ENHANCING MYANMAR TEXT-TO-SPEECH SYSTEM BY USING LINGUISTIC INFORMATION ON LSTM-RNN BASED SPEECH SYNTHESIS MODEL AND TEXT NORMALIZATION

AYE MYA HLAING

UNIVERSITY OF COMPUTER STUDIES, YANGON

JUNE, 2020

Enhancing Myanmar Text-to-Speech System by Using Linguistic Information on LSTM-RNN Based Speech Synthesis Model and Text Normalization

Aye Mya Hlaing

University of Computer Studies, Yangon

A thesis submitted to the University of Computer Studies, Yangon in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

June, 2020

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research and has not been submitted for a higher degree to any other University or Institution.

…..…………………………… …………........…………………………

Date Aye Mya Hlaing

ACKNOWLEDGEMENTS

I would like to thank the Ministry of Education for providing full facilities during the Ph.D. course at the University of Computer Studies, Yangon.

No research is possible without infrastructure and requisite materials. For this, I extend my sincere gratitude to Dr. Mie Mie Thet Thwin, Rector of the University of Computer Studies, Yangon, for giving the opportunity to do this research and providing me with the requisite institutional facilities throughout my research.

My greatest pleasure and the deepest appreciation to my supervisor, Dr. Win Pa Pa, Professor, Natural Language Processing Lab., University of Computer Studies, Yangon, for her kindly support, encouragement, excellent mentorship, invaluable guidance, constructive comments and practical advice throughout the research period.

I greatly appreciate and acknowledge the kindly support, motivation, encouragement and valuable suggestions received from Dr. Khine Khine Oo, Professor and Course-coordinator of the Ph.D 10th Batch, University of Computer Studies, Yangon.

I deeply and specially thank the external examiner, Dr. Aye Thida, Professor, University of Computer Studies, Mandalay, for her patience in critical reading, valuable comments and suggestions in the preparation of the dissertation.

I would like to mention my special appreciation and thanks to Dr. Khin Mar Soe, Professor, Head of Natural Language Processing Lab, University of Computer Studies, Yangon, for her insightful advice and suggestions, and providing me with the institutional resources.

I would like to extend my special thanks to Dr. Ye Kyaw Thu, Language and Speech Science Research Lab., Waseda University, Tokyo, Japan, for his suggestions and comments, especially related to the issues of the text normalization.

I am indeed obliged to Daw Aye Aye Khine, Associate Professor, Head of English Department, for her support in writing my research papers and my dissertation from the language point of view.

My sincere thanks also go to all my respectful teachers for giving us valuable lectures and knowledge during the Ph.D. course work.

I also thank my classmates from the UCSY Ph.D. 10th batch, my colleagues in Natural Language Processing Lab., and colleagues of my husband for their help in subjective evaluations throughout my research work and their true friendship along the way.

Last but not least, I would like to give my deep thanks to my parents for their support, continuous encouragement and endless love over the years. It cannot be expressed enough my unfailing gratitude to my husband who has supported me throughout this research and has constantly encouraged me to complete this thesis.

ABSTRACT

This thesis focuses on enhancing Myanmar Text-to-Speech (TTS) system to generate more natural synthetic speech for a given input text. Typical TTS systems have two main components, text analysis (front-end), and speech waveform generation (back-end). Both front-end and back-end parts are important for the intelligibility and naturalness of the TTS system. Therefore, this thesis is emphasized on both text analysis part and acoustic modelling part in Statistical Parametric Speech Synthesis (SPSS) system.

Text analysis part consists of a number of natural language processing (NLP) steps and text normalization is the first and crucial phase among them. Myanmar text contains many non-standard words (NSWs) with numbers. Therefore, Myanmar number normalization designed for Myanmar TTS system is implemented by using Weighted Finite-State Transducers (WFSTs). For grapheme-to-phoneme (G2P) conversion in text analysis part, the first large Myanmar pronunciation dictionary is built, and the quality of that dictionary is confirmed by applying machine learning techniques such as sequence to sequence modelling. With the purpose of extracting contextual linguistic features which can promote the quality of the synthesized speech of Myanmar TTS system, phoneme features and a large Myanmar pronunciation dictionary with syllable information are prepared on a general speech synthesis architecture, Festival. After that, a proposed Myanmar question set is applied in extracting linguistic features which will be used in neural network based speech synthesis. Finally, word segmentation, WFST based number normalization, G2P conversion, and contextual labels extraction modules are integrated into text analysis part of Myanmar TTS system.

The accuracy of acoustic model in SPSS is very important to achieve good quality synthetic speech. In this work, Hidden Markov Model based Myanmar speech synthesis is conducted with many contextual labels extracted from text analysis part and used as the baseline system. The state-of-the-art modelling techniques such as Deep Neural Network (DNN) and Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) have been applied in acoustic modelling of Myanmar speech synthesis to promote the naturalness of synthesized speech. The effectiveness of contextual linguistic features and tone information are explored in LSTM-RNN based

Myanmar speech synthesis using the proposed Myanmar question set. Furthermore, the effect of applying word embedding and/or Part-of-Speech (POS) features as the additional input features in acoustic modelling of DNN and LSTM-RNN based systems are investigated in this work. The effect of word vector features can be seen clearly in DNN based system in both objective and subjective evaluations. However, in LSTM-RNN based systems, it can be observed that applying word embedding features can only give little improvement in subjective results and it cannot lead to any improvement in objective results. Therefore, it can be concluded that contextual linguistic features extracted from our text analysis part and the proposed question set are good enough for acoustic modelling of LSTM-RNN based Myanmar TTS system to generate the more natural synthesized speech for Myanmar language. According to the objective and subjective results, the hybrid system of DNN and LSTM-RNN (i.e., four feedforward hidden layers followed by two LSTM-RNN layers) is the most suitable network architecture for Myanmar speech synthesis.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

CHAPTER 1
INTRODUCTION

Speech is the most important part for human communication and people want to communicate with machine via natural speech. There have been a great number of efforts in speech processing to build high quality human-computer interaction with speech. Text-to-Speech (TTS) also called Speech Synthesis, one of the main speech processing technologies, is a technique for generating intelligible, natural-sounding artificial speech for a given input text.

TTS system can be applied in a wide variety of application areas such as telephone based conversational agents, voice-over functions for visually impaired person, communication aid for speech impaired person, Speech-to-Speech translation system, automatic question and answering system, e-books readers, and communicative robots. Regardless of what application that TTS is applied, it is really important that the quality of the synthesized voice needs to be high and more natural sounding.

In physical nature, the text is a discrete message form and the speech signal is a continuous acoustic waveform. If the message is to be spoken aloud, there are many issues of how to pronounce and which prosody to use because the written text contains little or no prosodic information [48]. Therefore, extracting linguistic features that can relate the prosody of the speech is the important factor for TTS system.

Typical TTS systems have two main components: text analysis (front-end), and speech waveform generation (back-end). In the text analysis, given input text is transformed into a linguistic specification consisting of phonemes. The text analysis part is language dependent and includes a number of natural language processing (NLP) steps, such as word segmentation, text normalization, part-of-speech (POS) tagging, and grapheme-to-phoneme (G2P) conversion. In the speech waveform generation component, speech waveforms are generated from the linguistic specification produced by text analysis component.

This thesis is dedicated to enhance Myanmar TTS system that can output the more natural synthesized speech. Therefore, the research has emphasized on both front-end and back-end parts to promote the quality of the synthetic speech of Myanmar TTS system. Specifically, text normalization, building a large Myanmar pronunciation dictionary, contextual linguistic features extraction in the front-end part, and acoustic modelling of back-end part have been conducted to get high quality synthesized speech for Myanmar language. In addition to contextual linguistic features generated by front-end part, using word embedding features as the additional input ones is also highlighted in this research.

1.1 Motivation of the Thesis

Little research has been performed for speech synthesis on Myanmar language. One rule-based Myanmar TTS system [66], one diphone-concatenation based speech synthesis [47] and one statistical or HMM-based Myanmar TTS [51] are found publicly. The synthesized speech of diphone synthesis suffers from the sonic glitches of concatenative synthesis and likes robotic sound. The HMM-based Myanmar TTS system operates only at the syllable level and the quality of synthesized speech is still far below the natural speech. Some wrong insertion of pause within a word and erroneous word segmentation are found in synthesized speeches. This leads to the unnatural synthesized speech although the meaning can be understood by native speakers. Therefore, the quality of Myanmar TTS system should be enhanced by taking account the various aspects of promoting the quality of synthesized speech.

The first case is text normalization which needs to be handled Non-Standard Word (NSWs) in Myanmar language and there is no published work of implementing individual Myanmar text normalization system to integrate with Myanmar TTS system.

The second one is building a large Myanmar pronunciation dictionary for Grapheme to Phoneme (G2P) conversion. G2P conversion model is needed to generate accurate pronunciation for TTS. However, no enough amount of pronunciation data is available for Myanmar language to build G2P conversion model by applying machine learning techniques.

The third one is contextual factors used in acoustic modelling. Though many contextual factors such as phone identity factors, locational factors can affect spectrum, F_0 pattern and duration [53], previous works were not taken into account these factors. Therefore, extracting and applying contextual factors for Myanmar language is important for enhancing the quality of Myanmar TTS system. Moreover, none of the previously Myanmar TTS system has been used word embedding features for acoustic modelling. The effect of these features on acoustic modelling of neural network based Myanmar speech synthesis should be investigated.

The last one is acoustic modelling technique used in Myanmar speech synthesis. The decision tree clustered context-dependent HMMs has some limitations and one of the major factors of degrading the quality of synthesized speech is the accuracy of acoustic models [75].

Therefore, Myanmar speech synthesis should be performed by applying state-of-the-art modelling techniques to promote the naturalness of synthesized speech. To the best of our knowledge, this is the first attempt to apply neural network architecture in Myanmar speech synthesis.

1.2 Objectives of the Thesis

The main purpose of this thesis is to enhance Myanmar speech synthesis that can output the more natural synthesized speech. For promoting the accuracy of acoustic model, neural network architectures such as Deep Neural Network (DNN) and Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) have been applied in Myanmar speech synthesis and the most suitable neural network architecture for Myanmar speech synthesis has been investigated. The following are the other objectives:

Myanmar text contains many Non-Standard Words (NSWs) with numbers. To promote the quality of Myanmar TTS system, normalization of these NSWs needs to be performed in the initial step of the system. One of the objectives of this research is to implement Myanmar number normalization devised for Myanmar TTS system.

Constructing a large Myanmar pronunciation dictionary for G2P conversion in TTS front end is also involved in this research. Enough amount of pronunciation data is the most important factor for applying machine learning techniques on G2P conversion for TTS system.

Contextual linguistic information is an important feature for naturalness in speech synthesis and using that feature in various speech synthesis models improves the quality of the synthesized speeches for different languages. The other objective is to explore the effect of contextual linguistic features including tone information for the naturalness of Myanmar TTS system.

A question set is used for context clustering of HMM-based speech synthesis and extracting linguistic features of neural network based speech synthesis and it is also language dependent requirement. There is no publicly available question set for

Myanmar language yet. Therefore, proposing a question set for Myanmar language is one of the objectives of this research.

