Analysis of Word-Based and Unit-Based Diphone Concatenation for Myanmar Text-To-Speech

Ei Phyu Phyu Soe, Aye Thida
University of Computer Studies, Mandalay, Myanmar
eiphyu2soe@gmail.com

Abstract

Speech synthesis system is a popular field in natural language processing of computer science for various languages. The process of speech synthesis is to produce the human-like speech from corresponding language and it can be divided into two key phases such as high-level and low-level text-to-speech synthesis. In high-level synthesis, the input text is converted into such form that the low-level synthesizer can produce the output speech. Speech synthesis can be used three types: Articulatory, Format and Concatenative synthesis. Concatenative speech synthesis is the most popular technique among three types and it is the easiest technique to synthesis the speech. Diphone-concatenative method is one of the popular methods in nowadays because of speech quality is better than other techniques and complexity is less than others. This paper gives the analysis of two type of Diphone-concatenation; word-based and unit-based synthesis for Myanmar language by applying Pitch Synchronous Overlap and Add (PSOLA) algorithm to smooth the joints of the speech signals. This paper describes the Time-Domain Pitch Synchronous Overlap and Add method in Diphone-concatenation speech synthesis and it is to maintain the consistency and accuracy of the pitch marks of the speech signal and Diphones database with integrated vowels and consonants of Myanmar language. This paper also describes the building of Myanmar Diphone database for concatenation method. Diphone Database for Myanmar pronunciations is constructed in this paper to reduce the ambiguity in pronunciations.

1. Speech Synthesis

Natural language is an integral part of human lives. Language serves as the primary vehicle by which people communicate and record information. NLP process is important part of speech synthesis. There are other NLP techniques related to speech synthesis that, although they are not essential, their use may improve the overall quality of the resulting speech, yielding a more natural impression to the users.

Speech is the primary means of communication between people. Speech synthesis, automatic generation of speech waveforms, has been under development for several decades [8]. Recent progress in speech synthesis has produced synthesizers with very high intelligibility but the sound quality and naturalness still remain a major problem. However, the quality of present products has reached an adequate level for several applications, such as multimedia and telecommunications. With some audiovisual information or facial animation (talking head) it is possible to increase speech intelligibility considerably [2]. Some methods for audiovisual speech have been recently introduced by for example in [7] and [2].

The text-to-speech (TTS) synthesis procedure consists of two main phases. The first one is text analysis, where the input text is transcribed into a phonetic or some other linguistic representation, and the second one is the generation of speech waveforms, where the acoustic output is produced from this phonetic and prosodic information.

These two phases are usually called as high- and low-level synthesis. A simplified version of the procedure is presented in Figure1. The input text might be for example data from a word processor, standard ASCII from e-mail, a mobile text-message, or scanned text from a newspaper. The character string is then preprocessed and analyzed into phonetic representation which is usually a string of phonemes with some additional information for correct intonation, duration, and stress. Speech sound is finally generated with the low-level synthesizer by the information from high-level one.

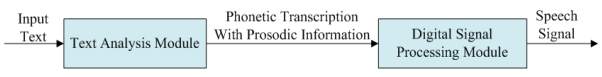


Figure1. Simple Speech synthesis procedure

2. Text-To-Speech Synthesis

The process of Myanmar text-to-speech synthesis system is the human-like speech from the input Myanmar text to spoken Myanmar language. Since this generally requires great language knowledge, the context where the text comes from, a deep understanding of the semantics of the text content and the relations. However, many research and commercial speech synthesis systems developed have contributed to our understanding of all these phenomena, and have been successful in various respective ways for many applications such as in speech-to-speech machine translation, interactive voice response systems, reading software for the blind, linguistic research and language teaching center [3].

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

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

3. Myanmar Language

Myanmar writing does not use white spaces between words or between syllables. Thus, the computer has to determine syllable and word boundaries by means of an algorithm such as finite-state and rule-based. Moreover, a Myanmar syllable can be composed of multiple characters. Syllable segmentation is the process of determining word boundaries in a piece of text.

Myanmar belongs to the Lolo-Burmese sub-branch of the Tibeto-Burmese branch of the Sino-Tibetan language family. Myanmar script draws its source from Brahmi script which flourished in India from about 500 B.C. to over 300AD. Myanmar is a tonal language. This means that all syllables in Myanmar have prosodic features that are an integral part of their pronunciation. Prosodic contrasts involve not only pitch, but also phonation. Standard Myanmar is based on the dialect spoken in the lower valleys of the Irrawaddy and Chindwin rivers. It is spoken in most of the country with slight regional variations. In addition, there are other regional variants that differ from standard Myanmar in pronunciation and vocabulary. All dialects are mutually intelligible. In addition, there are two registers: a formal and a colloquial one. The formal register is used in official publications, radio and TV broadcasts, literary works, and formal speech. The colloquial register is used in daily communications. In Myanmar there are 8 main races and 135 sub races under the main races. Myanmar (Burmese) is official language in Myanmar. We choose the formal register for Myanmar Speech Synthesis.

