A Myanmar Large Vocabulary Continuous Speech Recognition System

Hay Mar Soe Naing, * Aye Mya Hlaing, * Win Pa Pa, * Xinhui Hu,† Ye Kyaw Thu, †
Chiori Hori, ‡ and Hisashi Kawai,†
* Natural Language Processing Lab., University of Computer Studies, Yangon (UCSY), Myanmar
E-mail: {haymarsoenaing, ayemyahlaing, winpapa}@ucsy.edu.mm
† National Institute of Information and Communications Technology (NICT), Kyoto, Japan
E-mail: {xinhui.hu, yekyawthu, hisashi.kawai}@nict.go.jp
‡ Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA
E-mail: chori@merl.com

***Abstract*—This paper presents a large vocabulary automatic speech recognition (ASR) system for Myanmar. To the best of our knowledge, this is the first such system for the Myanmar language. We will report main processes of developing the system, including data collection, pronunciation lexicon construction, effective acoustic features selection, acoustic and language modelings, and evaluation criteria.**

Considering the fact that Myanmar being a tonal language, the tonal features were incorporated to acoustic modeling and their effectiveness were verified. Differences between the word-based language model (LM) and syllable-based LM were investigated; the word-based LM was found superior to the syllable-based model. To disambiguate the definitions of Myanmar words and achieve high reliability on the recognition results, we explored the characteristics of the Myanmar language, and proposed the Syllable Error Rate (SER) as a suitable evaluation criterion for Myanmar ASR system.

3 kinds of acoustic models; 1 Gaussian Mixture Model (GMM) and 2 Deep Neural Networks (DNNs) were explored by only utilizing the developed phonemically-balanced corpus consisting of 4K sentences and 40 hours of speech. An open evaluation set containing 100 utterances, spoken by 25 speakers, were experimented. With respect to the sequence discriminative training DNN, the results reached up to 15.63% in word error rate (WER) or 10.87% in SER.

I. Introduction

With the widespread of mobile phones, especially smartphones, demands for practical ASR applications have rapidly increased due to their convenience and user-friendliness for information access. Meanwhile, in recent years, ASR techniques have took a great leap forward with the help of DNN-based approaches. In the Siri system, for example, the ASR applications are used for searching information on the Web, and as many as 16 languages are currently supported [1]. In the VoiceTra4U system, as many as 12 languages are currently supported for speech-to-speech translations (S2ST) [2] [3]. In these systems, not only do languages with rich data sources such as English, Chinese, Japanese provide satisfied services with high recognition accuracies, but languages with low resources such as Thai and Vienamese, reveal great potentialities for ASR applications. With such systems, the efficiency of accessing information has greatly increased and communication between those who speak different languages has become convenient than ever before.

To the best of our knowledge after detailed investigations, there is still no such ASR system for the Myanmar language, although one prior study was found [4]. This study addresses a Hidden Markov Model (HMM) based speech recognition system, but its main focus is to verify the effective acoustic features among linear predicative coding (LPC), and gamma tone cepstral coefficient (GTCC). Moreover, the amount of data used in this study is very small, with only 558 utterances for training, and only 10 utterances of 1 female speaker for evaluations.

The primary goal of our study is to develop a robust speech recognizer for the Myanmar language, and to incorporate this system into the VoiceTra4U family for industrial S2ST translation applications. At present, the state-of-the-art ASR systems are generally developed with corpus-based approaches in which a large-scale speech data for AMs and textual data for LM are necessary. In most cases, the more data corresponding to the recognition task, the better the ASR system will be. However, creating such data is usually time-consuming and costly. For the Myanmar language, besides insufficient data, there is no prior system to be referred, therefore, to construct a system in a short period of time with limited budget would be a challenge.

This paper is organized as follows. After a brief introduction about the Myanmar language in Section II, the whole system configuration will be presented in Section III. The data construction method will be introduced in Section IV. The modeling approaches for LM and AM will be described in Section V. The ASR experiments will be reported in Section VI. Finally, discussions and conclusions on this work will be given in Section VII.

II. Myanmar language and standard dictionary

A. Tone, word and syllable structure

Myanmar (or Burmese) is the official and primary language of Myanmar, it is spoken by two thirds of the population; 32 million use it as their first language and 10 million as their second language. It is a tonal, pitch-register, and syllable-timed language, largely monosyllabic and analytic language; syllables or words with different tones will have different lexical meanings. There are four nominal tones transcribed in written Myanmar; low, high, creaky and checked. Myanmar tonemes are described with a variety of rates or durations. The length of a tone is defined as rate or duration [5]. The tones are indicated with diacritics or special letters in writing. The Myanmar tones are summarized in Table I.

TABLE I
Characteristics of Myanmar Tones

Register		Phonation	Length	Pitch
Low	a	Modal voice	Medium	Low
High	a:	Breathy voice	Long	High
Creaky	a.	Creaky voice	Medium	High
Checked	a?	Final glottal stop	Short	Varies

A word consists of one or more syllables which are composed of an initial component followed by zero or more medials, zero or more vowels with an associated tone. This means all syllables in Myanmar have prosodic features. Words in the Myanmar language can be divided into simple words, compound words and complex or loan words. A simple word is considered as a syllable, a compound word as a concatenation of several simple words, and loan words as transliterations mainly used for foreign words.