Another objective is building word vectors for Myanmar language which can give wide coverage and good performance, and applying these vectors as the input features to the acoustic modelling of Myanmar speech synthesis to explore their effectiveness.

1.3 Contributions of the Thesis

There are six main contributions in this thesis.

The very first contribution of this thesis is identifying semiotic classes for Myanmar language by the study of Myanmar text corpus and implementing Weighted Finite State Transducers (WFST) based Myanmar number normalization. This is the first published work of Myanmar number normalization system designed for Myanmar TTS system.

The second contribution is constructing a large Myanmar pronunciation dictionary and using that dictionary for G2P conversion in the text analysis part of Myanmar TTS system. Syllable information of each entry is also included in the dictionary.

The third contribution is exploring contextual linguistic features for Myanmar language and applying these features in modelling Myanmar speech synthesis. For extracting these features from text, phoneme features and pronunciation lexicon with syllable information for Myanmar language have been prepared in Festival, the general speech synthesis architecture.

Though question sets for other languages such as English, Japanese and Mandarin can be available publicly, question set for Myanmar language is not available. Therefore, the fourth contribution of this research is proposing a question set for Myanmar language and used this question set in linguistic features extraction for modelling neural network based Myanmar TTS system.

Word vectors can be obtained by unsupervised learning from large amount of raw text data and can be used as the input features to acoustic modelling of speech synthesis. Building monolingual corpus for the purpose of building word vectors for Myanmar language and applying these word vector features as the additional input features to contextual linguistic features in acoustic modelling of DNN and LSTM-RNN based Myanmar speech synthesis is the fifth contribution.

Recently, DNN-based acoustic modelling approach showed significant improvements over HMM-based approach. The sixth contribution is applying neural network based architecture in Myanmar speech synthesis for promoting the accuracy of acoustic model. Since Myanmar is a tonal language, the tone type of the syllable can influence prosodic features such as fundamental frequency and duration of that syllable. The effectiveness of tonal features in contextual information is also investigated on neutral network based Myanmar speech synthesis. The effect of different input features including word embedding features and part-of-speech (POS) features in acoustic modelling of neural network based Myanmar speech synthesis is examined in this research. The suitable network architecture for Myanmar speech synthesis is explored by doing experiments on DNN, LSTM-RNN, and a hybrid system of DNN and LSTM-RNN based acoustic models.

1.4 Organization of the Thesis

This dissertation is organized with nine chapters including introduction of TTS system, motivation, objectives, and contribution of this research.

The literature reviews on Text-to-Speech (TTS) techniques, related work of this research, previous researches of TTS system on Myanmar language, and evaluation metrics of TTS synthesis are described in Chapter 2. Detailed implementation of Myanmar number normalization system designed for Myanmar TTS system and experimental results of the system are presented in Chapter 3. The detailed process of building a large Myanmar pronunciation dictionary which has been applied in grapheme-to-phoneme conversion is illustrated in Chapter 4. Linguistic feature extraction, a proposed question set, and modelling word vectors for Myanmar language are described in Chapter 5. The phonetics of Myanmar language are also introduced. Chapter 6 shows the general overview of the Hidden Markov Models (HMM) based speech synthesis and the implementation of the baseline Myanmar speech synthesis using HMM. Moreover, Myanmar Speech Synthesis with CLUSTERGEN and the subjective results are also reported. The detailed implementation and experimental results of DNN based Myanmar speech synthesis with different input features is reported in Chapter 7. The proposed LSTM-RNN based acoustic modelling for

Myanmar speech synthesis is illustrated in Chapter 8 and experiments on different architecture of LSTM-RNN are conducted. The effect of contextual linguistic features, tone information, and word embedding features are also examined and analyzed in this chapter. Chapter 9 concludes the research work and mentions the advantages and limitation of the system, and the further directions to improve the naturalness of the TTS system.

CHAPTER 2
LITERTATURE REVIEW AND RELATED WORK

This chapter describes the literature review on Text-to-Speech (TTS) techniques, related work of this research and previous researches of TTS on Myanmar language. Evaluation metrics of TTS synthesis are also briefly presented in this chapter.

2.1 Text-to-Speech Synthesis

TTS synthesis is a technique for converting the natural input text into the intelligible, natural-sounding speech waveform.

Typically TTS system performs this conversion in two steps, first converting the input text into a phonemic internal representation and then converting this representation into a speech waveform [24]. The first step is text analysis and the second is speech waveform synthesis. Both of them are important for achieving more natural synthetic speech.

The text analysis part is language dependent requirement and includes many natural language processing (NLP) steps, such as sentence segmentation, word segmentation, text normalization, part-of-speech (POS) tagging, grapheme-to-phoneme (G2P) conversion, linguistic features extraction, and prosodic analysis.

There are different paradigms for waveform synthesis such as formant synthesis, articulatory synthesis, concatenative synthesis and statistical parametric speech synthesis. These are briefly described in the following sections.

2.1.1 Formant Synthesis

In formant synthesis, speech waveform generation component used very low-dimensional artificial spectra, including especially formants. Formant synthesizers attempt to mimic human speech by generating artificial spectrograms. The most well-known of the formant synthesizers were the Klatt formant synthesizer [29]. The evaluation of formant synthesis showed that the generated sound is intelligible, often "clean" sounding though it is far from natural sounding. [48].

2.1.2 Articulatory Synthesis

In articulatory synthesis, it attempts to synthesize speech by modeling the physics of the vocal tract and articulatory process [14]. The talking machine of von Kempelem is a famous articulatory synthesizer. The simple process like filtering or mechanical damping that is modelled the fact of moving the articulators with a certain inherent speed, controls the motion of the tubes. The problem is that the correct articulatory parameters cannot be found from recordings rather more-intrusive measures (e.g., MRI or EMA imaging, x-ray photography) were used [48]. In the early days, collecting this sort of data is more difficult.

2.1.3 Concatenative Synthesis

There are two kinds of concatenative synthesis: diphone synthesis and unit selection synthesis.

2.1.3.1 Diphone-concatenative Synthesis

In the 1980s, a small database of phoneme units called "diphones" were used as the acoustical units for waveform generation. These units were concatenated conforming to the given sequence of phonemes by using pitch-synchronous overlap-add (PSOLA) approach either in the frequency domain or time domain [36]. This is called as diphone synthesis.

2.1.3.2 Unit-selection Synthesis

With the increasing power of computer technology and the increase in amount of speech and linguistics resources, approaches for speech waveform synthesis has advanced from knowledge-based and rule-based ones to data-driven ones.

In unit-selection synthesis, synthesized speech can be produced by concatenating the waveforms of units selected from large, single speaker speech databases. A large number of units with various characteristics of prosody and spectrum are collected in the large databases and appropriate sub-word units are automatically selected from that databases to produce high-quality natural sounding synthesized speech with relevant prosody [23]. This approach is called "unit selection synthesis".

Unit-selection techniques have evolved to become the dominant approach to speech synthesis and commercial systems have developed using these techniques [4, 8, 10, 11].

In this technique, the quality of output derives directly from the quality of recordings. It seems that the larger the speech database, the better the coverage. However, no modification can be done on synthetic speech, this limits the generated speech to the same style as the original recordings. If speech variations can be controlled, the large databases with different styles of speech samples are needed. Recording large database with different styles is very time consuming and cost inefficient. Moreover, a bad join in a single utterance or inappropriate units in the utterance can degrade the quality of synthetic speech.

2.1.4 Statistical Parametric Speech Synthesis (SPSS)

For getting more control over speech variations, in the late 1990s, statistical parametric speech synthesis (SPSS) emerged and became popular in last decade [6, 53, 72, 78, 79]. SPSS might be simply described as generating the average of some sets of similarly sounding speech segments [77]. Figure 2.1 shows the classic three-stage pipeline of SPSS. Among these modules, statistical models have been built by applying different machine-learning algorithms. In waveform generator of SPSS, vocoders are used to extract acoustic features, such as spectral and excitation features, from the natural speech of training data and to generate synthetic speech from output acoustic features at synthesis time.

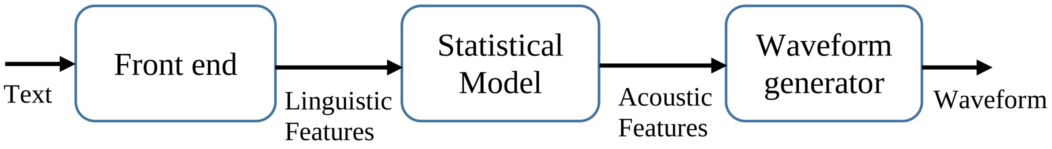


Figure 2.1 Pipeline of Statistical Parametric Speech Synthesis

2.1.4.1 Hidden Markov Model based SPSS

Hidden Markov model (HMM) based speech synthesis which uses HMM as its generative model was popular because of its flexibility in changing speaker identities, emotions, and speaking styles [57]. Many techniques for controlling speech variations have been applied in HMM-based speech synthesis and they can improve the quality of synthesized speech.

In [53], the authors applied an HMM-based speech synthesis system (HTS) to English speech synthesis using the general speech synthesis architecture of Festival. Many contextual factors related to phoneme, syllable, word, phrase and utterance are taken into account for English. The authors showed that the model size is already small enough for small devices such as PDAs without specific efforts to compress the file size and confirmed that the prosody of synthetic sound is fairly natural.

In [20], the authors built speech synthesis system for Bahasa Indonesia language with CLUSTERGEN method in FestVox. STRAIGHT and moving segment label with the new version of FestVox were added to the trial to improve the naturalization of the synthesized speech. 150 samples sound was used for DMOS test and number of respondents was 10. The result was average 3.74 score of Degradation Mean Opinion Score (DMOS).

2.1.4.2 Deep Neural Network based SPSS

Some limitations of decision tree clustered context-dependent HMMs are highlighted in [75]. One of the major factors of degrading the quality of synthesized speech is the accuracy of acoustic models. Therefore, HMM-based statistical models are replaced by DNNs.

Zen et al. proposed DNN to model the relationship between input features and their acoustic realizations [75]. The input features for the DNN-based systems included 342 binary features for categorical linguistic contexts and 25 numerical features for numerical linguistic contexts. In the objective evaluation, the authors explored the relationship between the performance and the architecture of the DNN; number of layers and units per layer. According to a subjective preference listening test, they confirmed that the preference of DNN-based systems is higher than HMM-based ones.

Although applying neural networks have been applied in speech synthesis since the 1990s [25], the advance both in hardware such as GPU, and in more efficient machine learning algorithms enables training DNN from a large amount of training data with more layers and computational resources. In recent years, artificial neural network-based acoustic models have become the state-of-the-art acoustic modeling in speech synthesis area.

Qian et al. examined DNN based TTS with a moderate size speech corpus of 5 hours. They showed that DNN can yield better synthesized speech than HMM [40]. The Root Mean Square Error of F_0 trajectories generated by DNN is improved by 2 Hz than that of HMM-based baseline.