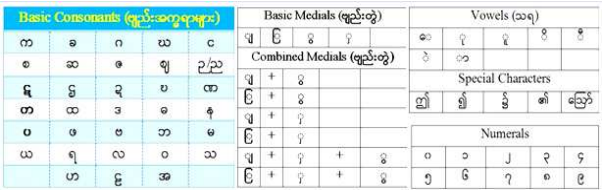
A Myanmar text is a string of characters without explicit word boundary markup, written in sequence from left to right without regular inter-word spacing, although inter-phrase spacing may sometimes be used. Myanmar characters can be classified into three groups: consonants, medials and vowels. The basic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

There are 34 basic consonants in the Myanmar script. They are known as “Byee” in the Myanmar language [13]. Consonants serve as the base characters of Myanmar words, and are similar in pronunciation to other Southeast Asian scripts such as Thai, Lao and Khmer.

Medials are known as “Byee Twe” in Myanmar. There are 4 basic medials and 6 combined medials in the

Myanmar script. The 10 medials can modify the 34 basic consonants to form 340 additional multi-clustered consonants. Therefore, a total of 374 consonants exist in the Myanmar script, although some consonants have the same pronunciation.

Table1. Myanmar Characters

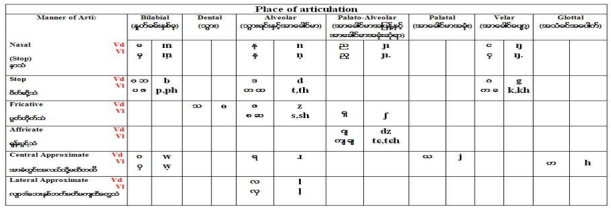


Vowels are known as “Thara”. Vowels are the basic building blocks of syllable formation in the Myanmar language, although a syllable or a word can be formed from just consonants, without a vowel. Like other languages, multiple vowel characters can exist in a single syllable. The various kinds of Myanmar characters are shown in Table1.

4. Phonology of Myanmar Characters

A phoneme is the smallest unit that distinguishes words and morphemes. Therefore, changing a phoneme of a word to another phoneme produces a different word or a nonsense utterance, whereas changing a phone to another phone, when both belong to the same phoneme, produces the same word with an odd or an incomprehensible pronunciation. Phonemes are not physical segments themselves, but mental abstractions of them [11]. Different acoustic realizations of a phoneme are called allophones. The acoustic characteristics of phonemes come from the vocal tract movement during their articulation. There are three types of phonetic parameters in phonology of Myanmar language: first is place of articulation, second is articulator and third is manner of articulation [4]. The pronunciation of Myanmar words depend on these parameters. A phoneme is a contrastive unit in the sound system of a particular language. It is a minimal unit that serves to distinguish between meanings of words. Phoneme can pronounce in one or more ways, depending on the number of allophones. It can represent between slashes by convention. Table2 describes the inventory of Myanmar consonant phonemes defined by the International Phonetic Association (IPA) [5] and [4].

Table2. Phonology of Myanmar Characters



4.1 Structure of Myanmar Phonological

The Myanmar language uses a rather large set of 50 vowel phonemes, including diphthongs, although its 22 to 26 consonants are close to average. Some languages, such as French, have no phonemic tone or stress, while several of the Kam-Sui languages have nine tones, and one of the Kru languages, Wobe, has been claimed to have 14, though this is disputed. The most common vowel system consists of the five vowels /i/, /e/, /a/, /o/, /u/. The most common consonants are /p/, /t/, /k/, /m/, /n/. Relatively few languages lack any of these, although it does happen: for example, Arbic lacks /p/, standard Hawaiian lacks /t/, Mohawk and Tlingit lack /p/ and /m/, Hupa lacks both /p/ and a simple /k/, colloquial Samoan lacks /t/ and /n/, while Rotokas and Quileute lack /m/ and /n/ [11]. Table3 shows the phonetic signs of 50 Myanmar vowels to pronounce the Myanmar words. These 50 phonemes show the basic symbol with four tone levels [4].

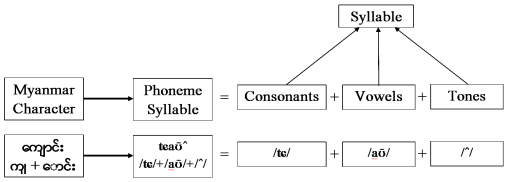


Figure2. Combination of Phoneme Syllable

Phonology is how speech sounds are organized and affect one another in pronunciation. The combination of consonant phoneme and a vowel phoneme produces a syllable in Figure2. The phonetic alphabet is usually divided in two main categories, vowels and consonants. Vowels are always voiced sounds and they are produced with the vocal cords in vibration, while consonants may be either voiced or unvoiced. Vowels have considerably higher amplitude than consonants and they are also more stable and easier to analyze and describe acoustically. Because consonants involve very rapid changes they are more difficult to synthesize properly [3].