The syllable is the base unit in Myanmar. Most syllables have meanings, and can be used as independent words. The syllable structure follows the pattern: C(G)V((V)C), which means the onset consists of a consonant optionally followed by a glide, and the rhyme consists of a monophthong alone, a monophthong with a consonant, or a diphthong with a consonant [6].

There are 33 consonants, 12 vowels, other medial and consonant diphthongs in Myanmar scripts. It should be noted that in the Myanmar language the syllable-to-pronunciation relationship depends on the context and position, for example the character "စ" in "သစ်စက် [saw mill]" is pronounced as /s/ and in "ဆန်စက် [rice mill]" is pronounced as /z/.

B. A Standard dictionary of Myanmar

To obtain a large word coverage and accurate pronunciations, a standard dictionary released by the Myanmar language commission (MLC) [7] is adopted as our baseline dictionary. There are 26.6K unique words in this dictionary and 105 phonemes are defined with tone information taken into account.

III. System configuration

The overall architecture of our Myanmar ASR system is shown in Figure 1. Because this is the first Myanmar ASR system, the amount of available data is insufficient, so we currently limit our ASR system to the travel domain. NICT owns a set of parallel multilingual text corpora - the Basic Travel Expression Corpus (BTEC) available for several languages including Myanmar. These corpora contain a wide coverage of expressions in the travel domain. In this study, we use this Myanmar BTEC data as the resource; for both speech and textual data to be based on.

A textual phonemically-balanced corpus (PBC) is deliberately constructed by selecting sentences from the BTEC data. A speech corpus is created by recording speech of the above texts. The AM is based on this speech corpus, while the LM is based on the whole BTEC texts. In prior to the construction of the LM, word segmentation was performed on the texts. Due to word segmentation errors, manual checks were also conducted on the segmented texts.

A grapheme to phoneme (G2P) converter was developed to obtain pronunciations of new words. The pronunciation lexicon consists of the original MLC lexicon and the pronunciations of new words appeared in the PBC text.

The speech decoding is performed by a weighted finite state transduer (WFST)-based decoder. The decoding graph $HCLG = H \circ C \circ L \circ G$ was built by the finite state transduer (FST) composition of lexicon L, LM G, context-dependent phones C, and HMM definition of AM H.

IV. Data constructions

The construction of an ASR system requires recorded speech to build a pertinent AM. In order to produce such AMs, two aspects must be taken into account when collecting training speech data; richness and balance. This means that the speech data must be rich enough to contain all the phonemes of the language, and the data must be well-balanced so that it preserves the phonetic distribution of the language. These two aspects are particularly important for low-resourced languages since the amount of available data is insufficient. Such data is referred to as PBC.

Traditionally, this PBC set is manually created; requiring a great deal of human effort. In this study, we adopt an automatic approach to select the sentences which are rich in phonemes from a large textual corpus, and construct the speech corpus by recording these selected sentences.

A. BTEC text

As introduced in Section III, the BTEC is a parallel multilingual corpus constructed by NICT. The text is in the domain of guidebooks for overseas travelers. The contents cover travel-related scenarios that are commonly used while traveling. Originally, it was a collection of parallel sentences in Japanese and their English translations written by bilingual travel experts [8]. Later, in collaboration with members of C-STAR (International Consortium for

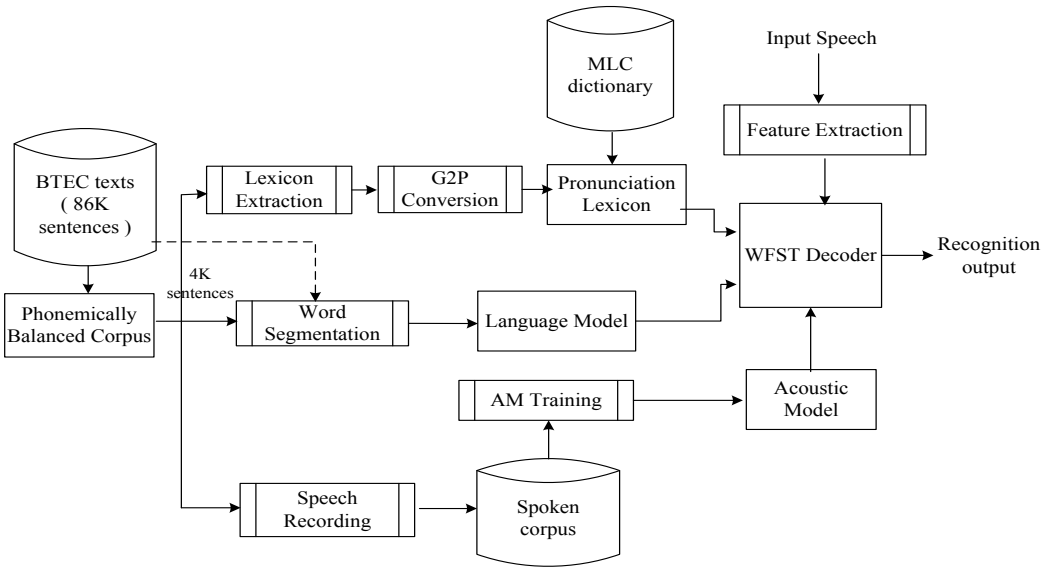

Fig. 1. Configuration of Myanmar ASR system

Speech Translation Advanced Research), this corpus was translated into several other languages including Chinese, Korean, Italian, Spanish, French, German [9][10][11]. Recently, it was also translated into Vietnamese, Indonesian, Thai and Myanmar; these languages are relatively small in terms of data resources [12][13][14].

