Research Article

International Journal of Advanced Computer Research, Vol 9(41)
ISSN (Print): 2249-7277 ISSN (Online): 2277-7970
http://dx.doi.org/10.19101/IJACR.2018.839001

Phoneme concatenation method considering half vowel sound for the Myanmar speech synthesis system

Chaw Su Hlaing[1*] and Aye Thida[2]
Research Scholar, Artificial Intelligence Lab, Faculty of Computer Science, University of Computer Studies, Mandalay and Myanmar[1]
Professor, Faculty of Computer Science, University of Computer Studies, Mandalay and Myanmar[2]

Received: 30-August-2018; Revised: 10-February-2019; Accepted: 12- February-2019

Abstract
Myanmar language is a tonal language and it has different written form and spoken form. Therefore, correct grapheme to phoneme conversion is one of the important steps in the developing of Myanmar text-to-speech system. Every Myanmar consonant has inherent vowel or half vowel, schwa vowel depends on the word. Therefore, the correct vowel insertion is also a critical task. If these vowels can be handled, the TTS quality will be higher so that schwa vowel handling rules are presented in this paper. Besides, this paper discusses the approach considered for the vowels 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. Since phoneme plays an important role, Myanmar phoneme inventory is presented in detail. After analysing the number of phonemes and half-sound consonants to be recorded, the Myanmar phoneme speech database which contains total 157 phoneme speech sounds have been created. It 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 highest level of intelligibility and acceptable level of naturalness.

Keywords
Text to speech, Myanmar language, Phoneme, Concatenative speech synthesis, Half-vowel sound.

1.Introduction
Text to speech (TTS) conversion is a system that can convert the written text into their corresponding speech. It is a very useful application for the visual and speech impaired person. The optimal character recognition (OCR) -based TTS system to help such visually challenged people by OCR has been proposed [1]. The resulting text from the OCR is converted into speech. They used the blind deconvolution method and 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. Nowadays, high quality TTS software can be commercially available for different languages. The most used speech synthesis approaches are articulatory synthesis, formant synthesis, concatenative synthesis and hidden Markov model (HMM)-based model approach. Each approach has their reasonable advantages and disadvantages based on the usage of languages. Among them, the concatenative synthesis approach is used in our system because it can generate natural sound as a consequence of pre-recorded sound. The speech quality and the size of the system is a 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 TTS systems proposed by [2−6] have been implemented by using 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 speech signal. They compare two Arabic text to speech systems: two screen readers, namely, non-visual desktop access (NVDA) and integrated bilingual solution for the blind or visually impaired, in the Arab (IBSAR) [8]. They tested the quality of two systems in terms of

*Author for correspondence

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 a competitive performance on the intelligibility tests. TTS system by using the concatenation method by using phoneme and defined as a speech unit for concatenation has been presented [9]. They also compared based on these two speech units by using the two different methods of join cost calculation. According to their result, diphones speech units achieved a more acceptable level from the subjective evaluations. In the research [10], they also used concatenative speech synthesis and part syllable for concatenative speech units was used to reduce the number of training sentences and the concatenation error. Therefore, part syllable transformation-based voice conversion (PST-VC) method was introduced which performs voice conversion with very limited data from a target speaker and simultaneously reduces concatenation error.

There are approximately a hundred languages spoken in Myanmar. Among them, the 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. Thus, syllable, word and phrase detections are most challenging tasks in Myanmar natural language processing research works. Typically, Myanmar speech and the 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, the 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 [11].

For Myanmar language, there has been considerable effort on speech processing in Myanmar natural language processing research works. Typically, text to speech systems in different languages have been developed by using different approaches as well as for Myanmar language. Rule-based Myanmar text-to-speech system (MTTS) system has been designed in which fundamental speech units are demi-syllables with level tone [12]. They used a source filter model and furthermore a log magnitude approximations 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 [13]. They used 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 unit is 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. Therefore, [14] proposed 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 speech out for all Myanmar texts. 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 consonants. Generally, Myanmar language is complicated not only in the number of basic sounds, but also in phonetic in language. It has different in pronunciation written form and spoken form. Moreover, every consonant has inherent vowel or half vowel, so that it is important to handle correct vowel defining in the process of phoneme conversion. Therefore, in this paper, we presented schwa vowel insertion rules. Myanmar language is a tonal language so that it can generate more variety of sounds so that the size of speech database become larger than other languages. In this situation, there may be some problem when transferring TTS to mobile devices because of limited resources in both storage capacity and processing. Therefore, in this paper, we proposed new phoneme concatenation method that extends the previous one. In this method, only 157 speech units are used to speech out for all Myanmar texts. So, it is very comfortable for mobile devices. In our method, the half-sound consonants are also considered so that we have to prepare the text for recording. After that, segment and label the recorded sound to get half-sound of consonant. Then, fetch the appropriate speech file 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 extending phoneme concatenate method. Then, they were tested on the already developed phoneme based MTTS system [14].