Table3. Phonology of Myanmar Vowels

Basic Symbols		Non-Nasalized								Nasalized					
		Tone1		Tone2		Tone3		Tone 4		Tone1		Tone2		Tone3	
i	IY	i˜	IY1	i˜	IY2	i˜	IY3	i˙	IY4	ĩ˜	IH1	ĩ˜	IH2	ĩ˙	IH3
e	AE	e˜	AE1	e˜	AE2	e˜	AE3	ei˙	EY1	eĩ˜	EY1	eĩ˜	EY2	eĩ˙	EY3
ɛ	EH	ɛ˜	EH1	ɛ˜	EH2	ɛ˜	EH3	ɛ˙	EH4	aĩ˜	AY1	aĩ˜	AY2	aĩ˙	AY3
								ai˙	AY4						
a	AA	a˜	AA1	a˜	AA2	a˜	AA3	a˙	AA4	ã˜	AH1	ã˜	AH2	ã˙	AH3
ɔ	AO	ɔ˜	AO1	ɔ˜	AO2	ɔ˜	AO3	aʊ˙	AW4	aũ˜	AW1	aũ˜	AW2	aũ˙	AW3
o	OO	o˜	OO1	o˜	OO2	o˜	OO3	oʊ˙	OW4	oũ˜	OW1	oũ˜	OW2	oũ˙	OW3
u	UW	u˜	UW1	u˜	UW2	u˜	UW3	ʊ˙	UH4	ũ˜	UH1	ũ˜	UH2	ũ˙	UH3

4.2 Structure of Myanmar Tonal

Myanmar language has four tones and a simple syllable structure that consists of an initial consonant followed by a vowel with an associate tone. This means all syllables in Myanmar have prosodic features. Different tone makes different meanings for syllables with the same structure of phonemes. In the Myanmar writing system, a tone is presented by a diacritic mark [5] and [4].

The fundamental frequency as shown in figure1 rises gradually from Tone 1 to Tone 4. Tone 1 starts at a relatively level range and tends to go down slightly; Tone 2 starts at a relatively level range, goes up, and then falls down relatively low; Tone 3 starts at a relatively high range, usually higher than or as high as the peak of Tone 2, and falls down relatively low; Tone 4 stats at a high range, frequently higher or as high as the peak of Tone 2 and falls low, but not as low as Tone 3 because it stops very suddenly before it can drop lower [4]. The general contrastive features of the four phonological tones offered by the analysis of their fundamental frequency can be described as Figure3:

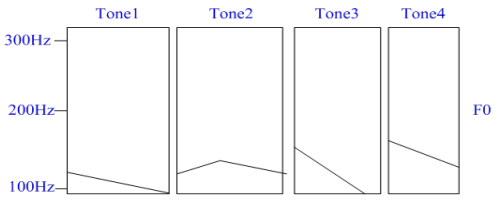


Figure3. Four Tones of Myanmar Language

There are four tones in Myanmar language. The lengths of tones are:

- Tone 1 has 18.50 Cs
- Tone 2 has 21.03 Cs
- Tone 3 has 15.44 Cs and
- Tone 4 has 10.35 Cs.

So Myanmar toneme is described with the variety of rate or duration. Length of the tone is defined as rate or duration. Tone 2 is defined as a longest rate and tone 4 is defined as a shortest rate in these four tones.

5. Overview of Myanmar Text-To-Speech

The TTS system has two major parts: a natural language processing (NLP) which reads the input text and translates it into a phonetic language and a digital signal processing (DSP) that converts the phonetic language into spoken speech. A TTS system generally consists of four modules, namely text analysis, phonetic analysis, prosody analysis and speech synthesis, as shown in Figure4.

In the step of text analysis, there have three parts: syllable normalization, number converter and syllable segmentation. Syllable normalization is performed from non-standard words to standard words. The purpose of number converter is to convert the number to textual versions. The non standard words are tokens like numbers, which need to be expanded into sequences of Myanmar words before they are pronounced. Number expands to string of words representing cardinal. The input text is analyzed to segmented Myanmar text like a syllable. Syllable segmentation is the process of identifying syllable boundaries in a text [14]. In this paper [14], syllable segmentation is based on the trigram statistics rule based algorithm. A rule based approach of syllable segmentation algorithm for Myanmar text is proposed. Segmentation rules were created based on the syllable structure of Myanmar script and a syllable segmentation algorithm was designed based on the created rules. The syllable structure of Myanmar sentence is expressed in finite state automatic (FSA) algorithm as shown in below:

Syllable::= C{M}{V}{F} | C{M}VA | C{M}{V}CA[F] | E[CA][F] | I | D

Phonetic analysis is also called G2P (Grapheme to Phoneme). Grapheme to Phoneme conversion translates the syllable of Myanmar text to phonetic sequence. It determines the pronunciation of a syllable based on its spelling. It also analyzes the best sequence of phonemes for words, numbers and symbols and converts into phonetic sequences. We have constructed the Myanmar phonetic dictionary to generate the phoneme sequence and to pronounce these phonemes [5].