Multi-task learning and stacked bottleneck features have employed into DNN-based TTS synthesis for multi-speaker modelling in [68]. In [31], the authors combined vector-space representations of linguistic context and DNN, which can directly accept such continuous representations as input. Different input features such as binary or continuous and different time scales such as frame or state are experimented on DNN-based TTS.

2.1.4.3 Long Short-Term Memory Recurrent Neural Network based SPSS

One limitation of the feed-forward DNN-based acoustic modeling is that the sequential nature of speech is ignored [76]. Although there is certainly a nature of correlation between continuous frames in speech, each frame is sampled separately by the DNN-based approach. Recurrent Neural Networks (RNNs) were applied for modeling sequential data that embodies correlations between consecutive frames in speech. In [59] and [26], the standard RNNs have been applied in speech synthesis area.

However, the standard RNNs has the problem that the influence of a given input on the hidden layer either decays or blows up exponentially around the network's recurrent connections [17]. To overcome this vanishing gradient problem, the most effective solution so far is Long Short-Term Memory (LSTM) architecture [22]. LSTM is the most widely used RNN in speech processing because LSTM is capable of learning long time-dependencies [18]. Recent studies demonstrated that LSTMs can achieve significantly better performance on SPSS than DNN.

In [13], RNNs with bidirectional Long Short-Term Memory (BLSTM) were adopted to capture the correlation information between any two frames in a speech utterance. The input feature vector contained 319 binary features for categorical linguistic contexts and 36 numerical linguistic contexts. Experimental results showed that a hybrid system of DNN and BLSTM-RNN can outperform either HMM, or DNN TTS system, both objectively and subjectively. The speech trajectory generated by the BLSTM-RNN TTS is fairly smooth and no dynamic constraints are needed.

The unidirectional LSTM RNNs with a recurrent output layer was proposed to apply acoustic modeling for SPSS to achieve low-latency speech synthesis in [76]. The authors used 291 linguistic contexts; only current and future two contexts at phoneme, syllable, word and phrase levels for the LSTM-RNNs as they can access the past information through their recurrent connections. The authors showed that LSRM-RNN based model could synthesize natural sounding speech and the speech parameter generation can be eliminated from the pipeline of speech synthesis. MOS scores showed that LSTM-RNNs offered more efficient acoustic modeling than feed-forward DNNs.

In [67], several variants of LSTM were examined and the forget gate and memory cell state of the LSTM were analyzed. The authors proposed a simplified architecture that can achieve similar performance in both objective and subjective evaluations though it has fewer parameters than the vanilla LSTM.

In [65], the authors used word embedding in Bidirectional LSTM-RNN (BLSTM-RNN) based TTS system. Four different kinds of published word embedding for English were tested and they proved that word embedding can improve the performance of the baseline BLSTM-RNN based TTS system without Part-of-Speech (POS) and Tone and Break Indices (ToBI) input features. However, it still has a gap to the upper bound system, which uses manually labeled POS and TOBI as input features.

In [63], the embedded vectors of various linguistic units such as phonemes, syllables, and phrases were used as the additional or alternative features in neural network based acoustic modelling. Preprocessing of the embedding vectors are done by applying a simple scaling method. The results indicated that using that features only lead to insignificant improvement of RNN based acoustic model. The objective results shows that although the word and phrase vectors encode useful information for F_0 modelling in DNN, it may be less useful for RNN systems.

2.2 Evaluation for Text-to-Speech

The quality and naturalness of synthesized speeches generated by speech synthesis are evaluated in terms of objective and subjective measures.

2.2.1 Objective Evaluation

Objective results are used to measure the quality of synthesized speech in terms of distortions between the synthesized speech and natural speech of the original speaker. The objective measures used in this research are Mel-Ceptral Distortion (MCD), F_0 distortion in root mean square error (RMSE) and voiced/unvoiced error (V/U) in percentage.

Mel Cepstral Distortion (MCD) is a measure of how different two sequences of mel cepstra are. If v^{syn} and v^{ref} are synthesized and reference waveforms, then MCD can be calculated with this Eq. (2.1).

$$MCD(v^{syn}, v^{ref}) = \frac{\propto}{T'} \sum_{\substack{t=0 \\ ph(t) \notin SIL}}^{T-1} \sqrt{\sum_{d=s}^{D} \left(v_d^{syn}(t) - v_d^{ref}(t) \right)^2} \tag{2.1}$$

where, v_d (t) are 60-dimensional mel frequency-scaled cepstral coefficients with a frame step size of 5 ms, d is the dimension index ranging from 0..59, t is time(frame index), T' is the number of non-silence frames, and $T = \min\left(|v^{syn}|, |v^{ref}|\right)$ frames in length. The expression $ph(t) \notin SIL$ means that frames inside silence region are excluded.

Root mean square error (RMSE) of F_0 distortion is calculated as:

$$RMSE = \sqrt{\frac{\sum_{t \in V}(f_{syn}(t) - f_{ref}(t))^2}{\#V}} \tag{2.2}$$

where, $f_{ref}(t)$ is the extracted F_0 observation of natural speech at time t, $f_{syn}(t)$ is the synthesized F_0 value at time t, t denotes the time indices when both natural speech and synthesized speech are voiced and #V is the total number of voiced frames.

If the $l_{syn}(t)$ and $l_{ref}(t)$ are voicing labels of synthesized speech and natural speech, then Voiced/unvoiced error (V/UV) or voicing classification error(VCE) can be calculated with the following equation:

$$V/UV = 100 \frac{\sum_{t=1,T}(1 - \delta(l_{syn}(t), l_{ref}(t)))}{T} \tag{2.3}$$

where, $\delta(l_{syn}(t), l_{ref}(t))$ is 1 if l_{syn} = l_{ref} and 0 otherwise and T is the total number of frames.

In all these metrics, the lower is the better. MCD measures the accuracy of generated spectral parameters and RMSE measures F_0 contours while U/UV measures the voicing accuracy in percentage.

2.2.2 Subjective Evaluation

Subjective evaluation is usually done to evaluate the quality and naturalness of synthetic speech of the TTS system. The following four subjective evaluation methods have been used in this research.

In Comparison Mean Opinion Score (CMOS) test, subjects are presented with a pair of synthesized speech samples generated by two systems. The judgment is made on 5-point CMOS scroes by comparing the synthesized speeches of two systems and the scores are shown in Table 2.1.

Table 2.1 CMOS Scores

Quality of the speech	CMOS Score
Much better	5
Better	4
About the same	3
Worse	2
Much worse	1

In Mean Opinion Score (MOS) test, subjects have to rate the naturalness of synthesized speeches on a scale from 1 to 5 and MOS scores are presented in Table 2.2.

Table 2.2 MOS Scores

Quality of the speech	MOS Score
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

In AB preference test, subjects are given synthesized speeches pairs and have to choose the more natural one in each pair or “neutral” if the difference between two speech samples cannot be perceived.

In MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor) listening test, the subjects are instructed to listen the speech samples generated by many systems and rate them using 0-100 scale on their naturalness. The rating scale are shown in Table 2.3.

Table 2.3 MUSHRA Scores

Quality of the speech	MUSHRA Score
Excellent	81-100
Good	61-80
Fair	41-60
Poor	21-40
Bad	0-20

2.3 Previous Text-to-Speech Researches on Myanmar Language

A few number of Text-to-Speech research on Myanmar language has been found publicly. These researches are done by applying rule-based, concatenative and Hidden Markov Model based approaches.

A Myanmar Text-to-Speech system with rule based tone synthesis was introduced in [66] and the authors implemented tone rules of a linear pattern based on two parameters, F_0 and duration. The speech unit used in the system was demisyllable. A syllable intelligibility test was carried out for 248 monosyllable words and average intelligibility score was 92.56%. There are issues on the improvement of speech naturalness and some modifications for continuous speech.

In [47], a Myanmar TTS system based on Diphone Concatenation method was proposed. Time Domain Pitch Synchronous Overlap and Add (TD-PSOLA) algorithm was applied to smooth the joints of the speech signals. The authors showed the testing results with the varieties of overlapping pitch marks for speech waveforms of Myanmar sentence. The comparisons of naturalness and intelligibility was done on 20 pairs of words of confusability.

A HMM-based Myanmar TTS system that operates at the syllable level was found in [51]. The authors used Myanmar tones included in defined phoneme symbols to achieve good quality synthesized speech. Twenty human judges were used to evaluate the quality of output speech in experiments. They used 65 sentences of five domains for testing and the average Mean Opinion Score (MOS) for the synthesized speech was 3.3.

2.4 Summary

In this chapter, the various techniques of TTS synthesis are reviewed and the state-of-the-art acoustic modelling techniques in SPSS such as DNN based and LSTM-RNN based techniques are highlighted. By reviewing the related work, it can be seen that these techniques can promote the quality of synthesized speech for other languages such as English, Japanese. According to the review of previous TTS researches on Myanmar language, there is no research on TTS for Myanmar language using neural network based architectures. Moreover, both objective and subjective evaluation methods of TTS are described in this chapter.

CHAPTER 3

MYANMAR NUMBER NORMALIZATION FOR TEXT-TO-SPEECH

This chapter presents Myanmar number normalization devised for Myanmar Text-to-Speech system and presents defined semiotic classes for Myanmar language. The detailed processes of implementing Weighted Finite State Transducer (WFST) based Myanmar number normalization system and experimental results of the system are presented in this chapter.

3.1 Myanmar Number Normalization

Text normalization is the first essential module in text analysis part of TTS system. In order to produce a phonemic internal representation of raw input text, it needs to be normalized or preprocessed in many ways. Input text is split into sentences, and deal with idiosyncrasies of abbreviations, numbers and so on. The first task in text normalization is sentence tokenization. There is no ambiguous case in sentence tokenization for Myanmar language because Myanmar sentences can be split by using sentence marker "။" which is normally used at the end of the sentence. The second step in text normalization is normalizing non-standard words (NSWs). NSWs are tokens like including numbers, abbreviations, dates, currency amounts and acronyms which need to be expanded into sequences of standard words before they can be pronounced. For example, "၂၀၁၉၊ ဇန်နဝါရီ" needs to be pronounced "နှစ် ထောင့် ဆယ့် ကိုး၊ ဇန်နဝါရီ", not "နှစ် သုည တစ် ကိုး၊ ဇန်နဝါရီ" and "၁.၁.၂၀၁၉" needs to be pronounced "တစ် ရက် တစ် လ နှစ် ထောင့် ဆယ့် ကိုး" instead of pronouncing "တစ် တစ် နှစ် သုည တစ် ကိုး". How to pronounce NSW is difficult because they are often very ambiguous. As an example, the number "၁၂၃" can be spoken in two different ways, depending on the context: တစ် ရာ့ နှစ် ဆယ့် သုံး (in "၁၂၃ သိန်း") and တစ် နှစ် သုံး (in "အိမ်နံပါတ် ၁၂၃"). Myanmar abbreviations such as "အ.မ.က", "အ.ဆ.ည" can be generally pronounced by pronouncing letter by letter. However, abbreviations of English units like "km", "°F" need to be pronounced as "ကီလိုမီတာ", "ဒီဂရီဖာရင်ဟိုက်", respectively.

