection of the most suitable speech units for Flite_Thai in order to improve its indelibility. They created a speech corpus that contains demi-syllable, diphone and hybrid diphone unit. They compared MOS value in intelligibility test of the three speech units. Their selected units, hybrid diphone unit obtained the highest score. In 2017, Mokgonyane et al. [10] discussed the development of a mathematical computer-assisted learning mobile application that integrates a text to speech synthesis module for South African low resourced languages, initially targeting the Sepedi language. According to their saying, 80% of participants were impressed by the developed mobile phone application. In 2009, Karabetsos et al. [11] presented efficient techniques that successfully address the challenging problems arising in embedded environments, such as database reduction, database compression by using an algorithm which leads to small footprint speech database, and a vector quantization approach was used for the spectral join cost calculation that significantly reduces the computational requirements of the unit selection module. They said that their results provide clear evidence of substantial improvement in the computational resources exploitation. In 2013, Gopi et al. [12] developed a TTS system of Malayalam language for android based mobile phones by using concatenative synthesis and diphone like segments (partneme) as the basic units for concatenation. Among the speech joint smoothing methods, the used epoch synchronous non-overlap and add (ESNOLA) method was used for smoothing the concatenation point. They achieved the acceptable level of speech quality with the MOS is 3.2.

3.Myanmar language

The Myanmar language is the official language and it is spoken as a first language by 32 million, primarily the Myanmar people and related sub-ethnic groups, and as a second language by 10 million, particularly ethnic minorities in Myanmar and neighboring countries like the Mon. It is classified into two categories: formal and colloquial. Formal is used in literary works, official publications, radio broadcasts, and formal speeches and colloquial is used in daily conversion and spoken. It is written from left to right and requires no spaces between words, although modern writing usually contains spaces after each clause to enhance readability. The Myanmar alphabet consists of 34 consonants but some has same pronunciations so that there are only 21 phonemes symbol (/k/, /kh/, /g/, /ŋ/, /s/, /sh/, /z/, /ɲ/, /t/, /th/, /d/,

/n/, /p/, /ph/, /b/, /m/, /j/, /l/, /w/, /θ/, /h/). There are eight basic vowel phonemes (/a/, /i/, /u/, /e/, /o/, /ɛ/, /ɔ/) in Myanmar, and these can be extended into 50 vowels according to tone level. Because, Myanmar language is a tonal language and there are 4 tone levels, which means phonemic contrasts can be made on the basis of the tone of a vowel. These contrasts involve not only pitch, but also phonation, intensity (loudness), duration, and vowel quality [13].

4.Architecture of MTTS system

MTTS system is composed of four main modules. They are (i) text analysis (ii) phonetic analysis, (iii) prosodic analysis and (iv) speech synthesis. In the text analysis module, the non-standard words such as numbers, abbreviations are normalized into readable form and syllable segmentation and irregular words normalization are also include in this section. Then, in the phonetic analysis module, the normalized texts are converted into their respective phoneme sequence in which we used Myanmar phonetic dictionary and apply the proposed phonological rules to achieve the correct phoneme sequence that can cause the high quality TTS system. After that, in the section of prosodic analysis, we consider to detect phrase boundary of Myanmar sentence to assign pause duration to get natural sounding Myanmar speech. In speech synthesis module, firstly, we create phoneme speech database that is most suitable for resource limited devices. Then the system generates the Myanmar speech output by using phoneme concatenative algorithm. The overall system architecture is described in *Figure 1* and the proposed phoneme based Myanmar TTS system is shown in algorithm1.

4.1Text analysis

The very first and most important module for any TTS system is text analysis. Only if the accurate result is achieved in this step, the higher quality of TTS system will be attained. In order to generate a phonemic internal representation, raw text first needs to be pre-processed or normalized in a variety of ways so that the input Myanmar texts are segmented into syllable and word for next processing. There are three main steps in text analysis for MTTS system. They are (i) loan word conversion, (ii) abbreviation expansion and (iii) number conversion.

In Myanmar language, some English words are directly used in the sentence. These words are mostly confined to loan words and they are speech out as the English pronunciation. Therefore, firstly, they are needed to detect from the sentence and then convert into Myanmar text to be able to read smoothly. Consequently, loan word dictionary is created that contains around 200 loan words and it can also be modified for new one. For instance – computer→ ကွန်ပျူတာ and camera → ကင်မရာ. Moreover, in TTS system, the input sentence may contain abbreviation so that they are needed to be expanded into readable form. Therefore, the abbreviation expansion is also considered in MTTS system.

For example-

- သံချောင်း (ဝါ) မောင်မြ→သံချောင်း ခေါ် မောင်မြ
  (Than chaung (alias) Mg Mya)
- စံပယ် (သို့) နှင်းဆီ→ စံပယ် သို့မဟုတ် နှင်းဆီ
  (jasmine (or) rose)

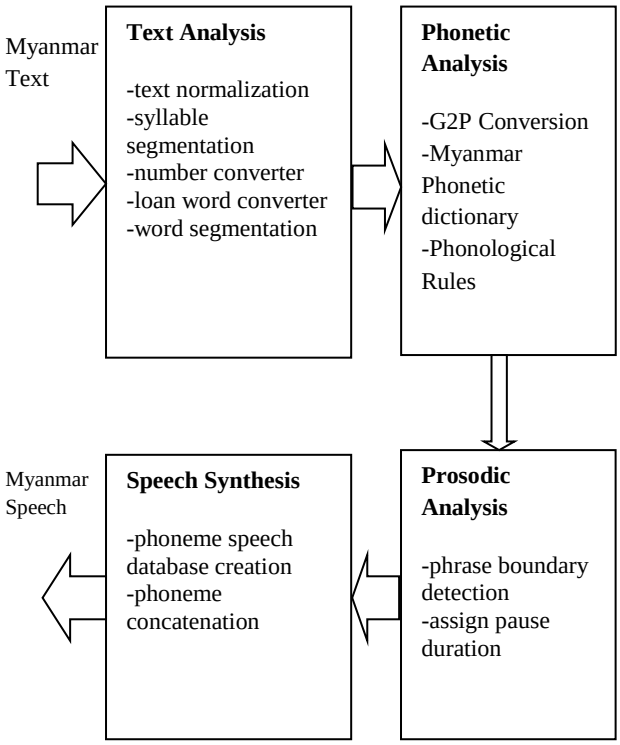

Figure 1 Block diagram of phoneme based Myanmar TTS system

4.1.1Myanmar number normalization

In TTS system, the input text can contain non-standard word (NSW) such as numbers, date, currency amounts etc. If normalization of these NSW is not performed in the early step, the quality of the system may be degraded. Therefore, Myanmar numbers are converted into readable word form by using rules based on regular expression. Although it seems easy to be handled Myanmar number, there are some ambiguous cases. However, in this system, nearly all kind of number formats such as decimal number, date, time, NRC number, phone number,

account number, ratio, and fraction can be converted into full form based on rules by considering their contextual information.

