HMM Based Myanmar Text to Speech System

Ye Kyaw Thu[1], *Win Pa Pa*[2], *Jinfu Ni*[3], *Yoshinori Shiga*[3],
Andrew Finch[1], *Chiori Hori*[3], *Hisashi Kawai*[3], *Eiichiro Sumita*[1]

[1]Multilingual Translation Lab., NICT, Kyoto, Japan
[2]Natural Language Processing Lab., UCSY, Yangon, Myanmar
[3]Spoken Language Communication Lab., NICT, Kyoto, Japan

yekyawthu, jinfu.ni, yoshinori.shiga, andrew.finch, chiori.hori, hisashi.kawai@nict.go.jp
eiichiro.sumita@nict.go.jp, winpapa@ucsy.edu.mm

10.21437/Interspeech.2015-132

Abstract

This paper presents a complete statistical speech synthesizer for Myanmar which includes a syllable segmenter, text normalizer, grapheme-to-phoneme convertor, and an HMM-based speech synthesis engine. We believe this is the first such system for the Myanmar language. We performed a thorough human evaluation of the synthesizer relative to human and re-synthesized baselines. Our results show that our system is able to synthesize speech at a quality comparable with similar state-of-the-art synthesizers for other languages.

Index Terms: Speech Synthesis, Text to Speech (TTS), Hidden Markov Model (HMM), Ossian, Myanmar

1. Introduction

The primary goal of this research was to develop a high-quality, efficient speech synthesizer for the Myanmar language. The demand for speech-to-speech (STS) translation has been increasing in recent years, and the synthesizer developed as a result of this work will be incorporated into the VoiceTra4U industrial STS translation application [1].

The main challenges stemmed from the almost total lack of linguistic resources for Myanmar. During the 9-month development period it was necessary define a set of phonemes and develop a grapheme-phoneme mapping system for the language. A corpus of Myanmar voice data was constructed from recordings of male and female speakers. It was also necessary to develop a system for text normalization.

Only two studies of Myanmar text-to-speech (TTS) have been reported so far[2][3]. There is much prior research on HMM-based speech synthesis, examples include for Japanese [4], Chinese [5], Vietnamese [6], and Korean [7]. HMM-based approaches can generate speech without the necessarity of large database. There is no statistical or HMM-based speech synthesis approach for Myanmar language.

2. Related Work

A TTS system for Myanmar was reported in [2]. Their approach was based on diphone concatenation with Time Domain Pitch Synchronous Overlap-Add (TD-PSOLA). They used a diphone database of over 7500 diphones extracted from 350 sentences of read speech recorded in reading style Myanmar from a female speaker. The main drawback of this approach is the size of diphone database. The experimental results are compared on two different pitchmarks of Hanning windows in terms of naturalness and speed.

The only speech synthesis methods that have been developed so far Myanmar have used diphone concatenation [2] which is of low speech quality and naturalness at the sentence level, or a rule-based approach [3] which used over 800 demi-syllables as the fundamental unit of speech, acquired by dividing syllables according to consonant-vowel boundaries. They conducted intelligibility tests and found that their system had a syllable correctness rate of over 90%.

In the recent decades, a significant improvement in speech synthesis quality has been achieved by using the HMM-based statistical parametric approach [8, 9, 10], which has proven as a promising method for the automatic generation speech from text. Moreover, a project called Simple4all [11] was launched to further develop the approach to construct TTS systems simply from audio and text data. In this paper, we utilize these tools, particularly using Ossian [12] to develop a HMM-based text-to-speech conversion system in Myanmar.