Since BTEC is elaborately designed by linguistic experts and constructed around the concept of providing a wide variety of samples, situations, and expressions, it has a large coverage of phonetics for each language. Therefore, these corpora serve as the primary sources for developing and evaluating ASR techniques. They are widely used in the current VoiceTra4U S2ST services [2][15].

After removing redundant sentences from the original Myanmar texts, about 90K sentences have remained. In this work, 86K sentences were finally used as the Myanmar BTEC textual resource, from which the Myanmar PBC sentences were selected. Table II shows some of the sentences that appeared in the Myanmar BTEC.

TABLE II
EXAMPLES OF SENTENCES IN THE MYANMAR BTEC

Myanmar	English
နေကောင်းရဲ့လား ။	How are you ?
ဒီဟာ ဘယ်လောက်ကျသလဲ ။	How much is it ?
ကျွန်တော် လက်ဆောင်ပစ္စည်း တချို့ ဝယ်ချင်လို့ပါ ။	I want to buy some souvenirs.
ဘူတာရုံက ဘယ်မှာလဲ ။	Where is the station ?
ကျေးဇူး အများကြီး တင်ပါတယ် ။	Thank you very much.

B. Sentence selection for PBC

To build the PBC, a greedy algorithm [16] is adopted to select the optimal sentences from the BTEC source texts. The selection is based on a metric that maximizes the coverage of multiple units. Here, syllable, diphone, and triphone are regarded as the target units for computing the coverage. The details of this algorithm are described as follows:

Let unit type $X = \{\mu_1^x, \mu_2^x, ..., \mu_{nx}^x\}$, where nx is the number of elements the X contains, and $X \in \{syllable, diphone, triphone\}$. Assume $p(\mu^{x_i})$ to be the occurrence frequency of unit μ^{x_i} in the source text corpus. Let S be a sentence set extracted from the BTEC source data. The coverage of S for X is defined as $C_S^X = \sum_{i=1}^{nx} p(\mu_i^x) \times \sigma_i^x$ where $\delta(\mu_i^x) = 1$ if $\mu_i^x \in S$, Otherwise, $\delta(\mu_i^x) = 0$. By definition, $\sum_1^{nx} p(\mu_1^x) = 1$. Here, C_S^{syb}, C_S^{di}, C_S^{tri} stand for the coverage of sentence set S for syllable, diphone, triphone, respectively.

We intend to maximize C_S^{syb}, C_S^{di}, and C_S^{tri} while designing a sentence set. Firstly, all the sentences having more than 40 syllables are filtered out. Then, a score is calculated for each of the current focused sentences, run through the entire text corpus during the loop process. More specifically, we selected the sentence that has the highest score among the source text corpus, each time after comparing the contributions of all the sentences to the current sentence set S in the following priority order.

① Maximizing set coverage of the syllables C_S^{syb}.
② Maximizing set coverage of the diphones C_S^{di} between two syllable boundaries.
③ Maximizing set coverage of the diphones C_S^{di}.
④ Maximizing set coverage of the triphones C_S^{tri}.
⑤ Select a question.

This algorithm aims to select the best sentence set that simultaneously maximizes the coverage of syllable, diphone spanning two syllables, diphone, and triphone.

With this algorithm, 4K sentences are finally selected as the PBC text data for ASR. According to our evaluations, these sentences respectively cover 99.8% of syllables, 90.9% of diphones and 88.8% of triphones. The selected sentences have resulted in having a wide coverage on foreign words or non-Myanmar words in each unit type. Among the PBC sentences, 25% of the sentences were interrogatives. This owes to the last item of the above list that it raises the priority of the general interrogatives. Such selection is effective when dealing with interrogative speeches that frequently appear in travel conversations.

C. Speech recording

Speech recordings were conducted for all sentences of the above PBC texts in two places. One was done in a sound-proof room at NICT. Here, 1 male and 3 female native speakers attended the recording; each speaker was asked to record the whole 4K sentences. Another was conducted in Myanmar, using the iPhone as recording devices in an open environment such as hotel lobbies and office rooms. 52 male and 48 female speakers attended the recording, each speaker was asked to record 100 sentences of the 4K PBC texts. Finally, a spoken PBC, a total of 25,861 utterances, about 40 hours, was constructed.

D. Building Pronunciation lexicon

In our system, the MLC dictionary is used as the basis for pronunciation mapping. In order to deal with the problem of out-of-vocabulary (OOV) words, for example; loan words for foreign proper nouns, we enhanced the dictionary by developing a G2P converter. To build the G2P converter, processes based on the following two phases were executed; extending the grapheme to phoneme mapping table, and building a model for a pointwise predictor.