In classifying text tokens (e.g. the time “နံနက် ၄:၃၀”), there is only one way to classify which is hour=“၄” minute=“၃၀”. However, in verbalization of that token, there are more than one way in which they can be rendered as words, e.g. “လေးနာရီ သုံးဆယ် မိနစ်” or “လေးနာရီခွဲ” and different users can have different preferences. With the semiotic model, verbalization component can be modified subjectively for other specific application. Therefore, semiotic classes for Myanmar language have been identified by analyzing the Myanmar text corpus, and classification and verbalization are accomplished depends on the semiotic classes.

Myanmar text has various NSWs including numbers. To promote the quality of Myanmar TTS system, these NSWs are firstly normalized into their standard words in the initial step of text analysis. In this research, normalization of NSWs with Myanmar numbers are more emphasized. Any parallel or annotated normalized corpus is not available publicly for Myanmar language and statistical approaches cannot be applied on Myanmar number normalization. The usefulness of weighted finite-state transducers (WFSTs) are found in the Kestrel text normalization system, a component of the Google TTS system [12]. Therefore, Myanmar number normalization is implemented by writing number normalization grammars that are compiled into libraries of WFST.

3.2 Semiotic Classes for Myanmar Language

For classifying NSWs, the semiotic classes for Myanmar language have to be identified. Therefore, a set of semiotic classes for Myanmar language are identified by investigating the Myanmar sentences from Asian Language Treebank (ALT) parallel corpus[1] [43] which comprises 20,000 sentences in the news domain and some Web data (3,150) sentences. The defined semiotic classes are described in Table 3.1 and bold font style indicates NSW with Myanmar number.

3.3 Weighted Finite State Transducer for Myanmar Number Normalization

Myanmar number normalization is attained by accomplishing two phases: classification and verbalization phases. In classification phase, tokens are identified by

[1] http://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/index.html

the defined semiotic classes, and in verbalization phase, tokens are expanded into standard words depending on their semiotic classes. Both classification and verbalization phases are accomplished by defining language-specific grammars that are compiled to WFST. A WFST consists of a set of states and transitions between states. Each transition is labeled with an input symbol from an input alphabet, an output symbol from an output alphabet, an origin state, a destination state, and a weight [44]. Finite State Transducer (FST) can represent certain sets and binary relations over the string. In this work, OpenGrm Thrax Grammar Compiler[2] [44] was used for the development of Myanmar number normalization system. The Thrax grammar compiler compiles grammars that consist of regular expressions, and context-dependent rewrite rules, into the archives of WFST. Grammars for classification and verbalization are presented in the following sections.

Table 3.1 Semiotic Classes for Myanmar Language

Semiotic Class	**Description**	**Examples**
DATE	date	၁၂-၂-၂၀၁၉, ၂၀၁၉ ဖေဖော်ဝါရီ ၁၂, etc.
TIME	time	၉:၁၀:၂၅, ၉ နာရီ ၁၀ မိနစ်, ဂျီအမ်တီ ၀၉၁၀, etc.
CURRENCY	currency amount	၃၄၅,၅၀၀ ကျပ်, $၁,၀၀၀, etc.
NUMBER	cardinal, decimal	၄၅ ယောက်, ၂၁၅.၇၅, ၅၀%, etc.
DIGIT	digit by digit	+၉၅-၉ ၁၁၁၂၂၃၃၃, အမှတ် ၆၉၅, etc.
RANGE	range	၃၅ - ၄၀ ဒီဂရီဖာရင်ဟိုက်, ဒေါ်လာ ၂၀၀ နှင့် ၃၀၀ ကြား, etc.
SCORE	score	၂-၃ ဂိုး, ၂:၃ ဂိုး, etc.
DIMENSION	dimension	ပေ ၄၀ × ၆၀, etc.
NRC	national identification number	၅/မရန(နိုင်) ၁၂၃၄၅၆, etc.

3.3.1 Grammars for Classification

For classifying the semiotic classes of NSWs, many kinds of grammars are compiled to the libraries of WFST and some preprocessing rules are applied.

Cardinal, decimal, and measure are marked as the member of NUMBER class. In Myanmar language, ordinal numbers are not necessary to be identified as one type

[2] http://www.openfst.org/twiki/bin/view/GRM/Thrax

of semiotic class because cardinal number or text are usually used in describing ordinal number. Examples are “၃ ကြိမ်” and “တတိယအကြိမ်” has the same meaning of ordinal number “3rd” in English. In the case that the clue that points the whole digit sequence as the NUMBER semiotic class is detected before the first digit (e.g. in “မီတာ ၁၀၀၊ ၂၀၀၊ ၃၀၀” sequence, the clue is “မီတာ”), spaces between these digit sequences are removed as the preprocessing step and the whole sequence can be identified as the NUMBER class. Unit symbols commonly used for measurement in Myanmar language (e.g. %, cm, °F, kg, km), 96 number prefixes (e.g. အသက်, နှစ်ပေါင်း, စုစုပေါင်း, အနည်းဆုံး, စာမျက်နှာ), and 275 number suffixes (e.g. သင်း, လုံး, ခု, ချောင်း, လက်မ) are used for classification of NUMBER semiotic class. These are extracted by collecting manually from Myanmar text corpus.

For DATE semiotic class, many rules are defined and implemented to cover many styles of writing date in Myanmar text. Table 3.2 shows some writing styles of Myanmar date.

In classification for CURRENCY semiotic class, two types of clues, currency symbol and currency text are included in writing grammars. As an example, “$ ၁၀၀”, “ဒေါ်လာ ၁၀၀” has the same meaning in English as "$100".

WFSTs are compiled from the grammars written for all defined semiotic classes and are used in the Classification phase. Different priorities of classification are accomplished by assigning different weights. Example input and output strings of classification phase are as follows:

- Input String: ရန်ကုန်တိုင်းဒေသကြီး အစိုးရအဖွဲ့မှ ကြီးမှူးကျင်းပသည့် ၂၀၂၀ ခုနှစ် နှစ်ဦး တရားအလှူတော် ဒုတိယနေ့ ဓမ္မသဘင်အခမ်းအနားကို ရန်ကုန်မြို့ရှိ ပြည်သူ့ရင်ပြင်တွင် ဇန်နဝါရီ ၂ ရက် ည ၇ နာရီခွဲ တွင် ကျင်းပခဲ့သည်။

  (The second day of the Dhama Ceremony at the beginning of Year 2020, organized by the Yangon Region Government, was held at the People's Square in Yangon on January 2 at 7:30 pm.)

Output String: ရန်ကုန်တိုင်းဒေသကြီးအစိုးရအဖွဲ့မှ ကြီးမှူးကျင်းပသည့် <DATE>year: ၂၀၂၀ ခုနှစ်</DATE> နှစ်ဦး တရားအလှူတော် ဒုတိယနေ့ ဓမ္မသဘင်အခမ်းအနားကို ရန်ကုန်မြို့ရှိ ပြည်သူ့ရင်ပြင်တွင် <DATE> month: ဇန်နဝါရီ day: ၂ရက်</DATE> ည <TIME>၇နာရီခွဲ</TIME> တွင် ကျင်းပခဲ့သည် ။

Table 3.2 Myanmar Dates

Myanmar	**English**
ဇန်နဝါရီ **၄၊ ၂၀၁၉**	January 4 2019
၂၀၁၉ ၊ ဇန်နဝါရီလ **၄**	2019 January 4
၂၀၁၉ ခုနှစ် ၊ ဇန်နဝါရီလ **၄**	2019 January 4
၂၀၁၉ ၊ **၄** လပိုင်း	4/2019
ဇန်နဝါရီလ **၄**	January 4
၄.၁.၂၀၁၉	4.1.2019
၄-၁-၂၀၁၉	4-1-2019
၄/၁/၂၀၁၉	4/1/2019
၁၃၈၀ ခုနှစ် ဝါဆို လဆန်း **၁၀** ရက်	10, waxing Warso, 1380
၁၃၈၀ ခုနှစ်၊ တန်ဆောင်မုန်း လပြည့်ကျော် **၁၀** ရက်	10, waning Tazaungmone, 1380
၂၀၁၈-၂၀၁၉ ခုနှစ်	2018-2019
၂၀၁၈-၂၀၁၉ ပညာသင်နှစ်	2018-2019 academic year

3.3.2 Grammars for Verbalization

Grammars for verbalization of tokens depending on the defined semiotic classes are also accomplished by using WFST. As an example, a score "၃-၂ ဂိုး" (3:2) is expanded into "သုံး ဂိုး နှစ် ဂိုး". In this case, "-" symbol is replaced by suffix "ဂိုး". Some symbols are neglected in verbalization. For example, in DIMENSION semiotic class,

“ပေ ၆၀ × ၈၀” (60 ft × 80 ft) can be expanded into “ပေ ခြောက် ဆယ် ရှစ် ဆယ်” and “×” symbol is neglected in verbalization.

In the pronunciation of Myanmar digit sequence like “တစ်ဆယ့်” (ten) and “တစ်ထောင့်” (one thousand), “တစ်” (one) is usually omitted. For example, the common pronunciation of “၁၁:၀၀ နာရီ” is “ဆယ့် တစ် နာရီ” and “၁၉၄၈” is “ထောင့် ကိုး ရာ့ လေး ဆယ့် ရှစ်”. Therefore, rules for fixing these cases are added in verbalization phase.

Example input and output strings of verbalization phase are as follows:

- Input String: မြန်မာနိုင်ငံသည် ၁၉၄၈ ခုနှစ် ဇန်နဝါရီလ (၄)ရက်နေ့ တွင် စစ်မှန်သော လွတ်လပ်ရေးကို ရရှိခဲ့သည်။

  (Burma gained genuine independence on 4 January 1948.)
- Output String: မြန်မာနိုင်ငံသည် ထောင့် ကိုး ရာ့ လေး ဆယ့် ရှစ် ခုနှစ် ဇန်နဝါရီလ လေး ရက်နေ့ တွင် စစ်မှန်သော လွတ်လပ်ရေးကို ရရှိခဲ့သည် ။

3.3.3 Myanmar Number Names Expansion

The expansion of Myanmar number names is applied in all grammars except digit by digit expansion. The number name grammars (rules) depend on the factorization of digit string into sum of products of powers of ten. The factorization is done according to the nature of the language. For example, most Western languages have no terms for 10^4, the factorization becomes $1×10^1 ×10^3$ and its verbalized text is “ten thousand”. In Myanmar language, a term “သောင်း” (ten thousand) is usually used for 10^4 and the factorization is $1×10^4$. Although there are terms “သန်း” (million) for 10^6 and “ကုဋေ” (ten million) for 10^7 in Myanmar language, it is not commonly used in pronunciation of CURRENCY class. “၁၀ သိန်း” (ten lakh) is commonly pronounced for 10^6 and “သိန်း ၁၀၀” (hundred lakh) for 10^7. Therefore, in factorization, $1×10^1×10^5$ is defined for the first case and $1×10^2×10^5$ for the second case. The factorization is converted into appropriate Myanmar standard words by the Myanmar number names grammar. As an example, ၁,၂၃၄,၅၆၇ (1,234,567) is expanded into standard words as “ဆယ့် နှစ် သိန်း သုံး သောင်း လေး ထောင့် ငါး ရာ့ ခြောက် ဆယ့် ခုနစ်”.