Some type of NSW is described in *Table 1*. Therefore rule based number normalization is implemented by using regular expression (RE).

Table 1 The sample type of NSW

Type	Example		
	English	**Myanmar**	**Myanmar Expansion**
Date	• 30-4-2018 • 6/3/2017 • May 3, 2017 • 2018-2019 Academic Year	• ၃၀-၄-၂၀၁၈ • ၆/၃/၂၀၁၇ • မေလ ၃ ရက် ၂၀၁၇ • ၂၀၁၈-၂၀၁၉ ပညာ သင်နှစ်	• သုံးဆယ်ရက်၊လေးလ၊နှစ်ထောင့်ဆယ့်ရှစ်ခုနှစ် • ခြောက်ရက်၊သုံးလ၊နှစ်ထောင့်ဆယ့်ခွန်ခုနှစ် • မေလသုံးရက် ၊နှစ်ထောင့်ဆယ့်ခွန်ခုနှစ် • နှစ်ထောင့်ဆယ့်ရှစ်၊နှစ်ထောင့်ဆယ့်ကိုးပညာသင်နှစ်
Time	• 3:45 • 8AM to 11 AM	• ၃:၄၅ • နံနက် ၈:၀၀ နာရီ မှ • ၁၁:၀၀နာရီအထိ	• သုံးနာရီ လေးဆယ့်ငါးမိနစ် • နံနက်ရှစ်နာရီမှဆယ့်တစ်နာရီအထိ
Number (cardinal)	• 50 kyats • 4.13 • 60%	• ၅၀ကျပ် • ၄.၁၃ • ၆၀%	• ငါးဆယ်ကျပ် • လေးဒသမတစ်သုံး • ခြောက်ဆယ်ရာခိုင်နှုန်း
Number as Digit	• 09-402663944 • ATM-99944433373201	• ၀၉-၄၀၂၆၆၃၉၄၄ • ATM ၉၉၉၄၄၄၃၃၃၇၃၂၀၁	• သုံညကိုးလေးသုံညနှစ်ခြောက်ခြောက်သုံးကိုးလေးလေး • အေတီအမ်ကိုးကိုးကိုးလေးလေးလေးသုံးသုံးသုံးခုနှစ်သုံးသုံညတစ်
Money	• 123456 Kyats • $100	• ၁၂၃၄၅၆ကျပ် • ၁၀၀ ဒေါ်လာ	• တစ်သိန်းနှစ်သောင်းသုံးထောင့်လေးရာငါးဆယ့်ခြောက် • တစ်ရာဒေါ်လာ
Range	• 4-3 or 4:3 Goals • Miles (60-50)	• ၄-၃ဂိုး (သို့) ၄:၃ ဂိုး • မိုင် ၆၀-၅၀	• လေး ဂိုး သုံး ဂိုး • မိုင် ခြောက်ဆယ် ငါးဆယ်
Fraction	• 1/306 • 2/3(fraction	• ၁/၃၀၆ • ၂/၃	• တစ်မျဉ်းဖောင်းသုံးရာ့ခြောက် • သုံးပိုင်းနှစ်ပိုင်း
NRC Number	• 8/PaKhaKa (N) 225349	• ၈/ပခက (နိုင်)၂၂၅၃၄၉	• ရှစ်မျဉ်းဖောင်း ပခက နိုင် နှစ်နှစ်ငါးသုံးလေးကိုး

Algorithm 1: Proposed phoneme based Myanmar TTS

Input: Myanmar Text $t = \{s_1, s_2, \ldots\}$
Output: Voice Files $v = \{v_1, v_2, \ldots\}$

```
v ← φ
for each si ∈ t do
    si ← sentence normalization(si)
    segment si into syllables sy = { sy1, sy2, …}
    segment word based on syllable w= {w1, w2, …}
    for each word wi ∈ w do
      pi ← wi is converted into phonemes from phonetic table
       for each phoneme p ∈ pi do
          pi ←modify phoneme by using phonological rules
          phrase segmentation and assign pause duration
          separate consonant ci and vowel vi phoneme
          ai←change each phoneme into arperbet a = {a1, a2,.}
            fetch corresponding voice file from phoneme speech database
          execute procedure phoneme_concatenation (ci,vi)
          generate speech output vi
       end for
    end for
end for
```

The RE for date and time format is as follow:

```
private static boolean isDate( String input) {
return input.matches("(၀?[၁-၉]|[၁၂][၀-၉]|၃[၀၁])[၊\\/-](၀?[၁-၉] |၁ [၀၁ ၂ ] ) [၊\\/-]*((၁၉|၂၀)\\D\\D)");}
private static boolean isTime(String input{
return input.matches("^([၀-၉]|၀[၀-၉]|၁[၀-၉]|၂[၀-၃]+):[၀-၅]*[၀-၉]$");}
```

4.1.2Evaluation for text analysis

To evaluate the result of text analysis, we calculated bilingual evaluation understudy (BLEU) score – the currently most popular automatic evaluation metric, has an elegant solution to the role of word order, word error rate (WER)-employs the Levenshtein distance, which is defined as the minimum number of editing steps – insertions, deletions, and substitutions needed to match two sequences and finally accuracy of text analysis is calculated. For evaluation, test corpus that contains some Myanmar sentences in which numbers are mostly contained (total number of sentences: 560 and the number of syllable are 67980) from web are used. The evaluation results for text analysis are shown in the following *Table 2*.