3. Myanmar Language

Myanmar language belongs to the Lolo-Burmese subbranch of the Tibeto-Burmese branch of the Sino-Tibetan language family. It is the official language of Myanmar where it is spoken by 32 million people. Like all Sino-Tibetan languages, Myanmar has a simple syllable structure consisting of an initial consonant followed by a vowel with an associated tone. There are no final consonants. There are 32 consonants but only 23 have distinct sound 23 and they can be unaspirated, aspirated and voiced, e.g. /p - p^h- b/ There are eight vowel phonemes (/a/, /i/, /u/, /e/, /o/, /ɛ/, /ɔ/) in Myanmar, i.e., sounds that distinguish word meaning. Myanmar is a tonal language and it has three tones, named Tone 1, Tone 2, and Tone 3. Myanmar tone is carried by syllable and is featured by both fundamental frequency and duration of syllable. Tones are associated with syllabic nuclei since the tone information is included in grapheme of that syllable. Figure 2 shows examples of Myanmar tones carried by the same phoneme string "ma". When "ma" is pronounced with different tones, it has different meanings: "hard" with Tone 1 (usually written as /ma/, "lift" with Tone 2 (/ma./), and "doctor" with Tone 3 (in word /tha-/ /ma:/ /do/). In words and sentences, the features of these tones are varied according tone context and intonation. Synthesis of appropriate tone features in contexts is important for achieving good naturalness of synthetic speech. In this work, we deal with this issue by carefully defining tonal phonemes and modeling the contextual tonal variations by a data-driven way through HMM training. This is a big difference compared with the existing methods [2] [3].

4. HMM-based Myanmar TTS

The overall structure of the proposed HMM-based Myanmar TTS system as shown in Figure 1. The HMM-based speech