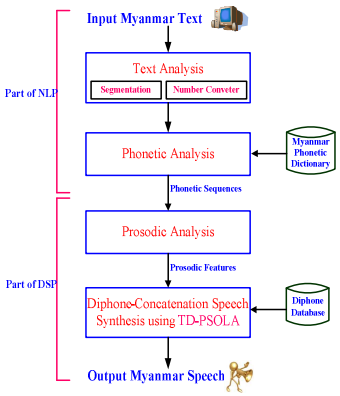


Figure4. Overview of Text-To-Speech Synthesis

The phonetic sequences are analyzed to produce the prosodic features by applying the phonological rules in prosodic analysis step. It is the module to analyze duration and intonation such as pitch variation, syllable length to create naturalness of synthetic speech. The combination of consonant phoneme and a vowel phoneme produces a syllable. The phonetic alphabet is usually divided in two main categories, vowels and consonants. Vowels are always voiced sounds and they are produced with the vocal cords in vibration, while consonants may be either voiced or unvoiced. Vowels have considerably higher amplitude than consonants and they are also more stable and easier to analyze and describe acoustically. Because consonants involve very rapid changes they are more difficult to synthesize properly [6].

Synthesized speech can be produced by several different methods. The methods are usually classified into three groups such as articulatory synthesis, formant synthesis and concatenative synthesis. The aim of this paper is to improve the quality of word-based and unit-based of Myanmar Text-To-Speech by applying the Concatenative speech synthesis using TD-PSOLA (Time Domain-Pitch Synchronous Overlap and Add) algorithms.

6. Diphone-Concatenation Speech Synthesis

The synthetic voice that imitates human speech from plain text is not a trivial task, since this generally requires great knowledge about the real world, the language, the context where the text comes from, a deep understanding of the semantics of the text content and the relations that underlie all these information. This chapter presents a detailed discussion of the Diphone inventory of Myanmar language that was recorded with "Wave Pad Sound Editor" and the concatenation method of speech file from Diphone inventory.

Concatenative Synthesis is the most used technique of today, where segments of speech are tied together to form a complete speech chain. The speech output is produced by coupling segments from the database to create the sequence of segments. This technique requires a bit of manual preparation of the speech segments [9].

A diphone synthesis uses a minimal speech database containing all the Diphones occurring in a given language. Diphones are speech units that begin in the middle of the stable state of a phone and end in the middle of the following one. The number of Diphone depends on the possible combinations of phonemes in a language. In Diphone synthesis, only one example of each Diphone is contained in the speech database. The quality of the resulting speech is generally not as good as that from unit selection but more natural-sounding than the output of formant synthesizers. Diphone synthesis suffers from the robotic-sounding quality of formant synthesis. In order to build a Diphone database, the following questions have to be answered and determined: What Diphone pairs exist in a language and what carrier words should be used? The answers for these questions are very language independent.

Concatenation has two types: word-concatenation and unit concatenation. Different sizes have different advantages and disadvantages. The technique used in speech synthesis part is Diphone concatenation technique, because this is the easiest method to implement among others and it can be considered as successful. Articulatory synthesis requires very good understanding of speech processing and an intensive research for obtaining necessary parameters. After obtaining parameters a good modeling and a lot of calculation is needed. Therefore, we decided that this kind of systems require more than what a 2 people team can do in about a year. Formant synthesis also requires a good background on signal processing and intensive research for obtaining necessary parameters. Obtaining signal processing background and these parameters may require years for our team, therefore we considered this method also as not suitable for us. However, concatenation methods do not require much background, by simply recording some words; a very simple system could be created.

Word concatenation is a good method for systems that require limited number of vocabulary. Diphone Concatenation method with word has the complete prosodic features because of recording the training data

and it has limited vocabulary. Only necessary words are recorded and they are used in runtime, no serious job is done in concatenation. Word-concatenation is limited the vocabulary and it cannot produce the every Myanmar words for every sentences but the pronunciation has the prosodic features and it has good naturalness.

The unit-selection synthesis uses large speech databases. More than one hour of recorded speech is usually used. During database creation, each recorded utterance is segmented into some or all of the following: individual phones, syllables, morphemes, words, phrases, and sentences. The division into segments can be performed using a number of techniques, for instance clustering, using a specially modified speech recognizer; it can also be done by hand, using visual representations such as the waveform and spectrogram. The unit-selection technique gives naturalness due to the fact that it does not apply digital signal processing techniques to the recorded speech, which often make recorded speech sound less natural. However, maximum naturalness often requires unit selection speech databases to be very large.