4.1.3Finding and discussion for text analysis

All the loan words and abbreviations can be converted for this test corpus. Therefore, it achieved the 100% accuracy. In the loan word conversion, the system can convert only the word from the predetermined loan word dictionary that can be updated for the new load words. For the abbreviation expansion, the system does not consider the abbreviations like that "ဆ.သ.ရ ၊ ဗ.ဝ.တ" . In the number conversion, Myanmar number "၇-7" is pronounced in two ways. For example – "၂၀၁၇-2017" is pronounced နှစ်ထောင့်ဆယ့်ခွန်" and the number "၁၇-17", ဆယ့်ခွန်. However, when the year suffix is combined, "၂၀၁၇ခုနှစ်-2017 year" is pronounced "နှစ်ထောင့်ဆယ့်ခွန်နခုနှစ်" . For this case, the WER rate for number conversion may be reduced.

Table 2 The evaluation result for text analysis

Type	BLEU Score				WER (%)	Accuracy (%)
	1-gram	2-gram	3-gram	4-gram		
Loan Word Conversion	100	100	100	100	0	100
Abbreviation Expansion	100	100	100	100	0	100
Number Conversion	98.05	97.68	97.42	97.05	7	93
Text Analysis	**98.05**	**97.68**	**97.42**	**97.05**	**7**	**93**

4.2Grapheme to phoneme conversion

The next step in MTTS system is to take the normalized word strings from text analysis and produce a pronunciation for each word. This process is named as grapheme to phoneme conversion, G2P conversion. The most important component here is a large pronunciation dictionary. We used Myanmar pronunciation dictionary. However, this dictionary alone turn out to be insufficient, because running text always contains words that do not appear in the dictionary and some texts may have different pronunciation based on their part of speech and there surrounding words. To solve this problem, we used both dictionary based approach and rule based approach based on Myanmar phonological rules. The proposed phonological rules for G2P conversion are discussed in the next section.

4.2.1Proposed phonological rules

Phonology is the study of the system of sounds or phonemes in a language. Myanmar phonology is generally constructed by means of combining the consonant phonemes, vowel phonemes and tone. Phonological rules translate phonemes to the real sounds (phones) and predict how a speech sound will change depending on its position in various speech environments. Different languages have different rules and they have to follow a specific set of rules that determine how we sound when speaking. Myanmar language has many phonological rules so that they have complex phonological structure discussed by [14]. In this paper, we proposed new phonological rules based sentence level Myanmar G2P conversion to get the correct phoneme sequence for high quality Myanmar TTS system. In our proposed system, the former proposed rules [7] are also considered. The proposed rules are presented in the following section and grapheme to phoneme conversion results for some Myanmar sentence by using proposed rules are described in *Table 3*.

Table 3 Myanmar G2P conversion by using phonological rules

Input Sentence	G2P Result	
	Original	Applied rules
ရေစစ်ဖြင့်ရေစစ်ပါ ။ (Filter water with water filter.)	je˗ si˒ phjĭ´ je˗ si˒ pa˗	je˗ **zɪ˒** phjĭ´ je˗ **sɪ˒ ba˗** **(rule 1 and rule 3)**
ထမင်းချက်ဒေါ်မြထမင်းချက်နေသည် (The cooker Daw Mya is cooking.)	th mĭˆ tɕhɛ˒ dəɔ˗ mj th mĭˆ tɕhɛ˒ ne˗ θi˗	th mĭɲˆ **dzɛ˒(rule 1)** dəɔ˗ mj th mĭˆ **tɕhɛ˒ (rule 3)** ne˗θi˗
ငါးခူဝယ်ခဲ့ပါ။ (Buy catfish)	ŋaˆ khu˗ wɛ˗ khɛ´ pa˗	**ŋə (rule 2)** khu˗ wɛ˗ **gɛ´ ba˗(rule 1`)**
အဖွားကျန်းမာရေး**ထူထူ**ထောင်ထောင်ရှိသည်။ (Grandmother's health become well.)	a phwaˆ tɕãˆ ma˗ jeˆ thu˗ thu˗ thaũ˗ thaũ˗ jhi´ θi˗	a phwaˆ tɕʌˆ ma˗ jeˆ **thu˗ du˗ thaõ˗ daõ˗ (rule 4)** jhi´ θi˗

Rule1 (Change voiceless to voice sound)

As rule1, for a given word, if the first syllable is ending with glottalized sound, ending with /ʔ/, the phoneme of next syllable does not change to voice sound. Moreover, we also consider that if it does not ending with stop vowel, the next syllable is changed voiceless to voice sound for all cases. However, if the next consonant is aspirated voiceless consonant

(/th,kh,ph,tɕh,sh,ɵ/), it does not change to voice one. For example – သဲဆပ်ပြာ(ɵ ɛˆ +sʰ aˀ+ p jaˉ) (sand soap)→သဲဆပ်ပြာ(ɵ ɛˆ +sʰ aˀ+ p jaˉ). In this example, သဲ- /ɵ ɛˆ/ is not ending stop vowel, normally, the ဆပ်- /sʰ aˀ/ should be changed into voice one. However the first syllale (sh) is aspirated consonant so it does not change. For the syllable ပြာ- /p jaˉ/ is also does not change to voice one because the previous vowel is ending with /aˀ/.

Rule2 (Derivative compound)

When the first syllable in a word does not have stop vowel sound and the pronunciation of two consonant and vowel compound nouns may not keep the original meaning, it changes into derivative compound noun and the vowel of first syllable moves to central vowel (ə). At that time, if this first syllable or next syllable is voiceless obstruent, they change to voiced one. Especially, it occurs in the consonants (/kh, ŋ, p, ph, d, w/) plus the vowel "aˆ" and (/sh, t/) plus vowel "ãˉ". As an exception, the semi vowel /w/ is destroyed for the syllable"/nwaˆ/ (cow) and /ɵwaˆ/ (tooth)". For example – ဓားမ-/d aˆ m a′/(knife) →ဓမ-/d ə m a′/. For this word, the syllable /d aˆ/ cannot keep its original sound and move to the central vowel into /d ə/.