3.3.4 Finalization

For the case in which token cannot be classified its semiotic class and verbalized into standard words, post-processing are done. The digit sequence is defined as cardinal number names if there is a comma or dot in digit sequence or it has one non-zero digit followed by the zero digits. If not, it will be scanned how many digits the sequence has. If it has three or more digits, it will be pronounced as digit by digit and if it has less than three digits, pronounced as the cardinal number. By applying these rules, there is no missing digit sequence to be verbalized in the Myanmar number normalization system.

3.4 Experiments

Experiments are done to evaluate the performance of classification and verbalization of WFST-based Myanmar number normalization system.

3.4.1 Test Data Preparation

Two test sets which needs to be normalized are prepared for evaluating the WFST-based Myanmar number normalization and shown in Table 3.3. For these two test sets, parallel tagged corpus is prepared for evaluating the accuracy of classification and normalized corpus for evaluating the overall performance.

Table 3.3 Test Sets for Number Normalization

Test Set	No. of Sentences	Source
TestData-1	1,000	ALT
TestData-2	947	Web

3.4.2 Experimental Results of Classification

Eq. (3.1) is used for evaluating the performance of classification.

$$tag\ accuracy\ (\%) = \frac{number\ of\ particular\ tags\ in\ test\ data}{number\ of\ particular\ tags\ in\ reference\ data} \tag{3.1}$$

where, the numerator is the number of output tags from the system and the denominator is the number of expected tags.

Table 3.4 shows the number of particular tags in test data and reference data and the percentage of tag accuracy for each test set. It achieves the overall tag accuracy **94.3%** on TestData-1 and **92.6%** on TestData-2.

Table 3.4 Classification Accuracy of Two Test Sets

Semiotic Class	**TestData-1**	**TestData-2**
NUMBER	755/810 (93.2%)	725/789 (91.9%)
DATE	396/404 (98%)	422/442 (95.5%)
TIME	78/80 (97.5%)	68/68 (100%)
CURRENCY	71/82 (86.6%)	98/118 (83.1%)
DIGIT	5/5 (100%)	3/3 (100%)
RANGE	2/2 (100%)	6/8 (75%)
SCORE	5/8 (62.5%)	N/A
DIMENSION	N/A	2/2 (100%)
Overall Accuracy	1312/1391 (94.3%)	1324/1430 (92.6%)

Classification errors are found in three cases. The first case is that there is no clue in the context such as “၅၀ က ပြန် စ သွားတာလား။” (Does it restart from 50?). In this case, the semiotic class for Myanmar number cannot be classified.

The second case is caused by some common prefixes or suffixes on both CURRENCY and NUMBER classes such as “သန်း”, “သိန်း”, “သောင်း”. As an example, in this sentence “လူဦးရေ သန်း ၅၀ ကျော်ရှိပါတယ်။” (There are over 50 million people.), “သန်း ၅၀” is misclassified as CURRENCY because “သန်း” (million) is common prefix in both NUMBER and CURRENCY classes.

The third case is that the clue in the context is not existed in the collected lists of prefix or suffix units. As an example, in this sentence “Wi-Fi ဖြာထွက်မှု ၂၄၁၂ မှ

၂၄၇၂ MHz ထိ ရှိနေပါသည်။” (The spread of Wi-Fi is ranged from 2412 to 2472 MHz.), “၂၄၁၂ မှ ၂၄၇၂ MHz” cannot be classified as RANGE class because the unit “MHz” is not included in the list.

3.4.3 Experimental Results of Verbalization

Simple rule-based number normalization system implemented by using regular expression (RE) in Perl programming language was used as a baseline system. Word error rate (WER) was used for comparing the results of two systems, a baseline system and WFST-based system. As shown in Table 3.5, WFST-based number normalization system achieves WER 0.5% for TestData-1 and 1.4% for TestData-2, and which is 5.0% and 6.3% better than the baseline system, respectively.

Table 3.5 Overall Performance

	TestData-1 (WER%)	**TestData-2 (WER%)**
Baseline system	5.5%	7.7%
WFST-based system	0.5%	1.4%

Though there are some tag errors in classification phase of WFST-based number normalization system, verbalization results of some misclassification have also correct pronunciations because of almost same pronunciation for CURRENCY and NUMBER class in Myanmar language and post-processing of the system. Therefore, the overall accuracy of verbalization is high in our experiments.

3.5 Summary

In this chapter, WFST-based number normalization system was implemented as the separate system. According to the experimental results, it can be concluded that this WFST-based approach can get acceptable results for Myanmar number normalization and can be used practically. This system has been integrated into the text analysis part of Myanmar TTS system to generate the satisfactory pronunciation for Myanmar numbers.

CHAPTER 4
PRONUNCIATION DICTIONARY FOR MYANMAR LANGUAGE

This chapter describes Grapheme to Phoneme (G2P) conversion which is one of the important modules in text analysis part of Myanmar TTS system. The detailed description of building a large Myanmar pronunciation dictionary is presented in this chapter.

4.1 Phonetic Analysis

Phonetic analysis is the stage to take the normalized word strings from text normalization module and produce a pronunciation for each word. The most crucial component in the phonetic analysis is a large pronunciation dictionary [24]. However, dictionary alone is not insufficient to pronounce all words in real text because real text contains words that do not appear in the dictionary. For pronouncing the unknown words, many G2P conversion methods have been applied.

4.1.1 Dictionary Lookup

For English TTS systems, the freely available CMU Pronouncing Dictionary[1] is one of the most widely-used lexicons. It contains over 134,000 entries and its pronunciations use a 39-phone ARPAbet-derived phoneme set. Another popular dictionary for English speech synthesis is 110,000 word UNISYN dictionary and it gives syllabifications, stress, and some morphological boundaries.

However, any large pronunciation dictionary for Myanmar language is not found in the web. Therefore, the first large amount of pronunciation dictionary for Myanmar language has been built for applying in Myanmar TTS system. The detailed process of building a large Myanmar pronunciation dictionary will be described in Section 4.2.

4.1.2 Grapheme-to-Phoneme Conversion

Grapheme-to-Phoneme (G2P) conversion is the process of converting a sequence of letters (graphemes) into a sequence of phones (phonemes). For example in

[1] http://svn.code.sf.net/p/cmusphinx/code/trunk/cmudict/cmudict-0.7b

Myanmar language, given a Myanmar word “မင်္ဂလာ” , the task is to generate its pronunciation “min ga- la”. It has a greatly critical role for TTS and Automatic Speech Recognition (ASR) applications.

Many G2P conversion methods have been applied for pronouncing unknown words. Letter-to-sound (LTS) rules are written as the earliest algorithms for G2P conversion. The decision tree is used for computing the phoneme string which has the highest conditional probability given the grapheme sequence.

Recently, machine learning methods such as neural sequence to sequence models have been applied in G2P conversion. The applicability of sequence-to-sequence models were explored to address the problem of G2P conversion [71]. Bi-directional Long Short-Term Memory (LSTM) neural networks using alignment information were applied to G2P problem and they achieved the state-of-the-art results. Unidirectional LSTM (ULSTM) with various kinds of output delays and deep bidirectional LSTM (DBLSTM) with a connectionist temporal classification (CTC) layer were experimented on the CMU dataset for US English [41] and the best result on G2P conversion was achieved by combining DBLSTM-CTC and 5-gram Finite State Transducer (FST). Attention enabled encoder-decoder models was applied to G2P conversion in [58] and they gained the best performance of G2P conversion on CMU Dictionary by ensembling five global attention models using different random initializations. A sequence to sequence architecture called Transformer[2], which has achieved superior performance in machine translation task [62], has been used for G2P conversion.

Previous works on Myanmar G2P conversion using data driven approaches have been done on limited amount of Myanmar pronunciation dictionary and selected sentences [49, 50, 52]. Currently, neural sequence to sequence models was applied for Myanmar grapheme to phoneme conversion in our work [21] to explore the applicability of the large Myanmar pronunciation dictionary. In sequence to sequence model, the conversion is done by jointly learning the alignment and translation of input (grapheme) to output (phoneme). There is no explicit alignment between input and output sequences like the traditional joint-sequence models [5].

[2] https://github.com/cmusphinx/g2p-seq2seq

4.2 Building Pronunciation Dictionary for Myanmar Language

To experiment the state-of-the-art machine learning techniques on the low-resource languages like Myanmar, building a large pronunciation dictionary is an essential work to explore the benefits of machine learning techniques. There is only one published standard Myanmar pronunciation dictionary, Myanmar Language Commission (MLC) dictionary, which has about 27,300 words and it is not large enough to apply machine learning techniques. Therefore, a large pronunciation dictionary in Myanmar language was built to enhance the naturalness of Myanmar TTS system. The size of this dictionary is 100,000 entries. As much as possible pronunciations of each word are collected in the dictionary.

Commonly used Myanmar phrases and sentences are mainly focused in collecting and preparing the data. Various types of data from different domains such as news, articles, daily conversations, interviews, and public announcements have been collected from the internet and books. They are written in both formal writing and spoken style. Word segmentation [37] are done on the collected data sources because there is usually no delimiter between words in Myanmar language. Words are manually checked and segmented to get meaningful words as there is still ambiguity in word segmentation for Myanmar language. Finally, 100,000 words are selected for tagging pronunciation.

Tagging pronunciation includes modelling and manual correction on automatic tagged pronunciations. First of all, half of the collected 100,000 words are tagged with their pronunciations by a G2P model. This first model was trained by applying WFST based G2P toolkit, Phonetisaurus[3] on the MLC dictionary. After that, the pronunciations of those 50,000 words were corrected manually and used as training data with Sequence-to-Sequence Transformer model. This second model was used to get the pronunciation of the rest 50,000 words and manual checking was done on those pronunciations. Finally, the large Myanmar pronunciation dictionary including the 100,000 pairs of words and pronunciations was achieved. Examples of selected words and their pronunciations in the pronunciation dictionary are described in Table 4.1. To the best of our knowledge, this is the first large pronunciation dictionary for Myanmar language. It

[3] https://github.com/AdolfVonKleist/Phonetisaurus

can be applied not only in Myanmar TTS system but also in Myanmar Automatic Speech Recognition system.

Table 4.1 Example of Words and Pronunciations

Type of word	Word/Entity	In English	Pronunciation in MLC symbols
Common regular words in Myanmar	ဖျော့ဖျော့	light color	hp j o. b j o.
	စက်မှုဇုန်	industry	s e' hm u. z oun
	စကားစမြည်	chatting	z a- g a: s a- m j i
Common Myanmar name entities	ရွှေတိဂုံ	the most sacred pagoda in Myanmar	sh w ei d a- g oun
	နေပြည်တော်	the capital city of Myanmar	n ei p j i d o
	ပုပ္ပါး	the name of famous mountain in Myanmar	p ou' p a:
Common loan words	ဖုန်း	phone	hp oun:
	ကာလာ	color	k a l a
	ကားပါကင်	car parking	k a: p a k in
Common Foreign name entities	မိုက်ခရိုဆော့ဖ်	Microsoft	m ai' kh a- r ou hs o. HP
	ဂူဂဲလ်	Google	g u g e: L
	ယူကျု	Youtube	j u ky u.