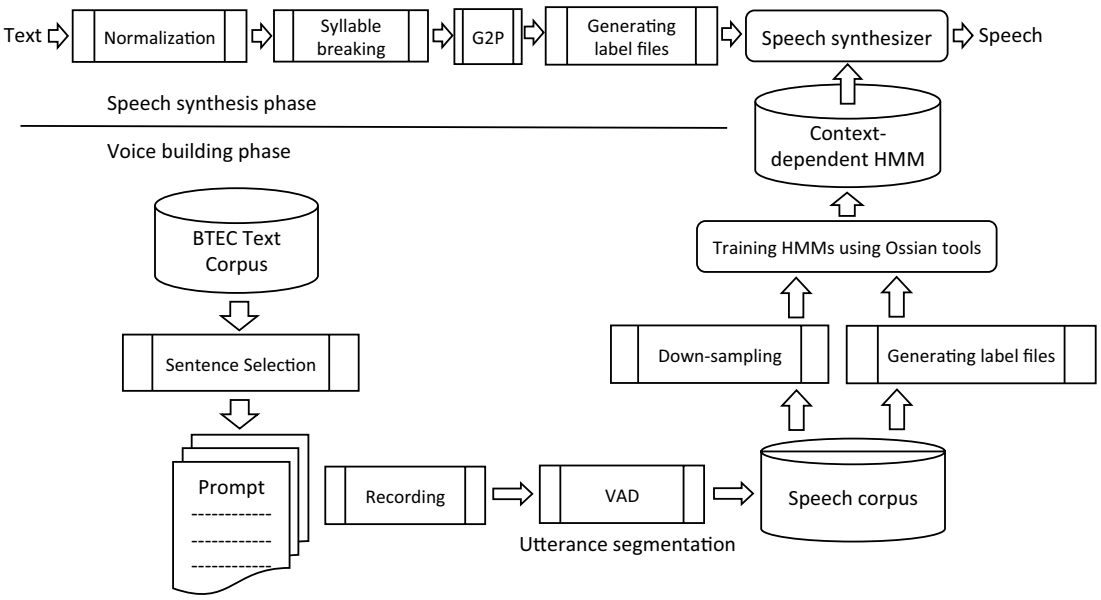

Figure 1: *Overview of HMM-based Myanmar TTS system.*

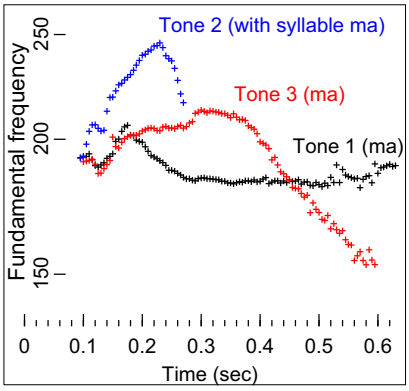

Figure 2: *An example of three tones of Myanmar syllable Ma.*

synthesis technique comprises a voice corpus building phase (Section 5.5) and a speech synthesis phase (Section 5.6). The overall system architecture can be seen in Figure 1. A HMM is a finite state machine which generates a sequence of discrete time observations. In HMM-based speech synthesis, the speech parameters of a speech unit such as the spectrum, fundamental frequency (F0), and phoneme duration are statistically modeled and generated using HMMs based on the maximum likelihood criterion [8, 9, 10]. By using Ossian toolkit [12], only a parallel corpus of text and voice data is required for training HMM-based speech synthesis engines. In this work, we treat Myanmar tones as part of carefully defined phonemes for achieving good quality synthetic speech. In comparison of conventional Ossian-based voice building. We encode each of these phonemes by a separated character in UTF-8. A module named "Generating label files" Figure1 completes this function.

5. Implementation

5.1. Sentence Selection

To collect a high-quality speech corpus for training the HMM-based acoustical models and to achieve high accuracy in Grapheme-to-Phoneme (G2P) conversion by taking into account the contexts of lexical and grammatical structures, we developed a greedy sentence selection tool to select a set of sentences from a large-scale text corpus. The goal was to maximize the coverage of the intended units according the statistics of these units in the text corpus and their contexts.

We adopted the BTEC corpus as the domain for sentence selection since our main purpose was to develop a TTS system for a speech-to-speech translation system for foreign travelers. The size of BTEC1 subset of the Basic Travel Expression Corpus (BTEC) corpus [13] used in these experiments was 160K sentences and manually phoneme-tagging all these sentences would be a time consuming task. Therefore a phonetically balanced sample was taken that contained all syllables by applying the greedy algorithm proposed in [14]. We briefly describe this algorithm below.

To select such a sentence set $\mathcal{S}$ from a large text corpus, it is necessary to define the metric of unit coverage of the sentence set. Let unit type, X, have elements $\{\mu_1^x, \mu_2^x, ..., \mu_{nx}^x\}$, where n_x is the number of elements. X can be a syllable, a diphone, or other defined unit. $p(\mu_i^x)$ is the occurrence probability of μ_i^x in the text corpus, and therefore, $\sum_{i=1}^{n_x} p(\mu_i^x) = 1$. The unit coverage of $\mathcal{S}$ to X, denoted by $C_{\mathcal{S}}^X$, is defined as $C_{\mathcal{S}}^X = \sum_{i=1}^{n_x} p(\mu_i^x) \times \sigma(\mu_i^x)$, where $\sigma(\mu_i^x) = 1$, if $\mu_i^x \in S$. Otherwise, $\sigma(\mu_i^x) = 0$. Given a text corpus and a required sentence set size $|\mathcal{S}| = n$ the goal is to select n sentences from the text corpus to maximize the coverage $C_{\mathcal{S}}^X$. In this paper, three types of units are considered namely: syllables, diphones (sequences of two syllables), and triphones. The coverages of these units were simultaneously maximized in a strict way: sentences were repeatedly selected from the whole text corpus, with syllable units taking the highest priority and the triphones the lowest.

We selected 5,276 sentences from BTEC1 (the text corpus). These covered 99.8% of syllables (excluding the coverages of punctuation counted in the statistics of the whole corpus), 90.9% of diphones and 88.8% of triphones. Furthermore, the selected sentence set contained foreign names and this should allow for the coverage of non-Myanmar words. The sentences were manually phoneme tagged to form a parallel training corpus for a statistical machine translation system for G2P

conversion.