The rest of this paper is organized as follows: in Section 2, the brief explanation of a Myanmar TTS system is presented and the schwa insertion rules are also described. Moreover, the detail explanation of Myanmar phoneme inventory is presented and the main contribution of this paper, phoneme concatenation method is also described. The experimental results of the proposed method are discussed in section 3 and the error analysis also discusses in section 4. Finally, the paper is concluded in section 5.

2.Materials and methods

In this section, the brief explanation of MTTS system is presented and used one of the most popular speech synthesis approaches, that is a concatenation speech synthesis. Concatenative method is a trade-off between the selection of speech units. Based on the selected speech units, the speech quality may be different. In this MTTS system, phoneme speech units are used for concatenation intended for the resource limited devices. Therefore, creating a phoneme speech database is very important so far detail explanations are described in this section. And then, the proposed extended phoneme concatenation method based on the created speech database is also explained specifically in the following sub-sections.

2.1Myanmar text-to-speech system

There are four main steps in the development of phoneme-based MTTS system. These are text analysis, phonetic analysis, prosodic analysis and speech synthesis respectively. In the system, firstly, Myanmar sentences are tokenized by using rule-based tokenizer [15] for the next processing. Typically, the input text can contain non-standard words (NSW) such as numbers, abbreviations and other than simple text. 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 step that is also called grapheme-to-phoneme conversion (G2P). The quality of TTS system also depends on the right G2P conversion. In the Myanmar language, actually, the words in writing and speaking are sometimes different. This problem is solved by using phonological rules to get the correct G2P conversion. Generally, fifty percent of Myanmar words have schwa vowel sound. Schwa is the only vowel that is permitted 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. Therefore, the schwa insertion rules are proposed and described in the following section. After that, in the prosodic module, assigning pauses are considered to get the more natural speech output. Finally, in the speech synthesis module, the corresponding speech files of phoneme sequences from the previous step are fetched from the created phoneme speech database for concatenation. Consequently, in this paper, the detail explanation of how to create phoneme speech database and proposed phoneme concatenation method is presented in the next sections.

2.2Inserting schwa vowel

In the Myanmar language, speech sound cannot be exactly expressed by written text. For example, the sound of "က" in the word "ကလောင်တံ (pen)" and "ကချေသည် (dancer)" is not same. The word "ကလောင်တံ" is pronounced as [**kə** laʊ̃ ˉ tʌ̃ ˉ] but pronounced as [**ka˝** tɕʰe ˉ ɵɛ ˉ] in the word "ကချေသည်". They have different vowel in the first syllable. Unlike English and other Asian languages, Myanmar consonant cannot be pronounced itself except nasal sound. Besides, they have to write combined with vowel to make a syllable. Actually, "က" is not /k/ but it may be /kə/ or /k**a˝**/ because every consonant has inherent vowel or schwa vowel.

In Myanmar language, when basic consonants (က-/k/, ခ-/k^h/, မ-/m/) is combined with other syllables (လေး, ရေ, နက်ဖြန်) to become words (ကလေး (baby), ခရေ (star flower), မနက်ဖြန် (tomorrow)), the speech sound of the first consonant is not fully pronounced. It turns into half-vowel or schwa as [**kə** leˆ], [**k^hə** ɹe ˉ], [**mə** nɛˀ p^hjʌ̃ˉ] instead of [ka˝ leˆ], [k^ha˝ ɹe ˉ], [ma˝ nɛˀ p^hjʌ̃ ˉ].

Therefore, if the basic consonant is situated in the first position of the word, it is pronounced as a schwa. Schwa is a very short neutral vowel sound, and like all other vowels, its precise quality varies depending on the adjacent consonants. In theory, virtually any written syllable that is not the final syllable of a word can be pronounced with the vowel [ə] (with no tone and no syllable-final [-ʔ] or [-N]) as its rhyme. In practice, the simple consonant letter alone is the most common way of spelling syllables whose rhyme is [ə].

Generally, a Myanmar phonological structure is complex and there are some rules for inserting schwa. If the schwa can be correctly considered, then the quality of MTTS system will be high in advance. Therefore, schwa inserting rules are proposed in the following sub-sections.

2.2.1Rule 1 (Internal closed juncture and unoriginal compound)

If the first consonant vowel (cv) is not the glottalized tone, a formation of the compound of the combination of two "cv" syllables, or the combination two of "v+cv" or "cv+v" syllables cannot keep their own original sound completely and form an internal closed juncture. The first property of an internal closed juncture is the tone of the first vowel slide to no tonal stage, which it represents with a / ə / symbol. For instance – the word "စာရေး - /sa- ɹeˆ /", the first syllable "/sa¯/" has no glottalized tone and is the cv combination. However, it cannot keep their original sound and it changes to tonal state as schwa vowel /sə/. *Table 1* shows phoneme sequence results of sample words that applied the Rule 1.