The main difference between the two types of Concatenative Synthesis lies in the size of the units being concatenated. Both methods store the pre-recorded speech units in a database, from which the concatenation originates. Those parts of utterances that have not been found, processed and stored in the database are built up from smaller units [10].

6.1 Time-Domain PSOLA

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

1. original speech signal is divided into separate short analysis signal
2. the modification of each analysis signal to synthesis signal and
3. the synthesis step where these segments are recombined by means of overlap-adding [10].

The purpose of TD-PSOLA (Time-Domain) is to modify the pitch or timing of a signal as shown in Figure5. The process of the TD-PSOLA algorithm is to find the pitch points of the signal and then apply the hamming window centered of the pitch points and extending to the next and previous pitch point. If the speeches want to slow down, the system defines the frame to double. If the speeches want to speed up, the system removes the frames in the signal.

Begin

Find the pitch points of the signal

Apply Hanning window centered on the pitch points and extending to the next and previous pitch point

Add waves back

To slow down speech, duplicate frames

To speed up, remove frames

Hanning windowing preserves signal energy

End

Figure5. TD-PSOLA Algorithm

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

7. Word-Based Diphone Concatenation

The basic idea behind building Myanmar Diphone databases is to explicitly list all possible phone-phone transitions in a language. One technique is to use target words embedded carrier sentences to ensure that the Diphones are pronounced with acceptable duration and prosody. Speech synthesis finds the corresponding pre-recorded sounds from its database and tries to concatenate them smoothly. It uses an algorithm like TD-PSOLA (Time-Domain Pitch Synchronous Overlap and Add) to make a smooth pass in Diphone. PSOLA method takes two speech signals. One of these signals ends with a voiced part and the other starts with a voiced part. PSOLA changes the pitch values of these two signals so that pitch values at both sides become equal. The advantage of this technique is to obtain a better output speech when compared to other techniques [12].

The structure of word-based Diphone database constructs with Arpabet signs to understand the retrieving phonemes. Firstly, training sentences pre-record with the wave pad sound editor and we

segment them according the TDPSOLA algorithm. And then we label them with pair of arpabet signs according segmental sound file. A Diphone represents the end of one phoneme and the beginning of another. This is significant because there is less difference in the middle of a phoneme than there is at the beginning and ending edges. The problem is that there is not complete for every Myanmar words because word-based Diphone database is created with the training Myanmar sentences and it greatly increases the size of database if we want to get the every Myanmar words. There are 1500 Diphone pairs for the 100 training sentences in the word-based Diphone-concatenation.

The task of word-based diphone labeling is a time giving task because the two words are labeled as the two diphones by segmenting with the time-domain method. The prerecorded sentences are attached with the labeled diphones in the word-based diphone database. The advantage of this concatenation is getting the sound like a human speech and the naturalness is high quality. The disadvantage of this is giving up many times to record together the segmented sound files and labeled the diphone pairs.

The example of the recording and segmenting the diphone for the word-based concatenation database is shown in Table4. The example sentence is "ကျိုက်ထီးရိုးတွင်တစ်တောင်တက်တစ်တောင်ဆင်းဖြင့်အဆင်းအတက်များသည်" and the arpabet pairs are "KYA-AY4- HT-IY2- Y-OO2- TH-IY1- D-EH- D-AW1- T-EH4- D-EH- D-AW1- Z-IH2- PHY-IH3 - AH- SA-IH2- AH-T-EH4- MY-AA2- TH-IY1".

Table4. Defining the Pitch Marks with Time-Domain

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

8. Unit-Based Diphone Concatenation

The idea of the unit-based Diphone database is not similar with word-based Diphone database because the unit-selection is based on the one word to cut the two sound files and it is to explicitly list all possible phone-phone transitions in a Myanmar language. Firstly, we list the possible words for unit-based diphone concatenation and the size of the diphone database has around 10496 diphones in Myanmar language. The unit-based diphone database is constructed on 14 types of diphone lists: Consonants-Consonants, Consonants-Exception Words, Consonants-Vowels, Exception Words-Consonants, Exception Words- Exception Words, Exception Words-Vowels, Vowels-Consonants, Vowels-Exception Words, Consonants-Silence, Exception Words-Silence, Vowels-Silence, Silence-Consonants, Silence-Exception Words and Silence-Vowels. The possible words for unit-based diphone are calculated in following Table5:

Table5. Calculation of Possible Unit-Based Diphone

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

The square of words reduce the square of vowels is equal to the possible unit-based diphone pairs. The 114 words combine with 22 consonants, 42 exception words and 50 vowels and 2500 vowels are the square of 50 vowels. The pair of vowels is not in Myanmar characters for Myanmar words.