1) Extension of the Grapheme to phoneme mapping: Table III shows the Myanmar consonant scripts grouped by their pronunciation types; un-aspirated, aspirated, voiced and nasal. This is also the mapping table used by the MLC dictionary. There are 23 phonemes for 33 consonant scripts, some scripts share the same pronunciations, for example "ဘ", "ဓ", "ဍ" and "ဗ". Besides these mappings for consonants, there are also 87 combinations for vowels defined in the MLC mapping table.

TABLE III
GROUPS OF MYANMAR CONSONANTS

Grouped consonants				
Unaspirated	Aspirated	Voiced		Nasal
က /k/	ခ /kh/	ဂ /g/	ဃ /g/	င /ng/
စ /s/	ဆ /hs/	ဇ /z/	ဈ /z/	ဉ/ည /nj/
ဋ /t/	ဌ /ht/	ဍ /d/	ဎ /d/	ဏ /n/
တ /t/	ထ /ht/	ဒ /d/	ဓ /d/	န /n/
ပ /p/	ဖ /hp/	ဗ /b/	ဘ /b/	မ /m/
ယ /j/	ရ /j/l/r/	လ /l/	ဝ /w/	သ /th/
	ဟ /h/	ဠ /l/	အ /a/	

TABLE IV
EXAMPLES OF PHONETIC MAPPING

Myanmar	MLC	Proposed mapping
ဘ	b	b
မ	m	m
ကျ	kj	ky
ချ	ch	ch
ဂျ	gj	gy
လှ	hla.	lha.
မှ	hma.	mha.
(စ်)	not defined	S
(ရှ်)	not defined	SH
(ချ်)	not defined	CH

In the MLC mapping, however, there are some phoneme sequences in which their occurrence orders are not the same as their real pronunciation orders. For example, the Myanmar syllable "မှ" is written as the order {မ,ှ}, and its phoneme mapping is represented as "hma." However, such writing is inconsistent with the order of how it is uttered; "mha." Therefore, we rewrote these kinds of phonemes according to their pronunciation orders so that they are easily understood and also easily processed by computers.

At the moment, considerations on mapping foreign pronunciations are still insufficient, there are many foreign pronunciations that cannot be found in the Myanmar language. To enhance the representations for these cases, 21 new symbols are newly defined. For example, the foreign name "George" is represented as "ဇော့်(ချ်)" which maps to "gyo.CH", here "CH" is for "(ချ်)". In total, 144 phonemes including consonants, vowels and new special symbols are defined in the proposed mapping table. Some examples of differences between the mapping of MLC and our proposed method are shown in Table IV.

2) Training model for the pointwise predictor: Due to its high performance and efficiency in the tagging tasks, the KyTea toolkit [17][18] is adopted for performing G2P conversion. Here, we consider finding pronunciations for a word as a tagging process. To utilize this toolkit, a model was trained by using a phoneme tagged corpus.

The phoneme-tagged corpus is composed of two parts; The first one is built by the MLC dictionary which contains pronunciations of 26,588 unique words. The second one is built from 3,000 sentences of manually tagged BTEC texts. Concrete steps to build these data are further shown in the following steps;

① Words from the MLC dictionary are broken into syllables using heuristic approach.
② MLC phonemes are mapped to the proposed phoneme set by using a manually built conversion table.
③ Alignment of syllables extracted from the BTEC text to their phonemes are firstly performed by using dictionary based model, and then corrected by manual annotations. The additional manual annotation is necessary because we found that only 80% of the words in the MLC dictionary can obtain unambiguous

alignments to their phonemes by using rules [19]. The alignment of the remainder needs to be completed by manual work. For example, ကာ/ka. ကာ/ga- are both single syllable words. The word ကာစား/ga-za: can be aligned as ကာ/ga- စား/za:. The alignment of "စား" is unambiguously given the alignment of "ကာ".

After the model for the pointwise predictor is trained using the aligned syllable-phoneme data, the pronunciations of all new words appeared in the PBC text are obtained by using the KyTea toolkit. Finally, the pronunciation lexicon for the ASR is obtained by combining them with the existing MLC lexicon. The vocabulary of the upgraded dictionary contains 33,576 unique words (2,578 unique syllables, 101 unique phonemes).

V. Modeling approaches

A. Acoustic model

In recent years, neural networks have again become inherent parts of the state-of-the-art ASR techniques. The DNNs act as AMs for HMM speech recognition system using the hybrid HMM approach whereas the Gaussian Mixture Model (GMM)-based AMs were conventionally used. It has been clear that the DNN-based AMs outperform the GMMs in different ASR tasks [20]. In this system, we focus on DNNs for acoustic modeling, and also explore GMM approaches. To obtain good AMs, we particularly investigated the optimal phoneme set and the effects of utilizing tonal features.