5.2. Text Normalization

Text normalization is a necessary first step for any TTS system. Our approach to Myanmar text normalization is fully rule based and was designed to cover common abbreviations, symbols and numbers in Myanmar language. To handle numbers and other ambiguous Myanmar words we used contextual information for decision making to get the correct reading. For example ၁၂:၁၀ နာရီ (12:10 o'clock) should be pronounced as ဆယ့်နှစ်နာရီ ဆယ်မိနစ် in the context နာရီ (o'clock) this is different from a mathematical expression ဆယ့်နှစ် အချိုး တစ်ဆယ် (twelve isto ten). Some examples of Myanmar language normalization are shown in Table 1:

Table 1: *Example of normalized text.*

Input text	Normalized text
၁၀၀ ဒေါ်လာ (100 dollar)	ဒေါ်လာ တစ်ရာ (dollar 100)
၂/၃ (2/3)	သုံးပိုင်းနှစ်ပိုင်း (lit. 2 of 3)
အထက (abbr of high school)	အခြေခံပညာအထက်တန်းကျောင်း

5.3. Syllable Breaking

In Myanmar text, words composed of single or multiple syllables are usually not separated by white space. Although spaces are used for separating phrases for easier reading, it is not strictly necessary, and these spaces are rarely used in short sentences. Syllable breaking is a necessary step in order to perform Myanmar syllabic grapheme-to-phoneme conversion. Generally, there are only 3 rules required to break Myanmar syllables if the input text is encoded in Unicode (where dependent vowels and other signs are encoded after the consonant to which they apply) [15]. Placing a word break in front of consonants, independent vowels, numbers and symbol characters is the primary rule. The second rule removes any word breaks that are in front of subscript consonants, Kinzi characters, and consonant + Asat characters. The third rule is concerned with break points for special cases such as syllable combinations in loan words and Pali words. This rule-based syllable breaking method is able to perform syllable breaking without error, since the process is unambiguous. Following are examples of Myanmar syllable breaking output:

Table 2: *Example of syllable breaking.*

Input text		Normalized text
မြန်မာနိုင်ငံသား	(citizen of Myanmar)	မြန်_မာ_နိုင်_ငံ_သား
မန္တလေးမြို့	(Mandalay city)	မန်း_တ_လေး_မြို့

5.4. Grapheme-to-Phoneme Conversion

Grapheme-to-Phoneme conversion is a necessary step for speech synthesis and it is the task of predicting the pronunciation of words given only the spelling. A grapheme is the smallest semantically distinguishing unit in a written language analogous to the phonemes of spoken languages. The correspondence between graphemes and phonemes of Myanmar language has ambiguity since the relationship between syllables and their pronunciation is context dependent, and there are many exceptional cases. The Myanmar Language Commission (MLC) [16] Pronunciation Dictionary (28,393 unique words) was used as a basis for G2P mapping and it was extended by adding some new phonemes for foreign pronunciation and modified for consistency of consonant and vowel order during syllable formation. Our approach differed from standard approaches in the method used for G2P conversion. In our approach we carefully defined a set of phonemes that included Myanmar tonal information and used these in the G2P conversion process.

[17] proposed four simple Myanmar syllable pronunciation patterns as features that can be used to augment the models in a Conditional Random Field (CRF) approach to G2P conversion. In [18] higher accuracy was achieved by using the Phrase Based Statistical Machine Translation (PBSMT), and we use this approach for the Myanmar TTS systems reported here. We used the MOSES toolkit for the PBSMT system [19]. The 5,276 Myanmar sentences were selected, syllable segmented, and annotated with phonemes using the procedure described in Section 5.1. The syllables were aligned to phonemes using GIZA++ [20], and phrase extraction used the grow-diag-final-and heuristics [21]. We used the SRILM toolkit [22] to train a 5-gram language model with interpolated modified Kneser-Ney discounting on phoneme training data [23]. In decoding, we adopted the default settings of the MOSES decoder. The SMT-based G2P convertor achieved a phoneme accuracy of approximately 91% [18].