4.3 Confirming the Quality of Myanmar Pronunciation Dictionary

Since the pronunciation dictionary is the warehouse of the knowledge concerning the orthography and pronunciation of words, the quality of the entries in the lexicon is needed as high as possible [48]. For the purpose of confirming the quality of Myanmar pronunciation dictionary for applying machine learning techniques, it has been currently deployed as the data source for training sequence to sequence G2P conversion models: joint sequence model, Transformer, simple encoder-decoder, and attention enabled encoder-decoder models [21]. The data setup for evaluating the pronunciation dictionary was shown in Table 4.2 and the entries in development and test sets were randomly selected from the dictionary.

Table 4.2 Data Setup for Evaluating the Pronunciation Dictionary

Set	Number of entries
Training set	93,000
Development set	5,000
Test set	2,000

The standard measures of phoneme error rate (PER) and word error rate (WER) were used for evaluating the performance of G2P models on the pronunciation dictionary and the results are presented in Table 4.3. According to the evaluation results, the quality of data is reliable and the amount of data is enough for exploring the benefits of machine learning techniques.

Table 4.3 Performance of G2P Conversion Models on the Pronunciation Dictionary

Model	PER(%)	WER(%)
Joint sequence model	1.7	10.0
Transformer	1.8	10.4
Simple encoder-decoder	2.5	13.5
Attention enabled encoder-decoder	2.1	12.6

4.4 Summary

The first large Myanmar pronunciation dictionary was constructed for Myanmar G2P conversion task and the detailed building processes for the pronunciation dictionary was reported in this chapter. The quality of the data was confirmed by applying neural sequence-to-sequence G2P models. This large Myanmar pronunciation dictionary has been used in G2P conversion of text analysis part for Myanmar TTS system.

CHAPTER 5
LINGUISTIC FEATURES AND WORD EMBEDDING FOR MYANMAR LANGUAGE

This chapter presents the introduction of Myanmar language, the phonetics of consonants, vowels, tones in Myanmar language, and linguistic features extraction for Myanmar language. A question set for Myanmar language used in Hidden Markov Model (HMM) based and neural network based speech synthesis is also described in this chapter. For the purpose of extracting linguistic features from real text, text analysis part of Myanmar language is also presented in this chapter. In addition to linguistic features that can be extracted by text analysis part, the way of getting word embedding for Myanmar language is also reported in detail.

5.1 Introduction to Myanmar Language

Myanmar language former known as Burmese is the official language of Myanmar and is spoken by 33 million people as a first language and by another 10 million people as a second language[1]. In Myanmar writing system, sentences are usually delimited by a unique sentence boundary marker "။". However, words are not always separated by spaces and spaces are sometimes used for phrase separation. A syllable is composed of one or more characters and one or more syllables can be formed as the word in Myanmar language. Myanmar words also have compound words and loan words. For some foreign words, a second killed consonant with parentheses such as (ဇ်), (လ်), (ပ်) are sometimes placed after the syllable to render a foreign sound.

5.1.1 Consonants in Myanmar Language

Myanmar script has 33 basic consonants, 4 basic medials, 12 basic vowels, other symbols, and special characters. The consonants have only 23 distinct pronunciation because some consonants have the same pronunciation in Myanmar language. For example, the consonants "ဂ" and "ဃ" have the same pronunciation /ga̰/. The place of articulation, the manner of articulation and the phonation gives the

[1] https://en.wikipedia.org/wiki/Burmese_language

consonant its distinctive sound. Table 5.1 shows Myanmar consonants with IPA symbols [60].

Table 5.1 Myanmar Consonants with IPA

Manner of articulation	**Place of articulation**						
	bilabial	**dental**	**alveolar**	**palato-alveolar**	**palatal**	**velar**	**glottal**
nasal (stop) voiced	မ /m/		န /n/	ည /ɲ/		င /ŋ/	
voiceless	မှ /m̥/		နှ /n̥/	ညှ /ɲ̊/		ငှ /ŋ̊/	
stop voiced	ဘ(ဗ) /b/		ဒ /d/			ဂ /g/	
voiceless	ပ /p/, ဖ /pʰ/		တ /t/, ထ /tʰ/			က /k/, ခ /kʰ/	
fricative voiced		သ /ð/	ဇ /z/				
voiceless		သ /θ/	စ/s/, ဆ/sʰ/	ရှ /ʃ/			
affricate voiced				ဂျ /dʑ/			
voiceless				ကျ /tɕ/, ချ /tɕʰ/			
central approximant voiced	ဝ /w/		(ရ) /ɹ/		ယ /j/		
voiceless	ဝှ /w̥/						ဟ /h/
lateral approximant voiced			လ /l/				
voiceless			လှ /l̥/				

5.1.2 Vowels in Myanmar Language

In Myanmar writing, there are basically 12 vowels and they are အ /a̰/, အာ /à/, အိ /ḭ/, အီ /ì/, အု /ṵ/, အူ /ù/, အေ /è/, အဲ /ɛ́/, အော /ɔ́/, အော် /ɔ̀/, အံ /a̰N/, အို /ò/. By extending these 12 basic vowels with tone markers and devowelizing consonants, all fifty vowels are occurred in Myanmar language and shown in Table 5.2.

Table 5.2 Myanmar Vowels with Myanmar Characters

basic symbol	non-nasalized vowels				nasalized vowels		
	Tone 1	Tone 2	Tone 3	Tone 4	Tone 1	Tone 2	Tone 3
အီ	အီ	အီး	အိ	အစ်	အင်	အင်း	အင့်
အေ	အေ	အေး	အေ့	အိတ်	အိန်	အိန်း	အိန့်
အယ်	အယ်	အဲ	အယ့်	အက်	အိုင်	အိုင်း	အိုင့်
				အိုက်			
အာ	အာ	အား	အ	အတ်	အန်	အန်း	အန့်
အော်	အော်	အော	အော့	အောက်	အောင်	အောင်း	အောင့်
အို	အို	အိုး	အို့	အုပ်	အုန်	အုန်း	အုန့်
အူ	အူ	အူး	အု	အွတ်	အွန်	အွန်း	အွန့်

Vowels differ only in the position of the tongue when voiced and they do not contain differences in voicing, manner, or place of articulation. Vowel quadrilateral of vowels in Myanmar language are described in Table 5.3 [60].

All the above distance features of consonants and vowels are taken into account for implementing Myanmar TTS system.

5.1.2 Tones in Myanmar Language

Myanmar is a tonal language and if the final glottal stop is regarded as a tonal feature and the non-final neural vowel /ə/ as an atonic vowel, it has four phonological tones [61]. Tone is the integral part of the pronunciation of syllable and can affect the meaning of that syllable. Table 5.4 shows an example of four phonological tones marked on Myanmar phoneme "ka".

Table 5.3 Myanmar Vowels in Vowel Quadrilateral

	Front	**Central**	**Back**
High	အိ /i/		အူ /u/
high	အေ /e/		အို /o/
Mid			
low	အယ် /ɛ/		အော /ɔ/
Low		အာ/a/	

Table 5.4 An Example of Four Phonological Tones in Myanmar

Tone	**IPA**	**Phonation**	**Length**	**Myanmar**	**Meaning**
Tone 1	kà	Normal	moderate	ကာ	cover
Tone 2	ká	Breathy	long	ကား	car
Tone 3	ka̰	Creaky	short	က	dance
Tone 4	kaʔ	Final Glotal stop	abrupt	ကပ်	stick

5.2 Linguistic Features Extraction for Myanmar Language

Many contextual features are taken into account in modelling SPSS because they can affect spectrum, F_0 pattern, and duration of the synthetic speech. These features are language dependent ones. Therefore, we have to extract linguistic features from plain text sentence for Myanmar language. The general speech synthesis architecture of Festival[2] has been configured for extracting contextual information from utterances for Myanmar language. However, there is no phoneme features file and pronunciation dictionary for Myanmar language in Festival. The detailed preparation will be presented in the following sections.

[2] http://www.cstr.ed.ac.uk/projects/festival/

5.2.1 Preparing Phoneme Features File

Phoneme features are prepared for consonants such as consonant type (nasal, stop, fricative, affricate, etc.), place of articulation (bilabial, dental, alveolar, etc.), consonant voicing, and lip rounding, and for vowels such as vowel frontness, vowel height, tone (Tone1, Tone2, Tone3 and Tone4), and nasality in phoneme features file which is used in Festival. Acoustic Phonetics and the Phonology of the Myanmar language book written by Dr. Thein Tun [60] was used as the reference.

5.2.2 Preparing Lexicon with Syllable Information

A large Myanmar pronunciation dictionary described in Chapter 4 was used the main entries of Myanmar lexicon in Festival. Moreover, Myanmar Language Commission (MLC) dictionary [80], words from training sentences, suffixes such as "ခဲ့ပါတယ်" (suffix of verb to past tense), "ပါရစေ" (suffix of verb to form modal verb), "လိမ့်မယ်" (suffix of verb to future tense), "ကြိမ်မြောက်" (suffix of noun to ordinal number) and syllables were included in Myanmar lexicon. The total number of entries in the lexicon is 117,400.

Syllable information is also included in the pronunciation dictionary for Myanmar language because syllable is the basic sound unit bearing tone information in Myanmar language. Standard Myanmar phoneme symbols and extended phoneme symbols for foreign words [50] were used in this lexicon. Table 5.5 shows the format of Myanmar lexicon applying in Festival.

Table 5.5 Myanmar Lexicon Format in Festival

Myanmar Word	Part-of-Speech (POS)	Pronunciation
စကားစမြည်	n	(((z a-) 1) ((g a:) 1) ((s a-) 1) ((m j i) 1)))
အံ့သြစရာ	adj	(((an.) 1) ((o:) 1) ((z a-) 1) ((j a) 1)))
ယူအက်စ်ဘီ	fw	(((j u) 1) ((e' S) 1) ((b i) 1)))

5.2.3 Contextual Linguistic Labels for Myanmar Language

By applying the phoneme features, Myanmar pronunciation dictionary, and preparing some configurations on Festival, the file with linguistic information for each utterance can be gained. After that, contextual linguistic labels for Myanmar language have been extracted from that file and formatted as HTS-style labels[3]. They are the language dependent requirements and the following contextual factors are taken into account for Myanmar language.