1) Optimal phoneme set: As discussed in the previous sections, to select PBC sentences, we have developed a pronunciation lexicon which contains 144 phonemes (denoted as $PHset_{144}$). However, in the process of acoustic modeling, long vowel phonemes such as /wei:/ have turned out to be not well-modeled. This maybe due to data sparseness because of the limited amount of training corpus. From our preliminary experiments, we found that shorter units of vowel phonemes are more helpful in acoustic modeling. For this reason, we divided 26 vowel phonemes into small parts, for example, /wei:/ was divided into /w/ and /ei:/. Furthermore we removed 10 phonemes that have the same pronunciations with different phoneme representation, for example /aei:/ and /ei:/ in $PHset_{144}$. 108 phonemes (denoted as $PHset_{108}$) are finally remained in the phoneme set. Table V shows performances of these two different phoneme sets by two AMs which will be further explained in the following sections. These results confirmed our anticipations that the phoneme set becomes optimized when it changes from $PHset_{144}$ to $PHset_{108}$.

TABLE V
WER [%] comparisons between two phoneme sets

AM	$PHset_{144}$	$PHset_{108}$
GMM(SAT)	41.05	36.34
DNN(CE)	31.31	27.75

2) Tonal features: Since Myanmar is a tonal language, the utilization of tone information is expected to be helpful to ASR in the same way as it is for other tonal languages such as Mandarin, Vietnamese, etc. In general, there are two methods to incorporate tone into acoustic modeling. One is to do modeling based on the phonemes in which all possible tone patterns are included. Here, these phonemes are regarded as tonal phonemes (or tonemes). For example, for the phoneme /ka/, if tone is taken into account, there are 3 different phonemes /ka/, /ka:/, and /ka./. There are 3 phonemes to be modeled for the tonal phoneme, whereas only 1 phoneme is modeled for the non-tonal phoneme. Another is to augment the conventional acoustic feature such as Mel Frequency Cepstral Coefficient (MFCC) with fundamental frequency (or pitch) feature because the pitch is regarded as being closely related to tone patterns. As previously shown in Table I, the pitch contour changes with four tones.

Here, the acoustic features extracted from the above tonal phonemes are referred to as **tonal features**. When the pitch feature is further added, it is referred to as **augmented tonal features**. On the other hand, those acoustic features extracted from non-tonal phonemes are referred to as non-tonal features.

The Kaldi pitch tracker [21] is used to extract the pitch feature. In this feature, besides the fundamental frequency (or pitch) value F_0, there are also two additional elements; one is the voicing probability (p_v) of current frame, the other is the F_0 delta of neighboring frames (ΔF_0).

3) Training process: The AM training flow is shown in Figure 2. All models are trained with the standard cepstral mean-variance normalized (CMVN) acoustic feature without energy, and its first and second derivatives. The augmented tonal feature (MFCC(dim=13) + F_0(dim=3)) which is mentioned in Section V-A2 is used as the acoustic feature. First, a GMM of speaker independent triphone (GMM(SI)) is trained by using the Linear Discriminant analysis (LDA) to project features of 9 successive frames to 40 dimensions, then by further applying Maximum Likelihood Linear Transform (MLLT). Second, the GMM with speaker adaptive training (GMM(SAT)) is obtained using feature-space Maximum Likelihood Linear Regression (fMLLR) transformations on the features estimated by the GMM(SI).

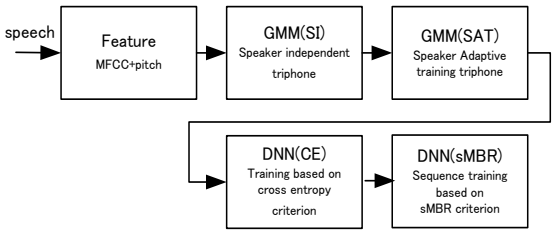

Fig. 2. Acoustic model training flow

Two kinds of DNN models are explored in this system. One is the DNN trained using the cross entropy criterion (DNN(CE)), the other is the DNN based on sequence discriminative training using state-level minimum Bayers risk criterion (DNN(sMBR)). DNN trainings are directly on top of the GMM(SAT), using 11 frames (5 frames on each side of current frame) of context windows of the fMLLR features. For parameters of the DNNs, 5 hidden layers, and 378 units in each layer are used. The output unit number of the respective DNNs is 2,556.

Kaldi speech recognition toolkit is used for the trainings of the above AMs and decodings of speech [22]. All the DNN trainings are performed on a GPU machine.

B. Language model

Language model is an important component in the modern ASR system. Until now, the n-gram LM has been the dominant technology used in the ASR systems since it is pragmatic and efficient. In general, to get an accurate estimation for a n-gram LM, a large textual corpus is required.

1) Data for LM - training, development and test sets: In this study, we use the BTEC texts as the n-gram LM training data. The whole BTEC containing 86K sentences is divided into 2 parts; one is the text data of the PBC containing 4K sentences (here it is denoted as T_{pbc}), and the other contains the remained sentences (denoted as T_{nopbc}).

The K-fold cross validation method was used to validate the texts of the PBC. Figure 3 shows the validation results. In this experiment, the 4K sentences of the PBC texts were randomly divided into 40 groups, with each group containing 100 sentences. The validation is on the perplexity of evaluation group with respect to the LM trained by the other 39 groups. From this graph, the perplexity demonstrates in a relative stable level throughout the whole validations. The 39th group is chosen as test set (T_{test}), and the 26th group is chosen as development set (T_{dev}) used for tuning the LM. The other 38 groups, together with T_{nopbs}, are used for training the LM.