5.5. Building Speech Corpus

This work used a subset of 4,000 sentences from the full corpus of 5,276 selected sentences. The work is ongoing and will eventually make use of the whole corpus. The sentences range in length from 1 to 40 syllables, and were read by one female and male native speaker. The speakers were not professional speakers, but they have knowledge of phonetics and standard Myanmar pronunciation. There are 3.59 hrs of female speech and 3.35 hr of male speech in the speech corpus. The utterances were recorded with a 48kHz sampling rate in a professional recording room using a Marantz Professional PMD661 recording device. The speakers read the text with Myanmar standard pronunciation from annotated phonemes in the text to achieve high quality. Voice activity detection (VAD) was used for the extraction of exact utterances and these clean utterances were used to build the Myanmar speech corpus. In voice building, we downsampled from 48kHz to 16kHz.

5.6. Training and Synthesizing

In this work we employed the Ossian [12] speech synthesis tool from the Simple4All toolkit which is capable of learning from data with little or no expert supervision. It is a collection of Python code for building TTS systems, with an emphasis on easing research into building TTS systems with minimal expert supervision [24]. The flow of the overall speech synthesis process is shown in the upper part of Figure 1.

All the utterances in speech corpus were used as training data.

6. Evaluation Method

To evaluate the performance of HMM-based Myanmar synthesized speech, a subjective evaluation of naturalness and understandability of the synthesized speech, resynthesized speech and natural speech was conducted using the Mean Opinion Score (MOS) (5-very good, 1-very poor) and also a listening comprehension test was performed. 20 human judges were used in both experiments.

We prepared three types of speech stimuli in the evaluation experiment. One is original speech, the second re-synthesized speech using extracted speech parameters from human speech, and HMM-based speech. All the sentences are open and most

Table 3: *Corpus Statistics.*

Data	Syllable		Word		Utterrance	Phoneme	
	total	uniq	total	uniq		total	uniq
Training	33,850	1,845	41,491	7,397	4,000	170,703	133
Test	1,327	488	805	466	65	4,550	92

Table 4: *MOS (Mean Opinion Score) of HMM-Synthesized Speech for Five Different Domains.*

	News		BTEC		Facebook		Blog		Mail	
	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
AVG	2.82	3.07	3.37	3.58	2.75	2.75	3.65	2.48	3.45	3.43
STDEV	1.07	1.23	1.23	1.09	1.21	1.16	1.01	1.16	1.13	1.11

of them are selected from other domains beyond the domain of travel. 65 sentences were randomly selected from 5 domains: the BTEC corpus, newswire text, Facebook posts, blogs and emails, were used in the evaluation. Male and female natural speech from 2 native Myanmar speakers was used as the baseline in the MOS experiment, and as a reference for the human judges in the listening test. Approximately equal proportions of male and female speech were used in both experiments, and the gender of the reference voices in the listening test matched the gender of the (re)synthesized voices.

The resynthesized speech was prepared by first analyzing the original speech using SPTK tools "http://sp-tk.sourceforge.net" with 24-order MGC (mel-generalized cepstrum). The speech was then resynthesized by using MLSADF (a tool of SPTK) with simple excitation.

7. Results and Discussion

Statistics on the training and test corpora are given in Table 3. The results from the MOS evaluation are shown in Figure 3 and Table 4. The error bars on the graph represent the standard deviation of the results. The subjective quality of the male and female voices was roughly equal. The average MOS score for the human voices was 4.7, and the average for the re-synthesized voices was 4.2. The average for the HMM-based synthesizer was 3.3, which is a respectable score relative to the scores from similar experiments on HMM-based synthesizers developed for other languages, for example [25] [26].

The MOS values for the HMM-based synthesizer varied according the domain of test data. The domain with the highest MOS (average of male and female voice scores) of 3.5 was the BTEC3, this was expected given it is similar in character to the training corpus. The lowest MOS of 2.9 was observed on sentences from the news domain, improving the performance on this domain remains a topic for future research. The syllable correctness rate in the listening test was 73.9% ($\sigma = 6.9$) overall; this value includes errors from whole utterances that the listener failed to recognize.