Table 1 Example words applied Rule 1

Before	After
/ saˉ ɹeˆ/ စာရေး(clark)	[sə ɹeˆ] စရေး
/ θuˉ ŋɛˉ/ သူငယ်(boy)	[θə ŋɛˉ] သငယ်
/ waˆ tɕʰãˆ / ဝါးချမ်း(half of bamboo)	[wə tɕʰãˆ] ဝခြမ်း
/ ŋaˆ kʰuˉ / ငါးခူ(catfish)	[ŋə kʰuˉ] ငခူ
/ daˆ maˊ / ဓါးမ(Knife with broad blade)	[də maˊ] ဓမ
/ pʰaˆ laũˆ / ဖားလောင်း(tadpole)	[pʰə laõˆ] ဖလောင်း
/ θaˆ aʰeĭˉ / သားအိမ်(ovary)	[θə aʰeĭˉ] သအိမ်

2.2.2Rule 2 (Initial obstruent voicing in internal closed juncture)

Rule 2 is the second property of an internal closed juncture. The second syllable consonant changes to voiced consonant if it is un-aspirated voiceless obstruent when two "cv+cv" syllables combination. For instance – in the word "ပန်းကန် - /pãˆ kãˉ/", the first syllable "/pãˆ/" is changed into schwa sound as /pə/ according to Rule 1 and the second one "kãˉ" is un-aspirated voiceless consonant "က, ဂ, တ, ဒ, ပ/ (/k/, /g/, /t/, /d/, /b/)". Therefore, it changes to voiced one "gãˉ". Some other example phoneme results from Rule 2 are shown in *Table 2*.

Table 2 Example words for Rule 2

Before	After
/ pãˆ kãˉ / ပန်းကန် (plate)	[pə gɅˉ] ဗဂန်
/ pãˆ teĭˉ / ပန်းတိမ် (silversmith)	[bə deĭˉ] ဘဒိမ်
/ pʰaˆ pjãˉ / ဖားပျံ (tree frog)	[pʰə bjɅˉ] ဖဗျံ

2.2.3Rule 3 (Voicing reflection in internal closed juncture)

The third property of an internal closed juncture is when a vowel of "cv" syllable reaches non-tonal stage, schwa (/ə/), the first consonant is changed to the voiced consonant if it is voiceless obstruent. For instance – in the word "ကြမ်းပိုး - /tɕãˆ poˆ/", the first syllable "/tɕãˆ/" is turn to schwa /tɕə/ and the second syllable "/poˆ/" is turn to voice one /boˆ/ according to Rule 1 and 2. Then, the first voiceless syllable (/tɕə/) has to change voice one (/ʤə/) also because it's vowel is schwa. The following *Table 3* shows some words that applied the Rule 3.

Table 3 Example word applied Rule 3

Before	After
/ tɕãˆ poˆ / ကြမ်းပိုး (bug)	[ʤə boˆ] ဂျိုး
/ sʰãˉ kɔˆ / ဆန်ကော (round bamboo tray)	[zə gɔˆ] ဇဂေါ
/ tãˉ taˆ / တံတား(bridge)	[də daˆ] ဒဒါး

2.2.4Rule 4 (Semivowel deletion in internal closed juncture)

The fourth property of an internal closed juncture is that if the beginning of the syllable contains "နွား" / nwaˆ / and "သွား" / θwaˆ /, the /w/ sound is disappeared. For instance – in the word "နွားထီး - /nwaˆ thiˆ/", the first syllable is "nwaˆ ". Therefore, the medial Wa hswe "ွ - /w/" is destroyed and its vowel becomes schwa, "nə". Some sample words from Rule 4 is shown in *Table 4*.

Table 4 Example words applied Rule 4

Before	After
/ nwaˆ thiˆ / နွားထီး (ox)	[nə thiˆ] နထီး
/ θwaˆ tɛʔ / သွားတက်(redundant tooth)	[θə dɛʔ] သဒက်
/ θwaˆ pʰoũˆ / သွားဖုံး(gun)	[θə pʰoõˆ] သဖုံး

2.2.5Rule 5 (The influence of the first syllable in internal closed juncture)

A vowel of the number counting "တစ် - /tiʔ/ is disappeared in the fifth property of an internal closed juncture even its vowel end with glottalized tone (Tone IV). Besides, in the sixth property of an internal closed juncture, the vowel of the counting number description word "နှစ်" - /hniˀ/ is also disappeared too. However, the next consonant syllable does not change to the voiced obstruent even if it is the voiceless obstruent due to the influential of "နှစ် - / hniˀ /.