The recoding for unit-based must record 10496 diphones into the unit-based diphone database and it performs to get the pronunciations of every Myanmar words. The problem is that it cannot get the good naturalness in the comparison of the word-based concatenation but the database complexity is not large.

The segmentation of sound files is an interesting task with time-domain method. The segmentation depends on the types of voiced and unvoiced characters to regard the length of the sound files. So the lengths of sound files are not same every consonants, vowels and exception words. The lengths of consonants are very short millisecond (ms), the lengths of vowels are short and the lengths of exceptions words are the longest millisecond (ms) among them. After recording the diphone with arpabet sings, we must label the unit-based diphone pairs according to the segmented sound files. In unit-based diphone concatenation, regarding the hanning window is the same as word-based but regarding of

speed is not. The pairs of consonants and other types like that vowels and exceptions are removed the sound files to speed up the speech. The pairs of pause and other types like that vowels, exceptions and consonants are duplicated the sound files to slow down the speech flows. The unit-based diphone concatenation is shown in Table6. The example sentence is "ကျိုက်ထီးရိုးသည် တစ်တောင်တက်တစ်တောင်ဆင်းဖြင့် အဆင်းအတက် များသည်" and the arpabet pairs are "KYA-AY4- HT-IY2- Y-OO2- TH-IY1- D-EH- D-AW1- T-EH4- D-EH- D-AW1- Z-IH2- PHY-IH3 - AH- SA-IH2- AH- T-EH4- MY-AA2- TH-IY1".

Table6. Labeling the Unit-Based Diphone and Sound Files

Myanmar Unit Pairs	Unit-Based Diphone Pairs	Sound Files
PAU-ကျ	PAU-KYA	1.wav
ကျိုက်	KYA-AY4	2.wav
ိုက်-ထ	AY4-HT	3.wav
ထီး	HT-IY2	4.wav
ီး-ရ	IY2-Y	6.wav
ရိုး	Y-OO2	7.wav
ိုး-သ	OO2-TH	8.wav
.....	…..	….
သည်	TH-IY1	9.wav
ည်-ဒ	IY1-D	10.wav
ား-သ	AA2-TH	32.wav
သည်	TH-IY1	33.wav
ည်-PAU	IY1-PAU	34.wav

9. Experimental Results

To test the intelligibility, naturalness and overall quality of the Myanmar Text-to-Speech system developed in this paper, a test for the Myanmar voice was designed. The only concern when choosing the test group is that they should have a good command of the Myanmar language. Since it is difficult to decide what a good command is, it was decided that the participators should have the Myanmar language as their mother tongue. The group consists of 9 people. The age distribution is showed in Table7.

Table7. Age distribution of the listeners

Gender	Age	Amount
Female	40+	1
Female	20-30	6
Male	20-30	1
Male	10-20	1

9.1 Analysis of Unit-Based Diphone Concatenation

The analysis of word-based is based on the corrected answer of the participants' listening sentences. This part contains 25 sentences of 24 pairs of words with different levels of confusability. The listener hears one sentence at a time and marks on the answering sheet. Figure6 shows the perception of the sentences rate after both listening. That is the results of how the listeners caught or heard each word after each listening.

The diagram shows how many words were understood by the listeners after each hearing. One can see that almost all words were recognized from the first hearing, except five words. Those words have the letters "ကျ" "စ" "လျ" "မှ" and "တ". That means that 75 % of these words were correctly understood.

Significance look at the results from the second listening one can see that there is no huge difference. The problematic words or sounds are the same after the second listening. 80 % of the words were correctly obversed. One person did not recognize the same word as he did recognize in the first listening. All listeners perceived the sentences that they already know names and public signboards. They recognized almost 100% in the short sentences and 65% are recognized in the long sentences.

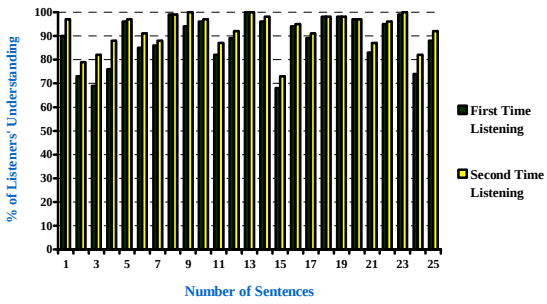


Figure6. Perception of the sentence in Unit-Based

In second part of sentence testing, the participants were supposed to write down on the answer sheet what they think they hear. One sentence is played at a time and they were only allowed to listen to the audio file one time. The sentences are resynthesized by words. Short sentences were chosen in the hope that a person's memory would not affect the end result.