Phoneme level:

- the current phoneme, and preceding and succeeding two phonemes
- the position of the current phoneme in the current syllable (forward, backward)

Syllable level:

- the number of phonemes in the preceding, current, and succeeding syllables
- the position of the current syllable in the current word (forward, backward) and in the utterance (forward, backward)
- the number of syllables before and after the current syllable in the utterance
- the vowel identity within the current syllable

Word level:

- the number of syllables in the preceding, current, and succeeding words
- the position of current word in the utterance (forward, backward)
- the number of words before and after the current word in the utterance

Utterance level:

- the number of syllables in the utterance
- the number of words in the utterance

However, Part of Speech (POS) information and intonation information such as tones and break indices (ToBI) are not taken as the features.

An example Myanmar utterance of word segmented, syllable segmented and its phonemes is shown in Figure 5.1 and its structure is shown in Figure 5.2. Contextual linguistic labels for the example utterance are extracted according to its structure and example contextual labels for the phoneme "ht" in that utterance are described in Figure 5.3. In the figure, (1) is the quinphone information, (2) is the position of phoneme in the current syllable (forward, backward), (3) is the structure of previous syllable, (4) is the structure of current syllable, and (5) is the position of current syllable in the current word (forward, backward) for phoneme "ht" in the example utterance.

[3] http://www.cs.columbia.edu/~ecooper/tts/lab_format.pdf

Utterance: နောက်ထပ် အားသာချက်ကတော့ မိသားစုနဲ့ ခရီးသွားဖို့ အဆင်ပြေတဲ့ ကျယ်ဝန်းမှုရှိတာပါပဲ ။

Word: နောက်ထပ် | အားသာချက် | က | တော့ | မိသားစု | နဲ့ | ခရီးသွား | ဖို့ | အဆင်ပြေ | တဲ့ | ကျယ်ဝန်းမှု | ရှိတာပါပဲ | ။

Syllable: နောက် | ထပ် | အား | သာ | ချက် | က | တော့ | မိ | သား | စု | နဲ့ | ခ | ရီး | သွား | ဖို့ | အ | ဆင် | ပြေ | တဲ့ | ကျယ် | ဝန်း | မှု | ရှိ | တာ | ပါ | ပဲ | ။

Phoneme: n au’ ht a’ a: th a gy e’ k a. d o. m i. th a: z u. n e. kh a- j i: th w a: b o. a- hs in pj ei d e. ky e w un: h m u. sh i. d a b a b e pm

Figure 5.1 Example of Syllable and Word Segmented Myanmar Utterance with its Phonemes

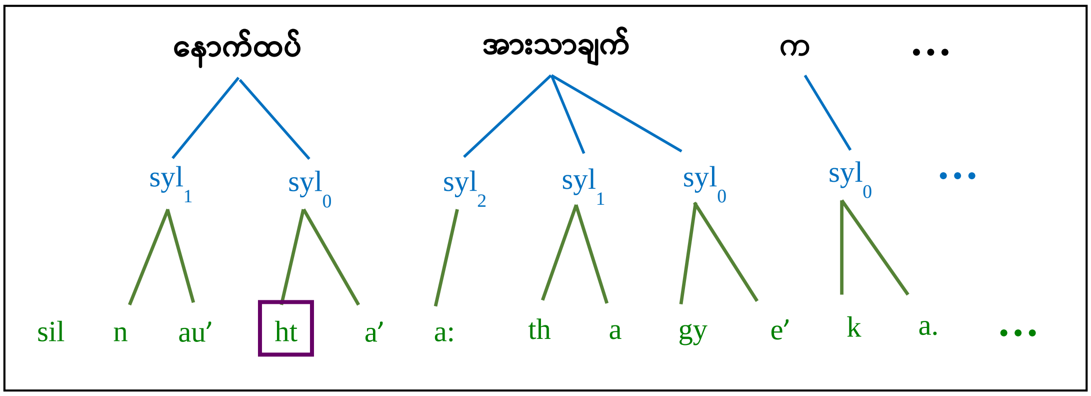


Figure 5.2 Example Myanmar Utterance with its Structure

n^au’-**ht**+a’=a:@1_2/A:1_1_2/B:1-1-2@2-1 …

(1) (2) (3) (4) (5)

Figure 5.3 Example Contextual Labels for Phoneme “ht” in the Utterance

5.2.4 Question Set for Myanmar Language

A question set is used for context clustering of HMM-based speech synthesis and extracting linguistic features of neural network based speech synthesis and it is also language dependent requirement. There is no publicly available question set for

Myanmar language yet. Therefore, questions were manually prepared for Myanmar language and the articulatory features of language such as the place of articulation and the manner of articulation were taken into account. Questions related to phonetic characteristics of consonants, vowels, diphthongs, and tones have been derived. Since Myanmar is a tonal language, tone dependent questions have been considered in the question set. The updated Myanmar question set has 635 questions including 622 phoneme questions and 13 related positional questions. Table 5.6 shows the group of phonemes that are taken into account for preparing the questions specific to Myanmar language and example entries in the question set are shown in Figure 5.4.

Table 5.6 Grouping of Phonemes for Myanmar Question Set

Group Name	**Phonemes**
Vowel	All vowels
Consonant	All consonants
Foreign	Phonemes used in some foreign pronunciations
Stop	k, kh, g, t, ht, d, p, hp, b
Nasal	ng, nj, n, m
Fricative	s, hs, z, th, dh, sh
Front Vowel	i, i:, i., i′, in, in:, in., ei, ei:, ei., ei′, ein, ein:, ein., e, e:, e., e′, ain, ain:, ain., ai′
Central Vowel	a-, a, a:, a., a′, an, an:, an.
Back Vowel	o, o:, o., au′, aun, aun:, aun., ou, ou:, ou., ou′, oun, oun:, oun., u, u:, u., u′, un, un:, un.
Tone1 Vowel	i, in, ei, ein, e, ain, a, an, o, aun, ou, oun, u, un
Tone2 Vowel	i:, in:, ei:, ein:, e:, ain:, a:, an:, o:, aun:, ou:, oun:, u:, un:
Tone3 Vowel	i., in., ei., ein., e., ain., a., an., o., aun., ou., oun., u., un.
Tone4 Vowel	i′, ei′, e′, ai′, a′, au′, ou′, u′
Neutralized vowel	a-
High Vowel	i, i:, i., i′, in, in:, in., u, u:, u., u′, un, un:, un.
Medium Vowel	ei, ei:, ei., ei′, ein, ein:, ein., e, e:, e., e′, ain, ain:, ain., ai′, o, o:, o., au′, aun, aun:, aun., ou, ou:, ou., ou′, oun, oun:, oun.
Low Vowel	a-, a, a:, a., a′, an, an:, an.

Rounded Vowel	o, o:, o., au′, aun, aun:, aun., ou, ou:, ou., ou′, oun, oun:, oun., u, u:, u., u′, un, un:, un.
Unrounded Vowel	i, i:, i., i′, in, in:, in., ei, ei:, ei., ei′, ein, ein:, ein., e, e:, e., e′, ain, ain:, ain., ai′, a-, a, a:, a., a′, an, an:, an.
Unvoiced Consonant	k, kh, s, hs, t, ht, p, hp, th, h, sh, ky, ch
Voiced Consonant	g, ng, z, nj, d, n, b, m, j, r, l, w, dh, gy
Front Consonant	p, hp, b, m, w, th, dh
Central Consonant	s, hs, z, nj, t, ht, d, n, j, r, l, sh, gy, ky, ch
Back Consonant	k, kh, g, ng, h
No Continuant	k, kh, g, ng, nj, t, ht, d, n, p, hp, b, m
Voiced Stop	b, d, g
Unvoiced Stop	p, hp, t, ht, k, kh
Affricate Consonant	ky, ch, gy
Voiced Fricative	dh, z
Unvoiced Fricative	th, s, hs, sh

```
QS "C-Stop" {-k+,-kh+,-g+,-t+,-ht+,-d+,-p+,-hp+,-b+}
QS "C-Nasal" {-ng+,-nj+,-n+,-m+}
QS "C-Fricative" {-s+,-hs+,-z+,-th+,-dh+,-sh+}
QS "C-Front_Vowel" {-i+,-ic+,-id+,-ia+,-in+,-inc+,-ind+,-ei+,-eic+,-eid+,-eia+,
-ein+, -einc+,-eind+,-e+,-ec+,-ed+,-ea+,-ain+,-ainc+,-aind+,-aia+}
QS "C-Tone1_Vowel" {-i+,-in+,-ei+,-ein+,-e+,-ain+,-a+,-an+,-o+,-aun+,-ou+,
-oun+, -u+,-un+}
CQS "Num-Syls_in_Utterance"                {/J:(\d+)+}
CQS "Num-Words_in_Utterance"               {+(\d+)-}
```

Figure 5.4 Example Entries in Question Set for Myanmar Language

5.3 Text Analysis for Myanmar Language

Figure 5.5 shows the process flow of text analysis part for Myanmar TTS system, which is also important for the quality of synthesized speech. Word segmentation is the first process for text analysis phase because Myanmar text generally lacks white space between words though spaces are sometimes used for separating among phrases for reading easily. We used word segmentation tool in [37] for word segmentation process. For text normalization, WFST based Myanmar number normalization system described in Chapter 3 was integrated into the pipeline of text analysis. A large Myanmar pronunciation dictionary reported in Chapter 2 with syllable information was applied in grapheme to phoneme (G2P) conversion. For contextual labels extraction, we configured and used Festival speech synthesis architecture with prepared phoneme features file, our large Myanmar pronunciation dictionary. Finally, HTS style contextual labels for the input text are extracted for utilizing in modelling speech synthesis for Myanmar language. Linguistic features for Myanmar language will be extracted from these contextual labels by applying our proposed question set in Section 5.2.4.

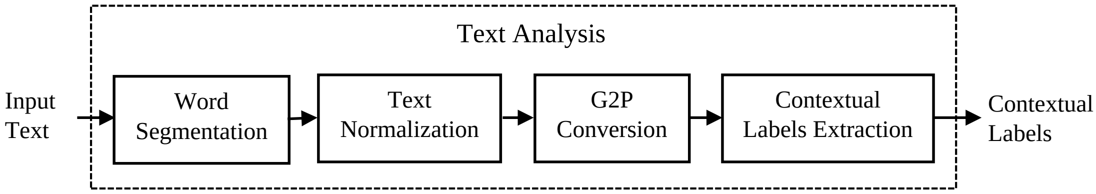


Figure 5.5 Process Flow of Text Analysis for Myanmar Language

5.4 Word Representations for Myanmar Text-to-Speech System

Some suprasegmental features for intonation prediction such as ToBI are not included in linguistic features generated by the text analysis part. They can only be achieved by manually annotated training corpus with high consistency among different annotators. This kind of annotation is time consuming and very expensive. Therefore, distributed word representations or word vectors which can be obtained by unsupervised learning from large amount of unstructured text data have been applied in Myanmar speech synthesis to improve the naturalness of the synthesized speech. Recently, word vectors have been applied in speech synthesis [42, 63, 64, 65]. These distributed representations of words in a vector space can capture a large number of precise syntactic and sematic word relationships [32]. Word vector features have been used in acoustic modeling of DNN-based and LSTM-RNN based Myanmar speech synthesis as the additional input features together with the linguistic features extracted from text analysis part. Firstly, we collected monolingual Myanmar text corpus for building word embedding for Myanmar language. After that, different dimensions of word vectors for Myanmar language are modelled by applying Continuous Bag-of-

Words (CBOW) and Skip-gram modelling methods. The experiments and results of these word vector features on Myanmar speech synthesis can be found in Chapter 7 and 8.