Table 5 Example words applied Rule 5

Before	After
/ tiʔ soũ¯ /	[də zoũ¯]
တစ်စုံ(one set)	ဒဇုံ
/ hniˀ tʰaˀ keĩˆ /	[hnə tʰʌˀ keĩˆ]
နှစ်ထပ်ကိန်း(square)	နှထပ်ကိန်း

2.2.6Predefined rules

Some of phoneme in Myanmar language has already been changed in allophone in any environment.

Table 6 Example words applied predefined rules

Before	After
/ tɕʰe¯ tʰauˀ /	[tɕʰi¯ daʊˀ]
ခြေထောက် (leg)	ချီဒေါက်
/ ljɛˀ sʰaˆ/	[jɛˀ sʰaˆ]
လျက်ဆား	ယက်ဆား
/ taˊ piˊ /	[də bɛˊ]
တပည့်	တဗဲ့

In the study of the phonology, the writing form is not the essence to be considered, but the pronunciation is an essential golden rule. Finally, there are also the words that do not take as a rule which have different sounds in writing and speaking. Such kind of words are pre-defined their phoneme sequences in the file. In this file, there are 1250 words and the system will find the words which are not covered by the proposed rules and fetch the appropriate phoneme sequences. *Table 6* shows predetermined phoneme sequences of some words.

2.3Building phoneme speech database

In the process of speech synthesis step, required speech units are fetched from the 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 phonemes, syllables, words, di-phones and tri-phones. If syllables are concatenated, the necessary syllables should be recorded. In this work, the smallest speech unit, necessary phonemes is recorded and stored in the speech database. Generally, there are five steps in building a phoneme speech database of our system:

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

2.3.1Myanmar phoneme inventory

A phoneme is the basic and a 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 sounds in the only first position such as start with /p/ in *pit* and /b/ for *bit*. These two phonemes /p/ and /b/ can distinguish two different sounds and meaning of words as they are the smallest unit of sounds and that cannot be separated again. In the word "*pet* and *pit*", they are only different in vowel sounds /e/ and /i/. Therefore, /p/, /b/, /e/, /i/ can be defined as phonemes in English language [16]. 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 words, 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 sounds /a/ and /o/. Therefore, /p/, /b/, /a/ and /o/ are phonemes in Myanmar language and the words, ပန်း, ဗန်း and ပုန်း are minimal pair because they have only one differ phone [17]. Myanmar phonemes that are used in this system are counted in detail in the following sub section.

2.3.1.1 Consonant 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 consonants, /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 [16] are as shown in *Table 7*.

Table 7 Consonant phoneme

Place of Articulation (အသံဖြစ်ရာဌာန်)								
Myanmar of Art		Bilabial (နှုတ်ခမ်းနှစ်ခု)	Dental (သွား)	Alveolar (လျှာဖျား)	Palato-alveodar (လျှာပြား)	Palatal (လျှာဖျားပြား)	Velar (လျှာရင်း)	Glottal (အခင်ပိတ်သံ)
Nasal (နှာသံ)		မ[m] မ့[m]		န[n] န့[n̥]	ည ည့[ɲ̥]		င င့[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[ʤ] ကျ[tɕ],ချ[tɕʰ]			
Central	Voiced Unvoiced	ဝ[w] ဖ့[w̥]		ရ[ɹ]		ယ[j]		ဟ[h]
Lateral (နှုတ်ဖက်ဖြစ်ခြင်း)	Voiced Unvoiced			လ[l] လ့[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 syllables that can be pronounced by combining with medials are only 47 syllables in the Myanmar scripts.

2.3.1.2Vowel 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-/ai?/, အိုင်-/ain/, အောက်-/au?/, အောင်-/aun/. There are two kinds of vowels: 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 8*.

Table 8 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 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 stopped 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)"- /kà/ or /kə/, "ကား (car)" - /k aˆ/, "ကာ (cover)" - /ka a¯/ and "ကတ် (card)"- /k aʔ/.

2.4Choose a speaker

For any speech-based systems, it is important to record the best sound quality possible since each minor distortion can often occur in complex speech. Therefore, the 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 have to be ensured. The speech quality depends upon the quality of the

recorded sound so a professional Myanmar female speaker is selected for recording.

2.5Preparing text for recording

Then, prepare the texts of the phonemes (consonants, vowels, combined medials 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 [kə leˆ], [kʰə ɹe¯], [mə n ɛˀ pʰj ɴ̃¯] instead of [kà leˆ], [kʰà ɹ e¯], [mà nɛˀ pʰjɴ̃¯]. 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¯ bɴ̃ˆ] (star flower)" in which "ခ-/ kʰə /" is minor syllable 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.

2.6Segment and label phoneme

Finally, the recorded sound files are segmented and labelled for the next processing. For this purpose, the sound editing software "*WavePad Sound Editor*" has been used. The phoneme sounds have been labelled manually one by one, after carefully listening and analysing 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 (ခရေပန်း) according to the above example word, only the very first syllable ("ခ") is segmented and label. 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 sounds of the word are segmented in this way to get extra 24 minor consonants sounds.