The answers are divided into three categories: correct answers, partially correct answers and incorrect answers. The correct-answers category consists only of complete and correctly recognized sentences with the more than of 95%. The partially-correct-answers category consists of the answers where only one word was misrecognized or not recognized at all and in those cases where I suspect there have been misspellings of the words in answer sheet and this part will record between 80% and 94%. The third part, the incorrect-answers part, consists of theses sentences where the participants could not recognize the meaning of the sentence or of those sentences where the participants recognize two or more words and it is regarded less than 80%.

The Figure7 lists the results of the first time of listening. 48% of these sentences were recognized correctly by all listeners. 33% were partially correctly recognized and the rest, 19%, were not correctly recognized by the listeners.

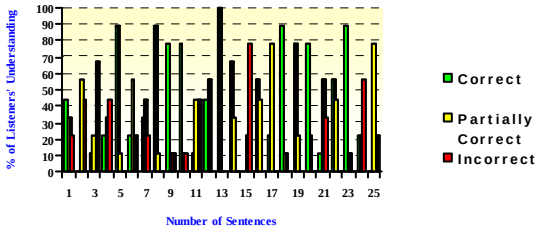


Figure7. Perception of the second part of the sentences in Unit-Based

9.1 Analysis of Word-Based Diphone Concatenation

The testing part of sentence observation for word-based also contains 25 sentences of 24 pairs of words with different levels of confusability. The testing style is the same as unit-based sentence perceptions. The listener hears one sentence at a time and marks on the answering sheet. Figure8 shows the perception of the sentences rate after both listening. That is the results of how the listeners caught or heard each word after each listening.

The following figure shows how many words were understood by the listeners after each hearing. Significance look at the results from the second listening one can see that there is no huge difference. The problem for word-based diphone concatenation is the prosodic features. One person did not recognize the same word as he did recognize in the first listening. All listeners perceived the sentences that they already know names and public signboards. They recognized almost 100% in the short sentences and 85% are recognized in the long sentences.

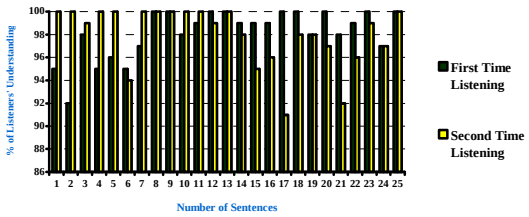


Figure8. Perception of the first part of the sentences

In second part of sentence testing, the participants were supposed to write down on the answer sheet what they think they hear. In this part, the listeners more understand in unit-based but the intelligibility of speech quality is not good. The advantage of this part can recognize the answers by guessing the next word first exactly heard speech. The answers are divided into three categories: correct answers, partially correct answers and incorrect answers. The Figure9 lists the results of the first time of listening. 84% of these sentences were recognized correctly by all listeners. 16% were partially correctly recognized and the rest, there were no percentage in correctly recognized by the listeners.

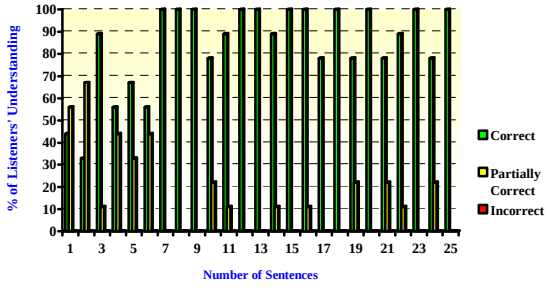


Figure9. Perception of the second part of the sentences in Word-Based

9.2 Testing Sound Quality with Categorical Estimation

The question for this part is “Do you consider the system to be of good sound quality?” and overall quality testing with categorical estimation (CE). There are five level scales in the test of sound quality; pronunciation, speed, naturalness, intelligibility and pleasantness.

9.2.1 Pronunciation

The question, considering pronunciation, is how understand the participants found the voice. 20% of the listeners found the voice good in word-based and 44% of the participants found the voice fair in unit-based and 60% fair in word-based. 56% of the participants thought the voice was audible in unit and 20% of the participants get the audible voice in word. The results of the speech of listening are shown in Figure10.

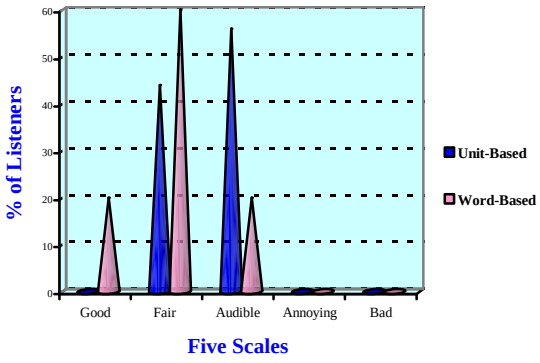


Figure10. Pronunciation of the Speech Quality

9.1.2 Naturalness

The question for the naturalness of the sound quality how much the listeners clear the voice or how much of what the voice said the participants understood, 30% of the listeners are good in word-based. 44% of the participants fair in unit-based and 40% get in word-based. 44% did understand clear in unit-based and 30% get clear in word-based.11% of listeners are neither much nor little that is not very clear is shown in Figure11.