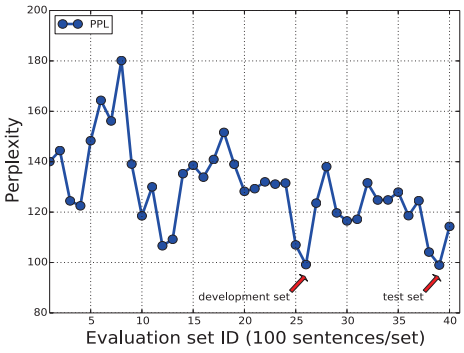

Fig. 3. K-fold cross validations on the PBC texts.

2) Word-based LM and syllable-based LM: Compared with other alphabetic writing systems such as English, Myanmar script is a syllabic system; syllable is the smallest linguistic unit. The Myanmar syllables can be defined in orthographic domain and the syllable-based model has been verified as having high regularities, and the syllable has been regarded as an effective unit for many applications of Myanmar natural language processing [6][23].

Based on these facts, in this study, we will compare the performances of two LMs, one is the syllable-based LM ($LM_{syllable}$), the other is the word-based LM (LM_{word}) which is extensively utilized in the ASRs of other languages.

The LM_{word} is trained using the training textual data which is segmented by word units, while the $LM_{syllable}$ is trained by the same textual data which is segmented by syllable units. The word segmentation is realized by using a word segmentor developed based on a pointwise approach [18] for which an existing 27K annotated BTEC is used. Compared with the word segmentation, the syllable segmentation is relatively simple, and is realized with a high accuracy (over 99.0%) by using simple rules [24].

3) LM training process: 3-gram models are built for both LM_{word} and $LM_{syllable}$, using the MITLM toolkit [25]. In building the LMs, the modified Kneser-Ney smoothing is used. For the evaluations of the resultant LMs, the perplexities of the T_{test} with respect to LM_{word} and $LM_{syllable}$ are 47.22 and 15.59, respectively.

VI. Experiments

We conducted ASR experiments using the data and modeling approaches described above. The main purposes of these experiments are to verify the effectiveness of acoustic features, particularly tonal phonemes, pitch feature in different AMs, and to compare the differences between the LM_{word} and the $LM_{syllable}$. The fMLLR transforms at test time are performed using the alignment model of the SAT model which is based on the speaker independent triphone GMM(SI).

A. Data set

Table VI summarizes the whole data sets used for the experiments.

TABLE VI
Data used for evaluations.

	Type	Size	Speakers (male/female)
AM	Train	24,512 utters. (39.5 hours)	104 (53/51)
	Dev	100 utters. (0.25 hours)	26 (13/13)
	Eval	100 utters. (0.25 hours)	25 (12/13)
LM	Train	86K sents.	-
	Dev	100 sents.	-

B. Evaluation criteria

In general, the WER is used as the evaluation criterion for ASR systems. However, in Myanmar language, the

definition of a word is not strictly defined in some cases. For example, a prefix or suffix can be regarded as a part of a word, or as an individual word. Such situations can cause words to be inconsistent if they are annotated by an unclear specification, or the segmented texts are from different data sources. Therefore, only using WER as the evaluation criterion sometimes does not reflect the real performance of the ASR system. On the other hand, the definition of a syllable is explicit in the Myanmar language. Moreover, as mentioned in the previous sections, the syllable is the base unit of the Myanmar language, almost every syllable has a meaning [6]. Thus, using syllables as the ASR output also provides meaningful and understandable results. This is very similar to Chinese characters which is the fundamental component of Chinese words. Instead of WER, the character error rate (CER) is generally used for Chinese ASR evaluations. Inspired by CER for Chinese, we propose the syllable error rate (SER) as an alternative criterion for Myanmar ASR evaluations.

C. Experimental results

1) Effects of tonal phoneme and pitch feature: To investigate the effects of tonal features including tonal phoneme and pitch on the recognitions, experiments using different feature combinations were conducted. Table VII and VIII respectively show the WERs and SERs of the evaluation set of when the tonal phonemes and pitch features were used/not used in acoustic modelings.

TABLE VII
WERs [%] with/without tonal features

	NT_NP	*NT_P*	*T_NP*	*T_P*
GMM (SAT)	35.60	29.91	31.27	27.60
DNN(CE)	21.09	19.52	19.10	17.10
DNN(sMBR)	20.15	17.94	18.26	15.63

NT - non tonal phoneme, T - tonal phoneme,
NP - no pitch, P - pitch

TABLE VIII
SERs [%] of with/without tonal features

	NT_NP	*NT_P*	*T_NP*	*T_P*
GMM (SAT)	21.48	23.05	22.93	20.79
DNN(CE)	14.32	13.32	13.38	13.19
DNN(sMBR)	13.44	13.19	12.69	10.87

From these results, we can see that both the tonal phoneme and pitch have big influences on the ASR performance. For example, the utilization of the tonal phoneme contributes to 4.33% absolute (or 12.2% relative) reductions in WER for GMM, 1.99% absolute (or 9.4% relative) WER reductions in DNN(CE) model. The influence of pitch feature on the recognition is also encouraging; it reduced the WER in GMM with 5.69% absolute (or 15.98% relative), and in DNN(CE) with 1.57% absolute (7.4% relative). After all, when using both the tonal phoneme and pitch feature, the WER reduction can reach up to 8.0% absolute (or 22.4% relative) for GMM, and 3.99% absolute (or 18.9% relative) for DNN(CE). As shown for other tonal languages such as Mandarin, Vietnamese, Thai, utilizations of tone information have again been verified greatly benefiting the ASRs for the Myanmar language.