7.1. Error Analysis

Visual inspection of the experimental results clearly showed that most of the errors were on vowel pronunciation; the tone of the vowels was pronounced correctly, however, the choice of vowel proved to be ambiguous. In other words, incorrect vowel pronunciations were chosen, but were pronounced with the correct tone. For example: ka: (pronounced as the 'ca'" English 'car') was recognized as ke: (pronounced as the 'ca' in the English 'care'). multiple times in our experiments. There a few examples of the converse case, where consonants of the same tone were confused, but these were far less frequent. Also,

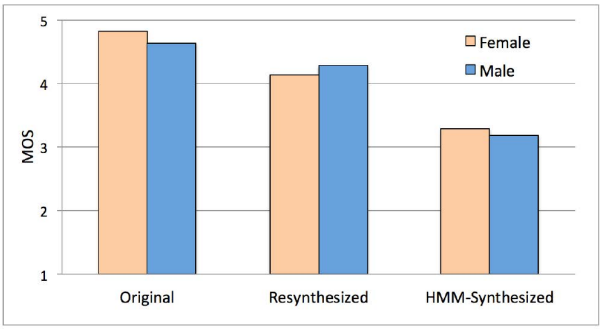

Figure 3: *Result of MOS (Mean Opinion Score) for Three Different Types of Speech.*

there were a few cases where the both the consonants and vowels were recognized correctly but with the wrong tone. Based on the above observations, we believe that the system is generally proficient in synthesizing tones, but there is room for improvement in the more ambiguous cases.

8. Conclusion

This paper describes the development, implementation and evaluation of the first statistical Myanmar TTS system. The system we propose an HMM-based synthesizer that operates at the syllable level, generating speech from sequences of phonemes produced from text by a phrase-based statistical machine translation-based G2P converter. We also proposed a rule-based method for text normalization. We evaluated our TTS system using a human evaluation based on MOS scores and an intelligibility test. Our results show that our system is generating Myanmar speech with a quality comparable to systems developed for other languages that are HMM-based. In future work, intend to increase the size of data we use to train the system and extend the the scope of the system into new domains.

9. Acknowledgements

We thank Ms. Aye Mya Hlaing (UCSY, Yangon, Myanmar) and Ms. Hay Mar Soe Naing (UCSY, Yangon, Myanmar) for their help in phoneme tagging and checking for MLC dictionary and selected 5,276 sentences.

10. References

[1] S. Matsuda, X. Hu, Y. Shiga, H. 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

[2] E. P. P. Soe and A. Thida, "Text-to-speech synthesis for myanmar language," *International Journal of Scientific & Engineering Research*, vol. 4, no. 6, pp. 1509–1518, 2013.

[3] K. Y. Win and T. Takara, "Myanmar text-to-speech system with rule-based tone synthesis," *Acoustical Science and Technology*, vol. 32, no. 5, pp. 174–181, 2011. [Online]. Available: http://ci.nii.ac.jp/naid/130001853285/en/

[4] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis." in *EUROSPEECH*. ISCA, 1999.

[5] Q. Yao, F. Soong, Y. Chen, and M. Chu, "An hmm-based mandarin chinese text-to-speech system," *In Proceeding of ISCSLP 2006*, vol. Springer LNAI Vol., no. 4274, pp. 223–232, 2006.

[6] T. T. Vu, M. C. Luong, and S. Nakamura, "An hmm-based vietnamese speech synthesis system," *2009 Oriental COCOSDA International Conference on Speech Database and Assessments*, vol. Springer LNAI Vol., pp. 116–121, 2009.

[7] S.-J. KIM, J.-J. KIM, and M. HAHN, "Implementation and evaluation of an hmm-based korean speech synthesis system(¡special section¿ statistical modeling for speech processing)," *IEICE transactions on information and systems*, vol. 89, no. 3, pp. 1116–1119, mar 2006. [Online]. Available: http://ci.nii.ac.jp/naid/110004719388/en/

[8] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from hmm using dynamic features," *Proc. ICASSP-95*, pp. 660–663, 1995.