Rule3 (Voicing & unvoicing for verb)

This rule is mainly focus on verb. Therefore, noun and verb phrases are identified by using rules and built lexicon. In Myanmar language, there are two types of compound verbs: noun (N) + verb (V) compounds and verb + verbs compounds. In this rule, firstly, the input word is analytic into two forms of compound verbs. If the input word is (N+V) compound, V is not changed into voice obstruent. If the word is (V+V) compound, the last V phoneme is changed when the syllable is unvoiced one. However, if the input word is N, individual V or other part of speech, it is considered with the above proposed rules. For example – လမ်းပြ(l ãˆ + p ja′)→လမ်းဗျ(noun) (l ãˆ + b ja′). လမ်းပြ(l ãˆ + p ja′)→လမ်းပြ(verb) (l ãˆ + p ja′). For the first one, the second syllable ပြ-/ p ja′/ changed into /b ja′ / because of its part of speech is noun. However, the second one does not change because of its POS.

Rule 4 (Voicing in reduplication)

When the syllable in a given word that may be verb or noun is duplicated, the second syllable changes to voiced sound regardless of any part of speech tag such as for the word "ထူထူထောင်ထောင်-/ thuˉ thuˉ thaũˉ thaũˉ/ (strong and upright), the syllable "ထူ-/thuˉ/ and ထောင်-/thaũ/" are duplicated. Therefore, the second syllable from this duplicated word is changed into voiced sound as "ဒူ-/duˉ/ and ဒေါင်-/daõˉ/" such that -/thuˉ duˉ thaũˉ daõˉ". However, if the duplicated verb is combined with adverbial suffix "ချင်း (/tɕhĩˆ/)" (ထောင်ထောင်ချင်း-/thaũˉ thaũˉ tɕhĩˆ), the second duplicated consonant does not change to voiced sound (tɕhĩˆ to dʑĩˆ)".

4.2.2The evaluation for G2P conversion

To evaluate the quality of G2P conversion, phoneme error rate (PER) is calculated with this equation: ***PER = (Insertion + Deletion + Substitution) * 100/N*** (number of words). Firstly, the G2P hypothesis (the output from the system) and reference (Myanmar Lexicon (version-2)- it is a Myanmar lexicon which contains pronunciation and meaning for each word) word strings are aligned. Then, perform a global minimization of the Levenshtein distance function which weights the cost of correct words, insertions (I), deletions (D) and substitutions (S).The PER for G2P conversion is shown in the following *Table 4*.

Table 4 Evaluation result of G2P conversion

	Number of Word	**Per (%)**	**Accuracy (%)**
G2P Conversion by using phonological rules	**22544**	**7**	**93.54**

4.2.3Finding and discussion for G2P conversion

The PER for G2P conversion is only 7. Although the acceptable level of PER is achieved, the exception is still exists. In rule 1, some aspired syllable in words cannot be converted into correct g2p form; such as "ရွှေဆိုင်(gold shop) စိန်ထည် (diamond) and မီးခိုး(smoke)". In rule 2, some word does not change into derivative compound noun and the vowel of first syllable moves to central vowel. For example "ဓားသွေးကျောက်, ငါးဆယ်သား, ငါးပါးသီလ, ဖားစည်, ဆံတော်, တံလှပ်". As an exception, some words that are not covered by rules are predefined. Therefore, as conclusion, our proposed rules also need around 17% exception to cover the whole language.

4.3Prosodic analysis

When we synthesized the speech, prosodic analysis is the most important module to get natural sounding TTS system. It includes the main features like duration, pitch, accent and prosodic phrasing. One major feature of the prosody of Myanmar speech flow is prosodic phrase grouping that is an utterance

has a prosodic phrase structure. Some words in spoken sentence look like to group naturally together and some words seem to have a noticeable break or disjuncture between them [15]. In Myanmar writing style, typically, there is no space between words, but insert space, sometime, between phrases based on user minded feeling. There are two main break characters in Myanmar: one or two downward strokes (၊ or ။), which respectively act as a comma and a full stop. The character "၏" used as full stop,"၍" used as a sentence connector,"၌" as locative marker,"၎း" used in columns and lists, are also used as punctuation mark in Myanmar. Wherever a space or some punctuation marks are present in the text, during parsing, the synthesizer will introduce a silence to represent it. This will improve the quality and even in some cases, the proper meaning can be conveyed. However, for Myanmar, in the formal written text, the phrase boundaries are not explicitly present thus the quality of the Myanmar TTS is found to be poor, in the sense that, the individual words in the sentence sound very good, but as a sentence, it does not sound natural. To perceive naturalness in the synthesized speech we need little longer silence in between the phrases than usual.

Therefore, in our proposed TTS system, the phrase boundary detection is considered based on the phrase particles that act as suffixes and they may be attached to a word that can become phrases in a sentence without having any effect on its role in the sentence. There are 455 words defined as particles according to Myanmar lexicon (Version-2)-Lexique Pro so that we create the particle corpus. To define phrase boundary, firstly, the normalized sentences are segmented into words by using longest matching approach [16] and these words are assigned their part of speech (POS) tag based on lexicon that created with 22544 words. We separate the input sentence into the phrase group based on these particles by using general rules and then assign pause duration behind the every phrase. Therefore, the synthesizer will give silence wherever there is pause. It gives longer pause for period (။) and medium pause for comma (၊) and very short pause for spaces. This will improve the quality of our Myanmar TTS system. *Table 5* shows the example of phrase boundary detection and separation and the algorithm for assigning duration between phrases is described in algorithm 2.

Table 5 Example for phrase boundary detection

Sentence level	သူသည်ငါးခူကိုချက်သည်။(He cooks fish.)						
Syllable level	သူ	သည	ငါး	ခူ	ကို	ချက်	သည်
Word level	သူ/he (pronoun)	သည် (particle)	ငါးခူ/catfish (noun)		ကို (particle)	ချက်/ cook (verb)	သည်။ (paricl-e)
Phrase level	သူသည် (Noun phrase)		ငါးခူကို (Noun phrase)			ချက်သည်။ (Verb phrase)	

Algorithm 2: Phrase segmentation and assign pause duration

Input: Syllable Sequence
Output: Sentence with segmented phrase and assigned pause duration
1: segment word Wi based on syllables Si, W= {W1, W2, …}
2: for each word Wi ∈ W do
3: If next word is particle then
4: Segment into phrase and assign pause 150 ms duration
5: If next word is ((၊) or space) then
6: Assign pause 100 ms duration
7: end for

4.4Speech synthesis

This module is the final one of the MTTS system and Myanmar speech is generated in this step. In the case of speech generation, we used concatenative speech synthesis technique. It has gained in popularity in recent years, due to its more natural sounding synthesized speech and it is based on the concatenation of speech signal that are pre-recorded in the database. Concatenating the natural recorded wave files is the better way to produce understandable and natural sounding synthesis speech.