2) Comparisons between $LM_{syllable}$ and LM_{word}: Table IX shows the ASR experimental results of using $LM_{syllable}$ and LM_{word}. In these experiments, both the tonal phoneme and pitch features which correspond to the cases of *T_P* of the table VII and VIII are used for acoustic modeling.

TABLE IX
WERs/SERs [%] of syllable-based LM and word-based LM

	$LM_{syllable}$	LM_{word}
GMM (SAT)	41.25/29.08	27.60/20.79
DNN(CE)	27.28/19.35	17.10/13.19
DNN(sMBR)	23.50/14.70	15.63/10.87

In Table IX, we can see that the recognition performance of the $LM_{syllable}$ is inferior to the LM_{word} with big differences of 10.18% and 7.87% in WER with DNN(CE) and DNN(sMBR), respectively.

These results are very similar to Mandarin ASRs where the character-based LM is verified being inferior to the word-based LM in most ASR tasks [26]. It is necessary to notice here that a Chinese character is also syllabic in nature. These facts confirmed the hypothesis that the syllable level constraints of a language are not as restrictive as word sequences [27], while providing additional linguistic informations. Due to such less restrictions, decoding in the syllable level is less effective than in the word level.

In conclusion, for the Myanmar language, the word-based LM can also be regarded as a suitable selection for an ASR system as with other languages, although the word segmentation is inevitable and word ambiguity remains in the word segmentation process.

VII. Discussions and Conclusions

In this paper we introduced the development work of our Myanmar ASR system. The main phases of development; including data collection, pronunciation lexicon construction, effective acoustic feature selection, language modeling and acoustic modeling approaches, are investigated and studied in detail. As an output of this development, we have constructed the following items :

① A PBC text data containing 4K sentences.
② A speech corpus of the above PBC texts, over 40 hours, recorded by 104 speakers.
③ A Myanmar pronunciation lexicon with a vocabulary of 34K words, together with a G2P converter.
④ An annotated and manually checked textual BTEC data, containing 86K sentences, 770K words.

By taking into account the Myanmar language being a tonal language, the effectiveness of tonal phoneme and

pitch have been investigated. They are verified to have high capabilities to improve the recognition performance. The syllable-based LM and word-based LM are also investigated, and the word-based LM is confirmed to be superior to the syllable-based LM, with great differences. Due to the characteristics of the Myanmar language; the ambiguous definition and distinct syllabic constitution of words, and the fact that most syllables having meanings, we proposed the SER as an alternative criterion for evaluating the Myanmar ASR system, so that it can be evaluated with a high reliability and low ambiguity. By evaluating an open test set consisting of 100 utterances from 25 speakers, the best WER of 15.63% (or SER of 10.87%) was obtained by the DNN approach.

Future works include extending the study focus from read speech to spontaneous speech, collecting more speech and textual data in a real environment, improving the quality of word segmentation to further reduce the inconsistencies in the textual data, and exploring new acoustic modeling approaches with full consideration of the special features of the Myanmar language. As a new member of the Universal Speech Translation Advanced Research (U-STAR) consortium, we will implement our ASR system into the VoiceTra4U family. This application will provide more practical data that will be fed back to the system, and further performance improvements can be expected.

Acknowledgment

The authors are grateful to all colleagues in NICT for their suggestions, discussions, and support for Myanmar database collection, speech recording, utilization of computing environment. We also would like to thank colleagues in the University of Computer Studies, Yangon for their support in speech recording.

References

[1] Siri, [Online] Available: en.wikipedia.org/wiki/Siri

[2] S. Matsuda, X. H. Hu. Kashioka, C. Hori, K, Yasuda, H. Okuma, M. Uchiyama, E. Sumita, H. Kawai, and S. Nakamura, "Multilingual speech-to-speech translation system: Voicetra," in *MDM (2), IEEE, 2013,* pp. 229-233, [Online]. Available: http://dblp.uni-trier.de/db/conf/mdm/mdm2013-2.html.

[3] U-STAR: The Universal Speech Translation Advanced Research, [Online] http://ustar-consortium.com/*app_ja*/other.html#languages

[4] I. Khaing, K. Z. Lin, "Design and implementation of speech recognition system for Myanmar," *Proceeding of international conference on computer science & human computer interaction (ICSSHCI 2014),* [Online]. Avaiable: http://www.ijitcs.com/volume%2013_No_1/Ingyin+Khaing.pdf

[5] E. P. P. Soe, "Grapheme-to-Phoneme Conversion for Myanmar Language", 11th International Conference on *Computer Applications (ICCA), Myanmar, 2013*, pp. 195-200

[6] H. H. Htay, K. N. Murthy, "Myanmar Word Segmentation using Syllable Level Longest Matching," In *proc. of the 6th Workshop on Asian Language Resources*, pp.41-48, 2008

[7] "Myanmar-English Dictionary, " Department of the Myanmar Language Commission, Yangon, Ministry of Education, Myanmar, 1993.