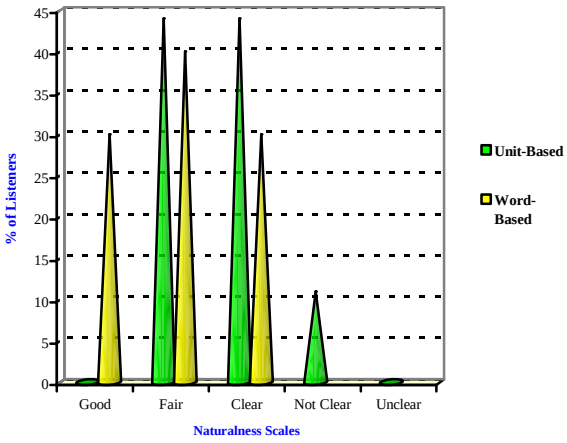


Figure11. Naturalness of the Speech Quality

9.1.3 Intelligibility

The question of this part is "Was the voice easy to understand?" The reason for this question was to establish if the difficulty in understanding is in the voice or in the listeners' lack of knowledge and lexicology. After the time of listening 21% get the good, 56% found very easy level and 23% of the listeners found the voice easy in word-based. 78% of the listeners found the voice easy to understand, while 32% found it neither hard nor easy. Figure12 shows the results of listening.

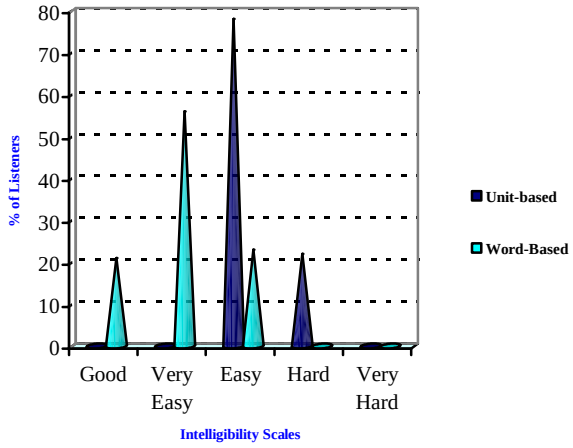


Figure12. Intelligibility of the Speech Quality

10. Conclusion and Future Extension

This paper shows the analysis results of the Myanmar Text-To-Speech synthesis system based on the unit-based and word-based diphone pairs in the Myanmar language. The different steps when building the diphone database have been explained, such as the recording process, the segmentation of the diphones and the database construction itself.

Testing and evaluating, where Myanmar speakers listened to the synthesized voice and decided whether the Myanmar synthesized voice produced good, intelligible and understandable Myanmar speech. The percentages of the pronunciation, naturalness and intelligibility of the word-based speech are better than the unit-based in analysis step of speech quality. The speech quality of Myanmar TTS is tested on the overall quality like that categorical estimation. This testing is simple and the participants interest in this testing by checking the rate of the quality of speech according the various quality scales. Categorical estimation knows the quality of speech with short time.

As a summary it can be said that the system provides satisfactory results after this initial testing

but extensive and continued work is required to develop the system further and to get a high level TTS-system by using coding method such as Linear Predictive Coding and Sinusoidal Modeling and emotional states like that happy, sad, angry and other human feelings.

References

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

[2] Beskow J. (1996). Talking Heads - Communication, Articulation and animation. Proceedings of Fonetik-96: 53-56.

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

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

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

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

[7] Lukaszewicz K., Karjalainen M. (1987). Microphonemic Method of Speech Synthesis. Proceedings of ICASSP87 (3): 1426-1429.

[8] Möbius B., Sproat R., Santen J., Olive J. (1997). The Bell Labs German Text-to-Speech System: An Overview. Proceedings of the European Conference on Speech Communication and Technology vol. 5: 2443-2446.

[9] Murray I., Arnott L. (1993). Toward the Simulation of Emotions in Synthetic Speech: A Review of the Literature on Human Vocal Emotion. Journal of the Acoustical Society of America, JASA vol. 93 (2): 1097-1108.

[10]Maria Moutran Assaf, "A Prototype of an Arabic Diphone Speech Synthesizer in Festival", 2005. 99.

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

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

[13]Tun Thura Thet, Jin-Cheon Na; Wunna Ko Ko, "Word segmentation for the Myanmar language".

[14] Zin Maung Maung and Y. MIKAMI, "A Rule-based Syllable Segmentation of Myanmar Text", in proceedings of IJCLNLP-2008 Workshop on Asian Language Resources, Hyderabad, India, 11-12 January 2008.