However, it requires more memory capacity. Therefore, the most important aspect in concatenative speech synthesis is to find correct speech unit length. The selection is trade-off between longer and shorter units. With longer unit high naturalness, less concatenation points are achieved, but the amount of required units and memory is increased. The shorter units less memory are needed, but the sample collection and labelling procedures become difficult and complex. The present systems units are: Words, Syllables, Demi- syllables, Di-phones, Tri-phones and Phonemes. For our proposed system, phoneme is used for concatenation.

4.4.1Phoneme speech database creation

The Phonemes are abstract units and their pronunciation depends on contextual effects, speaker's characteristics, and emotions. A set of phonemes can be defined as the minimum number of symbols needed to describe every possible word in a language. Phonemes are probably the most commonly used units in speech synthesis because they are the normal linguistic presentation of speech [17]. It is smallest compared to other units so that it is most suitable for embedded devices.

Myanmar syllable is composed of consonant and vowel phonemes. There are 34 consonants in Myanmar language but only 21 consonant (C) phoneme sound units need to be recorded because some has same pronunciation.

Moreover there are 4 basic medials (M) and 6 combined medials. Consonant letters may be modified by one or more medial diacritics (three at most), indicating an additional consonant before the vowel. Therefore, the 10 medials can modify the 34 basic consonants. For example - ပ (C) + ျ- (M) = ပျ (show), လ (C) + ွ(M) = လွ (beautiful). However, the medials cannot combine all of the consonants. For example: there is no combination for က(C) + ျ (M) + ွ (M) + ှ (M) = ကျွှ that cannot be pronounced. Therefore, after final counting the syllable that can be pronounced by combining with medials are only 39 syllables in Myanmar scripts that are what we need to be recorded in this system.

In Myanmar language, only consonant cannot stand to produce speech sound that can only be made by combination of vowel phoneme. There are 8 basic vowels defined in Myanmar language defined by [13]. Myanmar language is a tonal language and there are four tone levels. According to these tone levels, these vowels can be expanded into 50 vowels sounds that can be divided into 21 nasalized vowels, 22 non-nasalized vowels and 8 glottal stop vowels. These vowels are what we need to be recorded. These vowels are play an important role to construct syllable and to product Myanmar speech sound. These 50 vowels can made any syllables and any speech sound by multiplying consonant and medials. For example-"အား-aˆ" vowel can make the "ကား-kaˆ (car), စား-saˆ (eat), ထား-thaˆ (place)" syllable by combining "က-k (dance), စ-s(start), ထ-th (wake up) consonants + "အား-aˆ" vowel. Finally, the number of phonemes to be recorded for our proposed system is only 133 as shown in *Table 6*. These phonemes are recorded with 44100HZ sampling rate by native female speaker in LA Studio, Mandalay and it took one and half hours. Any raw text can be speech out based on these 133 phonemes. The phoneme speech database size is obviously reduced compared to the former diphone based Myanmar TTS proposed by Soe and Thida [7]. This is suitable for low computation performance devices.

Table 6 Phoneme speech database

Phoneme type	Number of phoneme	Cannot pronounce	Can pronounce
Consonants (C)	21	0	21
Vowels(V)	50	0	50
21(C)* 6 Medial	126	79	47
Special Character	5	0	5
Number	10	0	10
Total phonemes			**133**

4.4.2Phoneme concatenation

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 proposed system, we used concatenative method and phoneme is selected as basic unit for concatenation. Typically, Myanmar language is syllabic language and thus its spoken sentence form is based on the syllable that is made of consonant phoneme (CP) and vowel phonemes (VP). For example we can make sound for the syllable "ကျောင်း(school)-tɕaũˆ " into "KAUNG:" with two phoneme sounds "ကျ-/ tɕ/" (KYA) and "အောင်း-/aũˆ/" (AUNG:). Actually, 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. As an example, firstly, we separate the syllable phoneme "tɕaũˆ" into consonant and vowel phoneme "/tɕ/ and /aũˆ/" respectively. Then, fetch the corresponding speech .wav files from

the created database. Then, based on their duration, we modified the consonant phoneme and vowel phoneme according to threshold values. Then, these two wave files are normally concatenated so that we can get our desired speech output for syllable that can make word, phrase and up to sentences (see *Figure 2* to *Figure 6*). The phoneme concatenation method achieves the acceptable level of intelligibility and naturalness TTS.