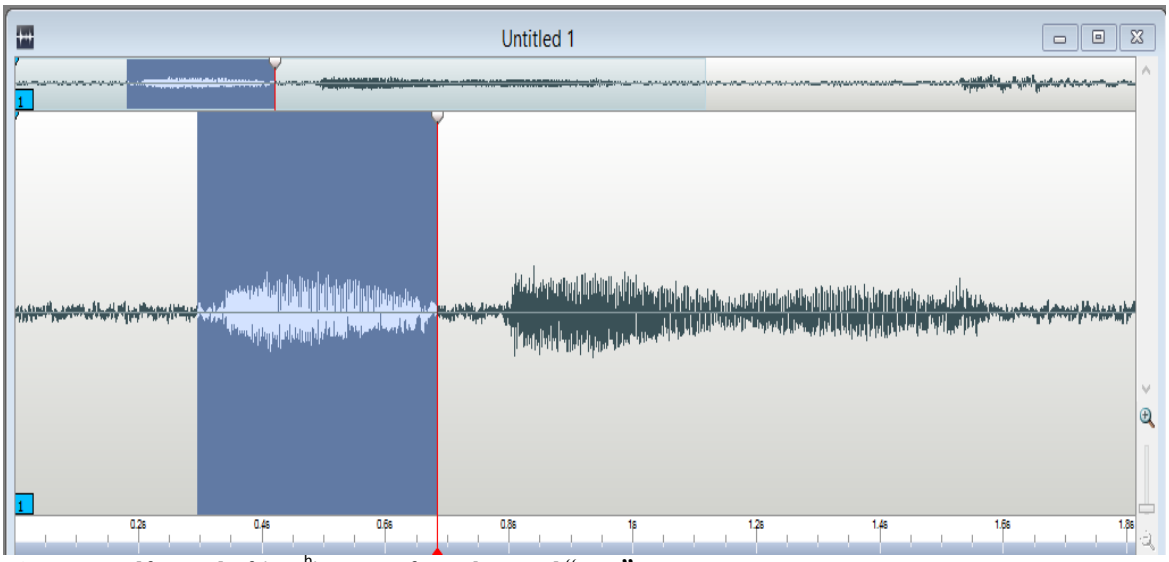

Figure 1 Half-sound of (ခ-kʰ) cutting from the word "ခရေ"

2.7Store the phoneme

In this step, the segmented speech wave files are stored in the Myanmar phoneme speech database. At first, there are only 133 phoneme speech sounds in [14] that did not consider half-sounds. Now, we consider this kind of sounds so that in this database, there are 157 phoneme speech files by adding extra 24 minor consonants as shown in *Table 9*. These phonemes can make sound for all kinds of Myanmar texts. Therefore, this amount of phoneme is

significantly reduced than the previous di-phone based MTTS system so that it is very suitable for resourcing limited devices.

Table 9 Number of sound files 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**

2.8Proposed phoneme concatenation algorithm

The concatenative speech synthesis is the most useful method in the field of TTS system and it can generate high quality natural speech because of the concatenation of selecting the pre-recorded speech segments [18]. In our proposed system, we used concatenative method and phoneme is selected as the basic unit 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/" comes from the combination of consonant phoneme as "ကျ-/tɕ/" (KYA) and vowel phoneme "ောင်း-/aũˆ/" (AUNG:). Therefore, every sound in Myanmar language is made up of these phonemes combinations. Phoneme based Myanmar TTS system has been developed in [14]. In their paper, the syllable that has half sound did not consider. Therefore, we extend the previous phoneme concatenation method as shown in ALGORITHM I.
According to the algorithm 1, 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 CP * threshold*). The threshold value may be different based on the consonant. If the consonant is aspirated consonant such as "ဈ, ဆ, ဏ, 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"), fetches the related half-sound voice file and do not modify anything as well as the vowel is "à" than fetch their corresponding consonant files.

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 words, phrases 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.

Algorithm 1: Phoneme based concatenation method
Input: consonant phoneme and vowel phoneme (CP_i, VP_i)
Output: concatenated voice file $Voice = \{Voice_1, Voice_2, \ldots\}$
Begin