[8] G. Kikui, E. Sumita, T. Takezawa, T and Y. Yamamoto, "Creating Corpora for Speech-to-Speech Translation," in *Proceeding of of Eurospeech, 2003*, pp.1:381-384.

[9] T. Shimizu, Y. Ashikari, E. Sumita, J. S. Zhang, and S. Nakamura, "NICT/ATR Chinese-Japanese-English Speech-to-Speech Translation System," in *Tsinghua Science and Technology, 2008*, pp.13(4):540-544 .

[10] X. H. Hu, R. Isotani, H. Kawai, and S. Nakamura, "Construction and Evaluations of an Annotated Chinese Conversational Corpus in Travel Domain for Language Model of Speech Recognition", in *Proceeding of Interspeech, 2010*, pp.1910-1913.

[11] M. Paul, H. Nakaiwa, M. Federico, "Towards Innovative Evaluation Methodologies for Speech Translation," in *Proceeding of NTCIR-4, 2004*.

[12] T. T. Vu, K. T. Nguyen, L. T. Ha, M. C. Luong, S. Nakamura, "Toward Asian Speech Translation: The Development of Speech and Text Corpora for Vietnamese Language," in *Proc. of TCAST 2009*.

[13] S. Sakti, T. T. Vu, A. Finch, M. Paul, R. Maia, S. Sakai, T. Hayashi, S. Matsuda, N. Kimura, Y. Ashikari, E. Sumita, S. Nakamura "NICT/ATR Asian Spoken Language Translation System for Multi-Party Travel Conversation," in *Proc. of TCAST 2009*.

[14] C. Wutiwiwatchai, T. Supnithi, K. Kosawat, "Speech-to-Speech Translation Activities in Thailand," in *Proc. of the Workshop on Technologies and Corpora for Asia-Pacific Speech Translation (TCAST) 2008*.

[15] S. Nakamura, K. Markov, H. Nakaiwa, G. Kikui, H. Kawai, T. Jitsuhiro, J. S. Zhang, H. Yamamoto, E. Sumita, and S. Yamamoto, "The ATR Multilingual Speech-to-Speech Translation System," *IEEE Transactions on Audio, Speech, and Language Processing, Vol.14, Mar.2006*, pp365 - 374.

[16] J. F. Ni, T. Hirai, and H. Kawai, "Constructing a Phonetic Rich Speech Corpus While Controlling Time-dependent Voice Quality Variability for English Speech Synthesis," in *Proceeding of ICASSP, pp. I811-814, 2006*.

[17] KyTea toolkit, [Online] http://www.phontron.com/kytea/index-ja.html

[18] G. Neubig,S. Mori, "Word-based Partial Annotation for Efficient Corpus Construction," in *Proc. of the seventh international conference on Language Resources and Evaluation (LREC), Malta. May 2010*.

[19] Y. K. Thu, W. P. Pa, A. Finch, A. M. Hlaing, H. M. S. Naing, E. Sumita and C. Hori "Syllable Pronunciation Features for Myanmar Grapheme to Phoneme Conversion", 13th International Conference on *Computer Applications (ICCA), Myanmar, 2015*

[20] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition," *IEEE Signal Processing Magazine, vol. 29, no. November*, pp. 82–97, 2012.

[21] P. Ghahremani, B. BabaAli, D. Povey, "A Pitch Extraction Algorithm Tuned for Automatic Speech Recognition," in *Proc. ICASSP*, pp.2513-2517, 2004.

[22] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, K. Vesely, "The Kaldi Speech Recognition Toolkit," *Proc. ASRU, 2011*.

[23] T. H. Hlaing, Y. Mikami, "Automatic Syllable Segmentation of Myanmar Texts using Finite State Transducer," *International Journal on ICT in Emerging Regions, Vol. 6, No.2*, pp.1-9, 2013.

[24] Y. K. Thu, A. Finch, Y. Sagisaka, E. Sumita, "A Study of Myanmar Word Segmentation Schemes" , 11th International Conference on *Computer Applications (ICCA), Myanmar, 2013*, pp.167-179.

[25] B. J. Hsu and J. Glass, "Iterative Language Model Estimation: Efficient Data Structure & Algorithms," In *Proc. of Interspeech, 2008*.

[26] J. Luo, L. Lamel, J. L. Gauvain, "Modeling Characters Versus Words for Mandarin Speech Recognition," In *Proc. of ICASSP*, pp.4325-4328, 2009.

[27] X. Y. Liu, J. L. Hieronymus, M. J. F. Gales, and P. C. Woodland, "Syllable Language Models for Mandarin Speech Recognition: Exploiting Character Language Models," *Journal of Acoustical Society of America, Vol133 (1)*, pp.519-528, Jan. 2013.