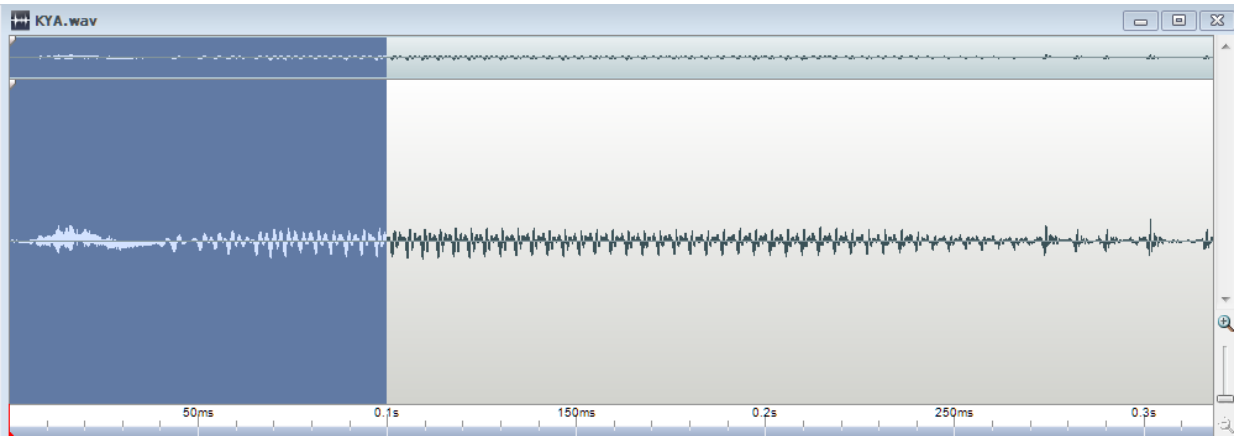

Figure 2 Original "ကျ-/tɕ/" (KYA) wave file

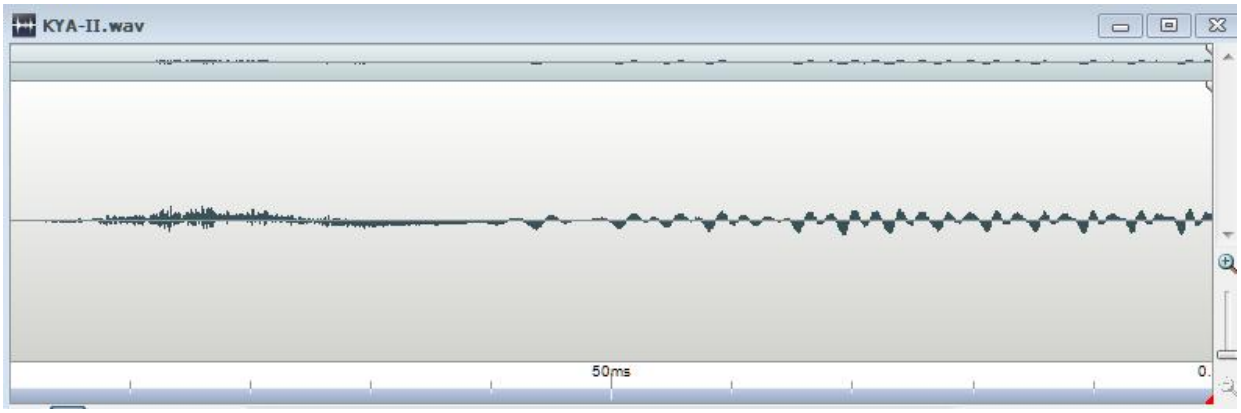

Figure 3 Modified "ကျ-/tɕ/" (KYA) wave file

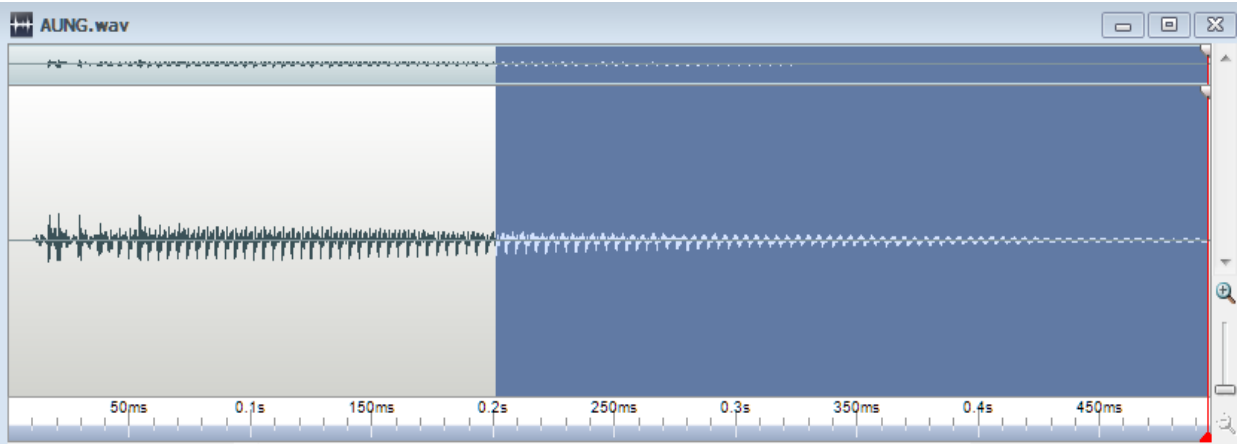

Figure 4 Original "အောင်း-/aũˆ/" wave file

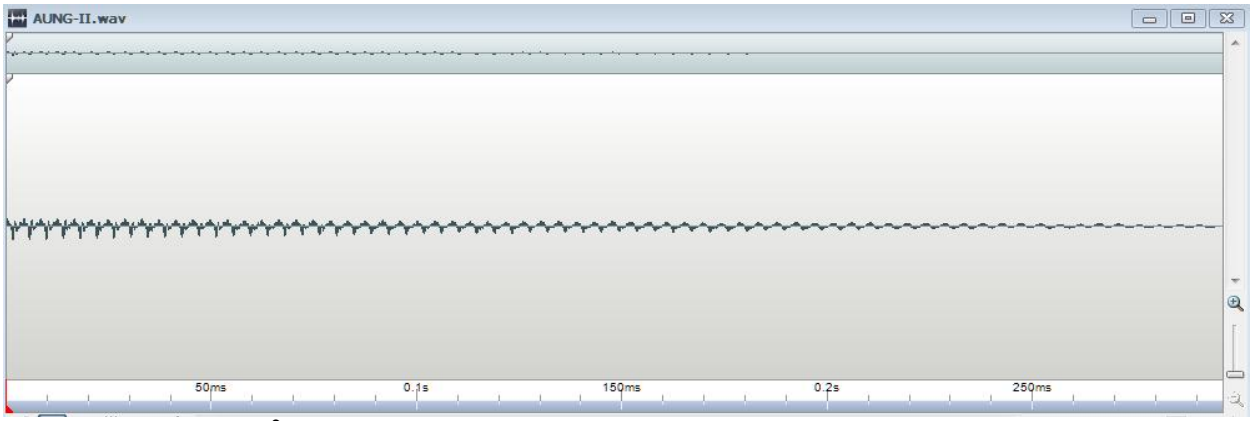

Figure 5 Modified "ဇော်ငိုး-/aũˆ/" wave file

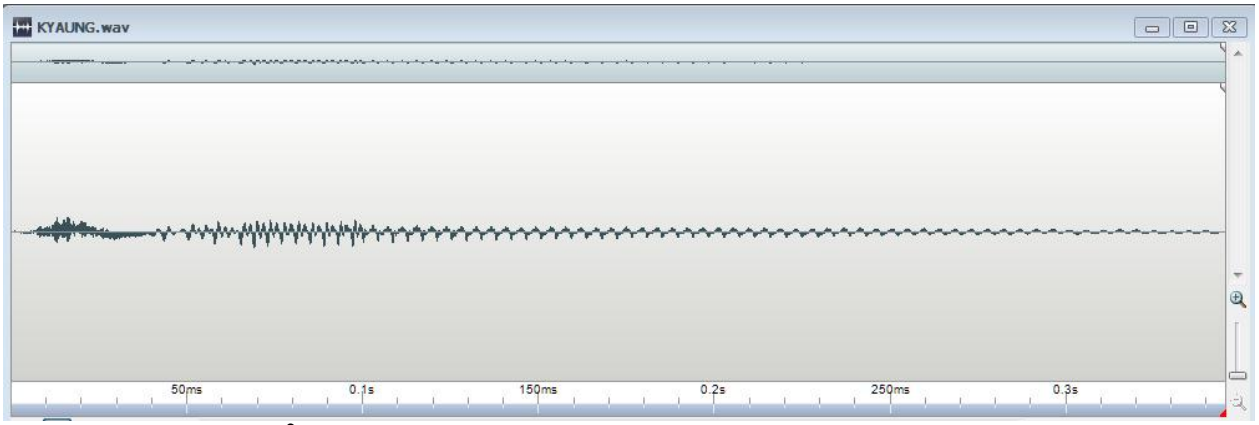

Figure 6 The result "ကျောင်း(school)-tɛaũˆ " wave file

4.4.3Evaluation for speech synthesis

The evaluation test for Myanmar TTS was designed according to existing methods that are to determine how much of the spoken output one can understand intelligibility and how much the speech output is near the real human speech naturalness. The proposed MTTS system's speech quality is evaluated by the six evaluators as shown in *Table 7*.

Table 7 Evaluator information for MTTS quality testing

No	No. of sentences	Gender/No	Age
1	400	Female (3)	19-25
2	400	Female(2) Male(1)	30-40

4.4.4Intelligibility test

For intelligibility test, it is one of the important factors affecting speech quality; we can calculate it either by mean opinion score (MOS) and word error rate (WER). In this test, we choose 400 sentences randomly with the average word length are 10. After the evaluators listen a sentence, they have to write whatever they hear, even if they don't understand the meaning. When we calculated WER for each person, the WER for intelligibility test is only 7% and the average MOS score for intelligibility is 3.7 and the result is shown in *Figure 7*.

$$\text{WER} = \frac{\text{substitutions} + \text{insertions} + \text{deletions}}{\text{reference} - \text{length}}$$

4.4.5Naturalness test

The most useful and simplest method to test naturalness is MOS. It has five level scales in MOS: bad (1), poor (2), fair (3), good (4) and excellent (5). Formula for calculation of MOS is as follow 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.

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

For the naturalness test, the evaluators are asked the system generated Myanmar speech is naturalness or not. According to their answers, 2% of listeners

though about the output speech is very natural, 21% considered the speech are natural and 51% of listener identified the voice are acceptable. Around 23% assumed the speech output is needed to get more naturalness. The average MOS score for 400 sentences is 2.96. Therefore, speech output of the system achieved the acceptable level of naturalness. The naturalness results from MOS evaluation of the system are shown in *Figure 8*.