1. CP: a set of consonant phoneme
2. VP: a set of vowel phoneme
3. WF_1: modified consonant file
4. WF_2: modified vowel file
5. C_SP: start position of consonant wave file
6. C_EP: end position of consonant wave file
7. V_SP: start position of vowel wave file
8. V_EP: end position of vowel wave file
9. S: a set of concatenated wave files
10. **procedure** *phoneme_ concatenation* (CP_i, VP_i) {
11. $C_i\_SP \leftarrow 0$
12. $C_i\_EP \leftarrow$ *duration of* C_i * *threshold value*
13. **if** (VP_i. equals ("ə")) **then**
14. $WF_1 \leftarrow$ *fetch the half sound file of* CP_i
15. **else if** (VP_i. equals ("à")) **then**
16. $WF_1 \leftarrow$ *fetch the sound file of* CP_i
17. **else**
18. {
19. **if** (CP_i is aspirated consonant) **then**
20. $WF_1 \leftarrow$ trim (C_i_SP, C_i_EP)
21. **else**
22. $WF_1 \leftarrow$ trim (C_i_SP, C_i_EP)
23. }
24. **end if**
25. $V_i\_SP \leftarrow$ *duration of* VP_i * *threshold value;*
26. $V_i\_EP \leftarrow VP_i$ *duration* - V_i_SP
27. $WF_2 \leftarrow$ trim (V_i_SP, V_i_EP)
28. $Voice_i \leftarrow$ create voice file by concatenating WF_1 and WF_2
29. predicting pause for each .mp3 file by using duration modelling
30. $Voice \leftarrow Voice \cup Voice_i$

End

3.Result

The proposed new phoneme concatenation method is tested in the already developed phoneme based MTTS system [14]. The two-quality measures are considered for testing the quality of MTTS system. They are intelligibility test which measures how much the user understands 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, 100 sentences which are from any domain, such as travelling, greeting, story and so on. The shortest length of the sentence is about 7 and the longest length is round about 24. To test these sentences, the 25 evaluators who are students and teachers from the University of Computer Studies Mandalay (UCSM) are requested to help and test the speech quality of MTTS system.

3.1Naturalness test

For the naturalness test, we used two methods for evaluation: mean opinion score (MOS) and degrade mean opinion score (DMOS). The MOS method is the most useful and simplest method to test the quality of MTTS system. It has five level scales in MOS: bad (1), poor (2), fair (3), good (4) and excellent (5) [19].

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.

$$\text{MOS} = \frac{\sum_{j=1}^{M} s_{ij}}{N} \quad (1)$$

3.1.1 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 2*.

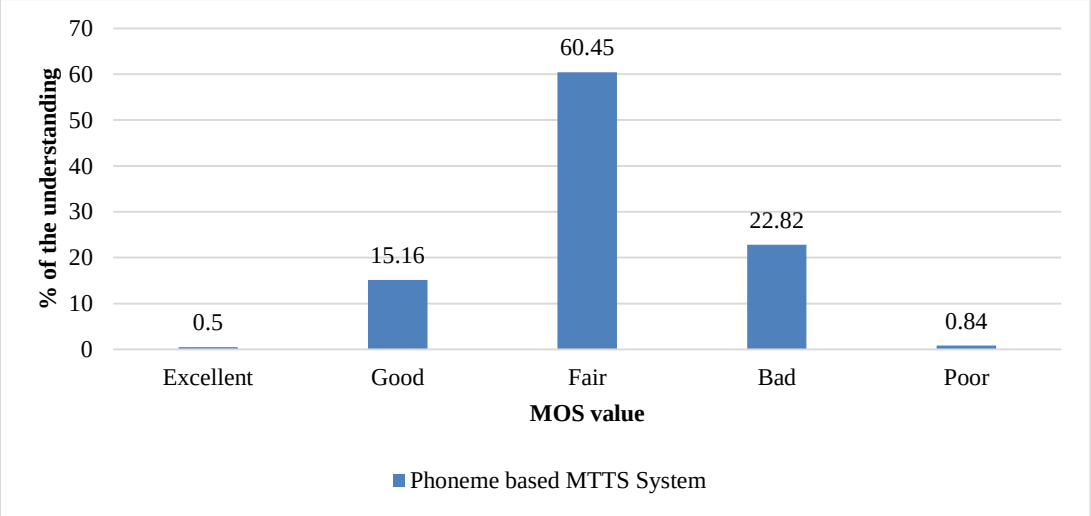

Figure 2 MOS value of naturalness test

3.1.2 DMOS for naturalness test

In DMOS method, evaluators need to listen to synthetic as well as a natural voice in random order without having prior information about the type of voice i.e. natural or synthetic, this is to avoid biased scoring. The motive of this method is to judge the voice in terms of naturalness. Average of scores given to natural sentences and synthesized sentences separately by each evaluator, will be calculated.