[9] K. T. T. Masuko, T. Kobayashi, and S. Imai, "Speech synthesis using hmms with dynamic features," *Proc. ICASSP-96*, pp. 389–392, 1996.

[10] K. Tokuda, T. Masuko, T. Kobayashi, and S. Imai, "An algorithm for speech parameter generation from hmm using dynamic features." *Acoust. Soc. Japan (J)*, vol. 53, no. 3, pp. 192–200, 1997.

[11] R. A. J. Clark, "Simple4all." in *INTERSPEECH*, F. Bimbot, C. Cerisara, C. Fougeron, G. Gravier, L. Lamel, F. Pellegrino, and P. Perrier, Eds. ISCA, 2013, pp. 2654–2656. [Online]. Available: http://dblp.uni-trier.de/db/conf/interspeech/interspeech2013.html

[12] S. consortium members, "Ossian speech synthesis toolkit," November 2013-2014. [Online]. Available:
http://homepages.inf.ed.ac.uk/owatts/ossian/html/index.html

[13] G. Kikui, E. Sumita, T. Takezawa, and S. Yamamoto, "Creating corpora for speech-to-speech translation," 2003.

[14] J. Ni, T. Hirai, and H. Kawai, "Constructing a phonetic-rich speech corpus while controlling time-dependent voice quality variability for english speech synthesis," *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006*, pp. I–811–I–814, 2006.

[15] Y. K. Thu, A. Finch, Y. Sagisaka, and E. Sumita, "A study of myanmar word segmentation schemes for statistical machine translation," *Proceeding of the 11th International Conference on Computer Applications*, pp. 167–179, 2013.

[16] *Myanmar-English Dictionary*. Department of the Myanmar Language Commission, Ministry of Education., 1993.

[17] 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," *Proceeding of the 13th International Conference on Computer Applications*, pp. 161–167, 2015.

[18] Y. K. Thu, W. P. Pa, A. Finch, J. Ni, E. Sumita, and C. Hori, "The application of phrase based statistical machine translation techniques to myanmar grapheme to phoneme conversion," in *Proceedings of PACLING (to appear)*, 2015.

[19] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ser. ACL '07. Stroudsburg, PA, USA: Association for Computational Linguistics, 2007, pp. 177–180. [Online]. Available: http://dl.acm.org/citation.cfm?id=1557769.1557821

[20] F. J. Och and H. Ney, "Improved statistical alignment models," in *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ser. ACL '00. Stroudsburg, PA, USA: Association for Computational Linguistics, 2000, pp. 440–447. [Online]. Available: http://dx.doi.org/10.3115/1075218.1075274

[21] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation," in *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL)*, Edomonton, Canada, May 27-June 1 2003.

[22] A. Stolcke, "Srilm - an extensible language modeling toolkit," 2002, pp. 901–904.

[23] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," in *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, ser. ACL '96. Stroudsburg, PA, USA: Association for Computational Linguistics, 1996, pp. 310–318. [Online]. Available: http://dx.doi.org/10.3115/981863.981904

[24] O. Watts, A. Stan, R. Clark, Y. Mamiya, M. Giurgiu, J. Yamagishi, and S. King, *Unsupervised and lightly-supervised learning for rapid construction of TTS systems in multiple languages from 'found' data: evaluation and analysis*. ISCA-INST SPEECH COMMUNICATION ASSOC, 2013, pp. 101–106.

[25] J. Latorre, K. Iwano, and S. Furui, "New approach to the polyglot speech generation by means of an hmm-based speaker adaptable synthesizer." *Speech Communication*, vol. 48, no. 10, pp. 1227–1242, 2006.

[26] T. T. T. Nguyen, C. d'Alessandro, A. Rilliard, and D. D. Tran, "Hmm-based tts for hanoi vietnamese: issues in design and evaluation." in *INTERSPEECH*, F. Bimbot, C. Cerisara, C. Fougeron, G. Gravier, L. Lamel, F. Pellegrino, and P. Perrier, Eds. ISCA, 2013, pp. 2311–2315.