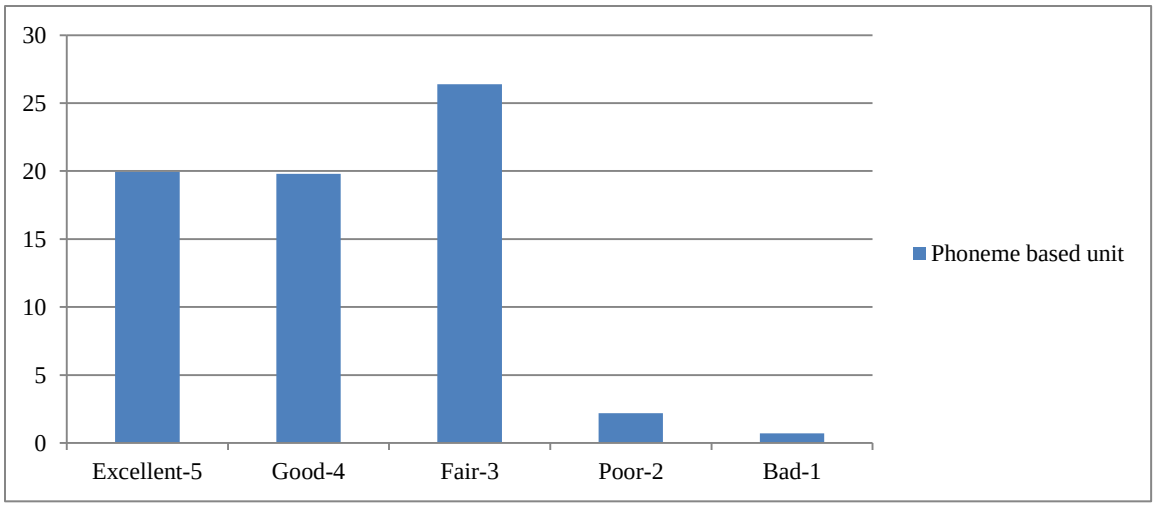

Figure 7 MOS for intelligibility test

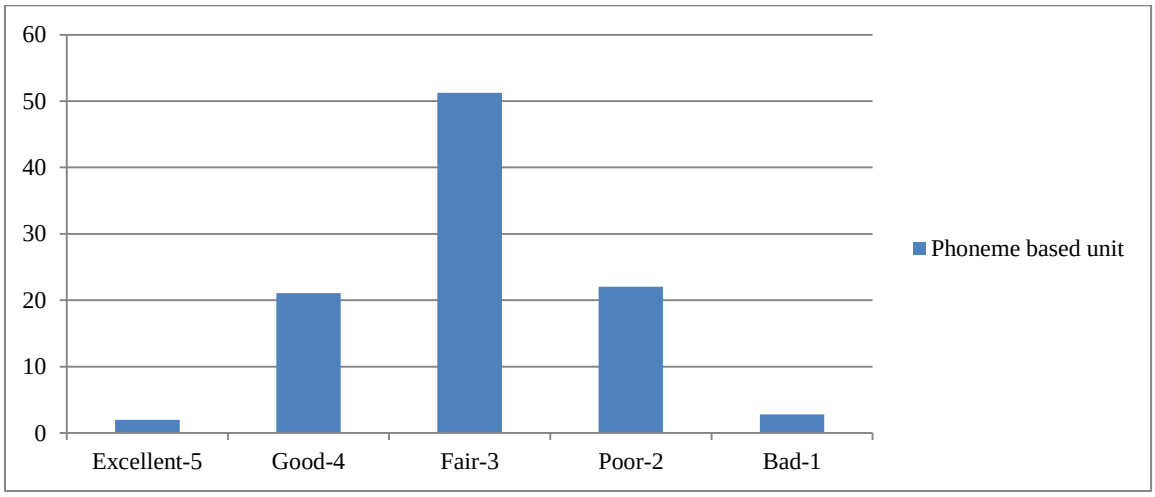

Figure 8 MOS for naturalness test

5.Conclusion

We presented the detailed modules for developing MTTS system. For the text analysis modules, we proposed rule based number converters that can convert nearly every Myanmar number based on their context. Then, we described the proposed phonological rules that can support for getting high quality TTS system. MTTS system is developed by using a concatenative speech synthesis method using phoneme as a speech unit for concatenation. The phoneme speech database is created and it contains only 133 phonemes that can speech out for all Myanmar texts. Therefore, this system is suitable for resource limited devices. Each module is evaluated by experimental results. According to our experimental results, MOS score (intelligibility & naturalness) for our proposed algorithm is 3.7 and 2.96. Therefore, it achieved an acceptable level of intelligibility, but still need naturalness. For future work, the system can be accommodated by using digital signal processing approaches such as TD-PSOLA and Wave Similarity Overlap and Add (WSOLA) for smoothing concatenation points to get more high quality Myanmar TTS system.

Acknowledgment
None.

Conflicts of interest
The authors have no conflicts of interest to declare.

References
[1] Black AW, Campbell N. Optimising selection of units from speech databases for concatenative synthesis.1995.
[2] Conkie A. Robust unit selection system for speech synthesis. The Journal of the Acoustical Society of America. 1999 .
[3] Hunt AJ, Black AW. Unit selection in a concatenative speech synthesis system using a large speech database. In international conference on acoustics, speech, and signal processing 1996. (pp. 373-76). IEEE.
[4] Toda T, Kawai H, Tsuzaki M, Shikano K. Unit selection algorithm for Japanese speech synthesis based on both phoneme unit and diphone unit. In international conference on acoustics, speech, and signal processing 2002 (pp. 465-8). IEEE.
[5] Douke M, Hayashi M, Makino E. A study of automatic program production using TVML. Eurographics.1999:42-5.
[6] Win KY, Takara T. Myanmar text-to-speech system with rule-based tone synthesis. Acoustical Science and Technology. 2011; 32(5):174-81.
[7] Soe EP, Thida A. Text-to-speech synthesis for Myanmar language. International Journal of Scientific & Engineering Research. 2013; 4(6):1509-18.
[8] Wongpatikaseree K, Ratikan A, Thangthai A, Chotimongkol A, Nattee C. A real-time Thai speech synthesizer on a mobile device. In international symposium on natural language processing 2009 (pp. 42-7). IEEE.
[9] Wongpatikaseree K, Ratikan A, Chotimongkol A, Chootrakool P, Nattee C, Theeramunkong T, et al. A hybrid diphone speech unit and a speech corpus construction technique for a Thai text-to-speech system on mobile devices. In international conference on electrical engineering/electronics computer telecommunications and information technology 2010 (pp. 1089-93). IEEE.
[10] Mokgonyane TB, Sefara TJ, Manamela PJ, Manamela MJ, Modipa TI. Development of a speech-enabled basic arithmetic m-learning application for foundation phase learners. In AFRICON 2017 (pp. 794-9). IEEE.
[11] Karabetsos S, Tsiakoulis P, Chalamandaris A, Raptis S. Embedded unit selection text-to-speech synthesis for mobile devices. IEEE Transactions on Consumer Electronics. 2009; 55(2):613-21.
[12] Gopi A, Shobana PD, Sajini T, Bhadran VK. Implementation of Malayalam text to speech using concatenative based TTS for android platform. In international conference on control communication and computing 2013 (pp. 184-9). IEEE.
[13] Myanmar language commission, Myanmar grammar. 30th year pecial edition. University Press, Yangon, Myanmar; 2005.
[14] Tun DT. Acoustic phonetics and the phonology of the myanmar language. School of Human Communication Sciences, La Trobe University, Melbourne, Australia. 2007.
[15] Zhao Z, Ma X. Active learning for the prediction of prosodic phrase boundaries in Chinese speech synthesis systems using conditional random fields. In IEEE/ACIS international conference on software engineering, artificial intelligence, networking and parallel/distributed computing 2015 (pp. 1-5). IEEE.
[16] Htay HH, Murthy KN. Myanmar word segmentation using syllable level longest matching. In the workshop on Asian language resources, IJCNLP 2008 (pp. 41-8).
[17] Jurafsky D, Martin JH. Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition. Prentice Hall; 2000.

Chaw Su Hlaing is a Ph.D. student at Artificial Intelligence Lab, Faculty of Computer Science, University of Computer Studeies, Mandalay, and Myanmar. She had received her B.C.Sc. and M.C.Sc. from Faculty of Computer Science, University of Computer Studeies, Mandalay, and Myanmar. Her current research interests are Web Data Mining, Digital Signal Processing, Natural Languge Processing and Liguistic Research. She is currently working in the research of Speech Synthesis for Myanmar Language.
Email: chawsuhlaing.ucsm.mm@gmail.com

Aye Thida is a Professor at Artificial Intelligence Lab, Faculty of Computer Science, University of Computer Studeies, Mandalay, and Myanmar. She was one of the leaders of Natural Language Processing Projects. Her team has developed Myanmar to English Translation System in 2011. Her research interests include High Performance Computing, Distributed Processing, Queuing and Natural Language Processing. She had received B.Sc(Hons)Maths 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.
Email: ayethida.royal@gmail.com