For each evaluator, the following equation is used to calculate DMOS. It has also five scales from 1-5 ((5)system sounds like humna, (4) robotic sound but reading correctly, (3) reading sentences with less broken words in robotic manner, (2) almost every word broken and (1) extremely intolerable).

$$\text{Normalized score of synthesized to natural} = \frac{\text{Synthetic voice score}}{\text{Natural voice score}} \times 5 \quad (2)$$

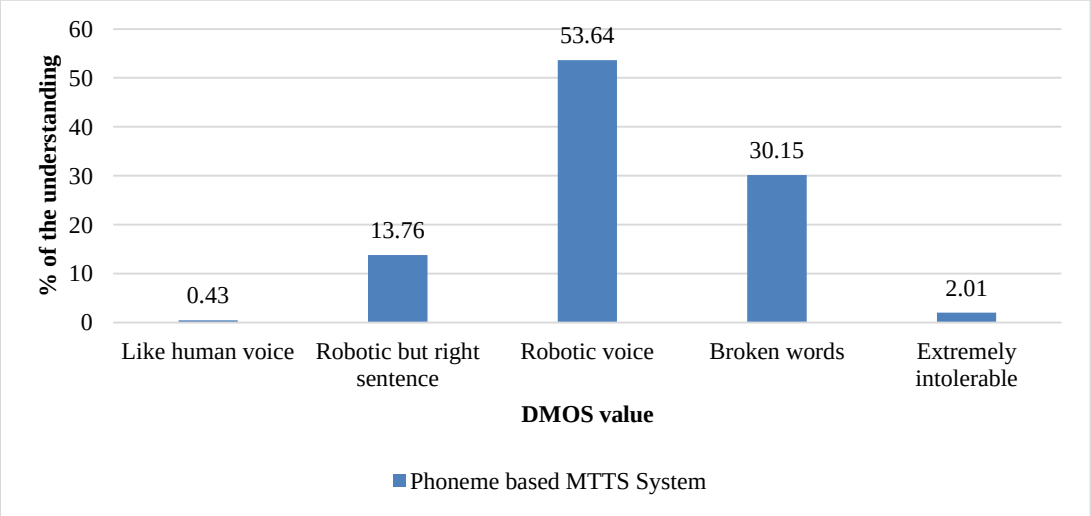

Figure 3 DMOS value of naturalness test

For the DMOS test, the evaluators are asked the question of the sound quality how much the listeners feel the voice is similar to the real person by listening the mixture of synthetic and natural voice. Regarding their answers, 0.43% of listeners thought about the output speech is very natural, 13.76% considered the speech are natural and 53.64% of listener identified the voice are acceptable. Around 30.15% assumed the speech output is needed to get more naturalness and only 2.01% though the worst. The average score for DMOS test is 2.26 as shown in *Figure 3*.

3.2Intelligibility test

Intelligibility is one of the important factors affecting speech quality. We can calculate intelligibility either by MOS as shown in Equation 1 and word error rate (WER) as shown in Equation 3.

3.2.1 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 cases, 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 of intelligibility test is 3.58.

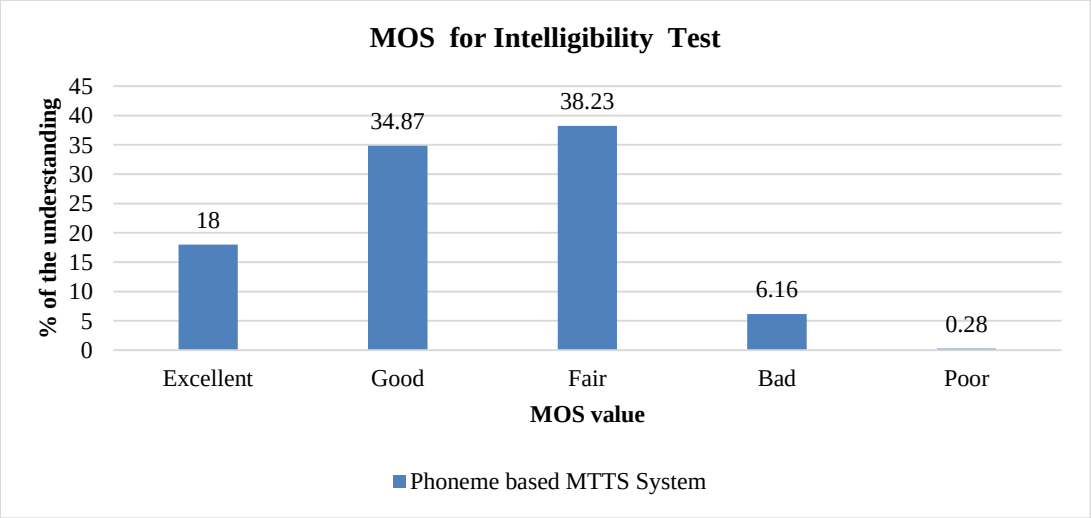

Figure 4 MOS value of intelligibility test

3.2.2 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. According to their results, we calculated WER by using the following equation and the WER value is only 8% on intelligibility test.

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

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 9.79 MB. However, it is not significant and it is still suitable for the resource limited devices.

4.Discussion

Actually, the Myanmar speech could be synthesized by using the defined 133 phoneme speech units with the proposed phoneme concatenation method. In this case, the half sound, schwa vowel sound did not consider. So, this paper extended the previous method to get this kind of half sound. Therefore, final MOS scores of intelligibility test is increased up to 3.58 and the MOS score of naturalness test is also increased up to 2.91 and DMOS is 2.26. According to the evaluation results, the system got very 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 analysed, some vowel combination cannot generate the clear voice, especially, the non-nasalized vowel such as “အေ၊အေး၊အော့၊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 "မေးပေးတို့ကယ်" is generated as "AYE, PAY, OOE, AEL" as the vowel sound instead of "MAY, PAY, TOE, KAL". Because of such vowels, the speech quality needs to go high MOS scores. Therefore, in the future work, we will consider to get the more quality speech file for such vowel by considering the extra speech units.

5.Conclusion

This paper presented the proposed extended phoneme concatenation method in Myanmar text to speech system by considering half sound of the consonant, so that inserting rules for schwa vowel are described. A 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 supports MTTS system for resource limited devices. According to the experimental results, we achieved the highest intelligibility and naturalness results than the previous one. To get more natural output speech, we are considering to use signal processing techniques or adding extra sound files such as onset sound in the future research work.

Acknowledgment

The authors wish to thank all members in the Artificial Intelligence Lab and all the evaluators at the University of Computer Studies, Mandalay and Myanmar.

Conflicts of interest

The authors have no conflicts of interest to declare.

References

[1] Verma A, Singh DK. Robust assistive reading framework for visually challenged. International Journal of Image, Graphics and Signal Processing. 2017; 9(10):29-37.

[2] Black AW, Campbell N. Optimising selection of units from speech databases for concatenative synthesis. CSTR; 1995.

[3] Conkie A. Robust unit selection system for speech synthesis. In Joint Meeting of ASA/EAA/DAGA, Berlin, Germany. 1999.

[4] 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. 373-6). IEEE.

[5] 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.

[6] Douke M, Hayashi M, Makino E. A study of automatic program production using TVML. Short Papers and Demos, Eurographics. 1999; 99:42-5.

[7] Ramteke GD, Ramteke RJ. Efficient model for numerical text-to-speech synthesis system in marathi, hindi and english languages. International Journal of Image, Graphics & Signal Processing. 2017; 9(3):1-13.

[8] Bakhsh NK, Alshomrani S, Khan I. A comparative study of Arabic text-to-speech synthesis systems. International Journal of Information Engineering and Electronic Business. 2014; 6(4):27-31.

[9] Kasparaitis P, Kančys K. Phoneme vs. diphone in unit selection TTS of Lithuanian. Baltic Journal of Modern Computing. 2018; 6(2):162-72.

[10] Jannati MJ, Sayadiyan A. Part-syllable transformation-based voice conversion with very limited training data. Circuits, Systems, and Signal Processing. 2018; 37(5):1935-57.

[11] Myanmar language commission, Myanmar grammar, 30th year special edition. University Press, Yangon, Myanmar; 2005.

[12] Win KY, Takara T. Myanmar text-to-speech system with rule-based tone synthesis. Acoustical Science and Technology. 2011; 32(5):174-81.

[13] Soe EP, Thida A. Text-to-speech synthesis for Myanmar language. International Journal of Scientific & Engineering Research. 2013; 4(6):1509-18.

[14] Hlaing CS, Thida A. Phoneme based Myanmar text to speech system. International Journal of Advanced Computer Research. 2018; 8(34):47-58.

[15] Maung ZM, Mikami Y. A rule-based syllable segmentation of Myanmar text. In proceedings of the IJCNLP-08 workshop on NLP for less privileged languages 2008 (pp. 51-8).

[16] Acoustic phonetics and phonology of the Myanmar language. School of Human Communication Sciences, La Trobe University, Melbourne, Australia, 2007.

[17] Myanmar Language Commission, Myanmar Grammar, 30th Year Special Edition, University Press, Yangon, Myanmar, 2007.

[18] Lemmetty S. Review of speech synthesis technology. Helsinki University of Technology. Department of Electrical and Communications Engineering. Master's Thesis. 1999.

[19] http://tdil-dc.in/undertaking/article/449854TTS_Testing_Strategy_ver_2.1.pdf. Accessed 12 May 2018.

Chaw Su Hlaing is a Ph.D. student at Artificial Intelligence Lab, Faculty of Computer Science, University of Computer Studies, Mandalay and Myanmar. Her research interests are Web Data Mining, Digital Signal Processing, Natural Language Processing and Liguistic Research. She is currently working in the research on Speech Synthesis in Myanmar Language. She received Bachelor of Computer Science and Master of Computer Science from the Computer University, Mandalay, and Myanmar.
Email: chawsuhlaing@ucsm.edu.mm

Aye Thida is a Professor, at Faculty of Computer Science, Artificial Intelligence Lab, University of Computer Studies, Mandalay and Myanmar. Her research interests are Machine Translation, Text-to-Speech System and Big Data Management. She is currently working in 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.
Email: ayethida@ucsm.edu.mm