5.4.1 Building Word Embedding for Myanmar Language

Though many pre-trained word vectors can be retrieved from the Web, only two sets of word vectors for Myanmar language are found publicly in [1, 16]. In [1], the size of word vectors is small, and it contains about 55K entries for Myanmar language and can be downloaded from the link[4]. In [16], the word vectors are trained on Common Crawl and Wikipedia using fastText. The size of pre-trained word vector for Myanmar language (Burmese) is about 335K entries[5]. Myanmar word vector of fastText contains different encodings such as Zawgyi and Unicode. The coverage of Polyglot and fastText Myanmar word vectors on our training corpus used in speech synthesis is hardly enough to apply in our speech synthesis. That is why word vectors are built for Myanmar language with standard encoding Unicode for more coverage and better performance.

5.4.1.1 Data Collection

Firstly, a large monolingual Myanmar corpus is collected for the purpose of building high quality word vectors with wide coverage. Myanmar data from Asian Language Treebank (ALT) parallel corpus [43] is used as one of the data sources. It comprises 20,000 sentences translated from English texts sampled from English Wikinews and is an annotated corpus including word segmentation. Another one is Myanmar data of ASEAN-MT parallel data [38], and it also consists of 20,000 sentences in travel domain with segmented words. Another large dataset is collected from Myanmar websites and blogs, and the data size has about 436,000 sentences. It is general domain including news, business, health, politics, tourism, education, arts, technology, sport, and religion. The text data from the training speech corpus (4,000 sentences) is also included in the monolingual Myanmar corpus. Finally, it contains about 480,000 sentences. The statistics of that monolingual corpus is shown in Table 5.7.

[4] https://polyglot.readthedocs.io/en/latest/Download.html

[5] https://fasttext.cc/docs/en/crawl-vectors.html

5.4.1.2 Preprocessing

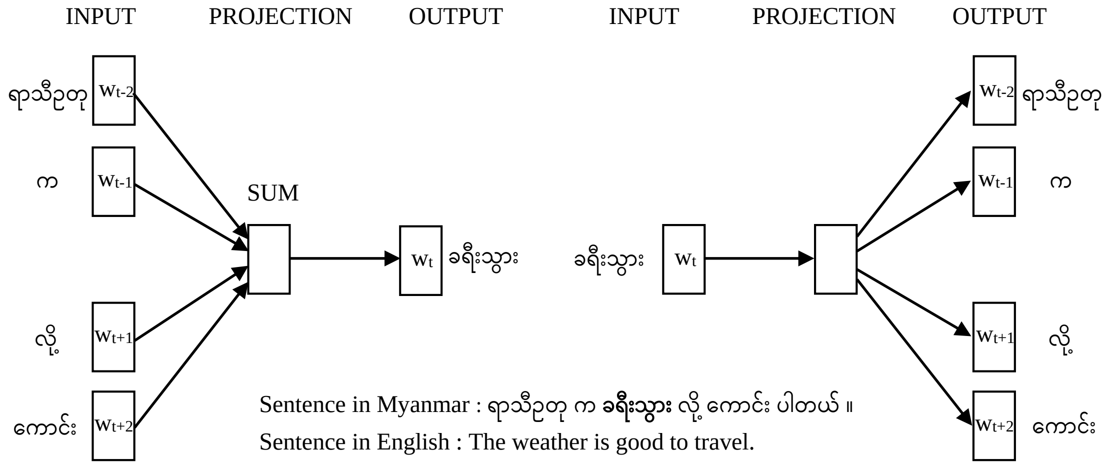
There are three steps in the preprocessing: data cleaning, standardizing encoding, and word segmentation. Some characters such as “...”, “[”, “]”, “*”, special characters used in some web pages such as characters for telephone icon, emotional icons, and Myanmar sentence marker “။” were removed from the data source. The second step is standardizing encoding, and this is converting Zawgyi, partial unicode commonly used in the Myanmar blogs, to standard Unicode to make the data processing more easily. In the final step, word segmentation [37] was done on the data collected from websites and blogs. This corpus was used in modelling word vectors for Myanmar language.

Table 5.7 The Statistics of Monolingual Myanmar Corpus

Data Source	Domain	Ratio
ALT	Wikinews	4%
ASEAN-MT	Travel	4%
data from speech corpus	Travel	1%
Webs	General	91%

5.4.1.3 Modelling Word Vectors

Word embedding is a low dimension continuous-valued vector used for representing word. Representation of words as continuous vectors can be learned by using neural network language model (NNLM) [3, 33]. Two particular models for learning word representations that can be efficiently trained on large amounts of text data are CBOW and Skip-gram models [32, 34]. They can be trained for getting improvements in accuracy at much lower computational cost. Therefore, we learned distributed word representations of Myanmar language by applying CBOW and Skip-gram models. In the CBOW model, continuous distributed representation of the context (surrounding words) are combined to predict the word in the middle. The Skip-gram architecture is similar to CBOW, but instead of predicting the current word based on the context, it tries to maximize classification of a word based on neighboring words within a sentence. Distributed representation of current word is used to predict words within a certain range before and after the current word. Figure 5.6 shows the architectures of CBOW and Skip-gram models for an example Myanmar sentence.



(a) CBOW (b) Skip-gram

Figure 5.6 Model Architectures of CBOW and Skip-gram

More formally, given a sequence of training words $[w_1, w_2, \dots, w_T]$, modelling Skip-gram is done by maximizing the average log probability of the word w_{t+j} given the center word w_t,

$$\frac{1}{T}\sum_{t=1}^{T} \sum_{-c \le j \le c, j \ne 0} \log p(w_{t+j}|w_t) \tag{5.1}$$

where c is the size of training context. The basic Skip-gram defines $p(w_{t+j}|w_t)$ using softmax function and the formulation is expensive. Therefore, an efficient Negative sampling (NEG) is used to replace over $log\, p(w_{t+j}|w_t)$ term in Skip-gram formulation [32]. This NEG is also used for training the CBOW model.

Using the monolingual corpus for Myanmar language, we trained word vectors with word2vec[6], a tool for computing continuous distributed representations of words. The CBOW and Skip-gram models with different choice of word vector dimensionality (100, 200, and 300) were trained on the monolingual Myanmar corpus. The training process was iterated 15 times with negative sampling. The size of training context for all models was set to 8. The word vector set covers 97.1% of the training corpus for speech synthesis. The trained word vectors are shown in Table 5.8 and these word vectors have been applied in neural network based Myanmar speech synthesis. The aliases in Table 5.8 will be used for referring these word vectors in next chapters.

[6] https://code.google.com/p/word2vec/

The number of total words in training is 9,068,590 and the vocabulary size is 197,307. Analogy datasets on Myanmar language for evaluating these models is not currently available. Examples of the nearest five neighbours of Myanmar words that can be generated from Myanmar word vector model are shown in Table 5.9 and input words are shown in the bold style.

Table 5.8 Trained Word Vectors for Myanmar Language

Alias	Training Method	Dimension
W1	CBOW	100
W2	CBOW	200
W3	CBOW	300
W4	Skip-gram	100
W5	Skip-gram	200
W6	Skip-gram	300

Table 5.9 Examples of the Nearest Five Neighbors for Input Words "ဒေါ်လာ" (Dollar) and "အနီရောင်" (Red)

Word	(English)	Word	(English)
ဒေါ်လာ	Dollar	**အနီရောင်**	Red
ရူပီး	Rupee	အဝါရောင်	Yellow
အမေရိကန် ဒေါ်လာ	American Dollar	မီးခိုးရောင်	Grey
မြန်မာငွေ	Myanmar money	အပြာရောင်	Blue
ပီဆို	Peso	အနက်ရောင်	Black
မြန်မာကျပ်	Myanmar Kyat	အစိမ်းရောင်	Green

5.5 Summary

In this chapter, first, the way of how to prepare phoneme features file and lexicon for the language-specific purpose has been presented and contextual linguistic features for Myanmar language have been extracted by configuring and preparing in Festival. The nature of Myanmar language is also briefly described. Finally, integration of various modules in the text analysis part for Myanmar TTS system are described and contextual labels generated from the text analysis have been applied in all modelling of speech synthesis which will be reported in next chapters. A question set for Myanmar language has also been proposed for extracting linguistic features of neural network based speech synthesis. Training detail and trained word vectors which can be used as the additional input features for Myanmar speech synthesis are also presented in this chapter. Contextual linguistic features from text analysis part and word embedding features from our trained word vectors will be applied in Myanmar TTS system to promote the naturalness of the synthetic speech in next chapters.

CHAPTER 6

HIDDEN MARKOV MODELS BASED MYANMAR SPEECH SYNTHESIS

Statistical parametric speech synthesis which uses a hidden Markov models (HMM) as its generative model is typically called HMM-based speech synthesis. HMM-based speech synthesis system (HTS) was applied to Myanmar speech synthesis using the general speech synthesis architecture of Festival configured for Myanmar language described in Section 5.2. This chapter describes the basic structure and the algorithms of the Hidden Markov Models (HMM) based speech synthesis and the implementation of the baseline HMM-based Myanmar speech synthesis. As the preliminary experiment of examining the importance of word information, CLUSTERGNE model has been trained for Myanmar language and results are reported in this chapter.

6.1 Hidden Markov Model

A hidden Markov model (HMM) is a statistical Markov model in which the system being modelled is assumed to be a Markov process with unobservable states. At each time instance, the HMM changes the states at Markov process in agreement with a state transition probability, and then produces observational data ***o*** in accordance with an output probability distribution of the current state.

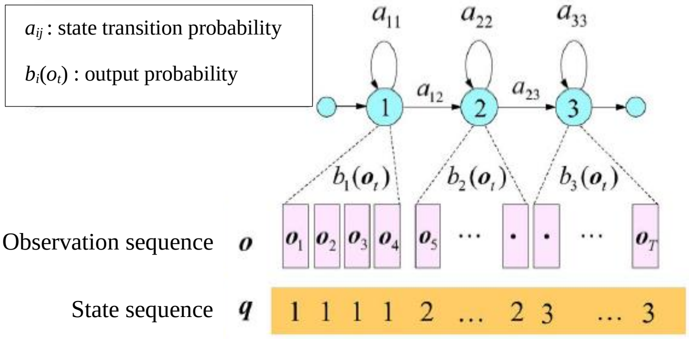

a_{ij} : state transition probability
$b_i(o_t)$: output probability
a_{11}
a_{22}
a_{33}
a_{12}
a_{23}
1
2
3
$b_1(\boldsymbol{o}_t)$
$b_2(\boldsymbol{o}_t)$
$b_3(\boldsymbol{o}_t)$
Observation sequence $\boldsymbol{o}$
$\boldsymbol{o}_1$ $\boldsymbol{o}_2$ $\boldsymbol{o}_3$ $\boldsymbol{o}_4$ $\boldsymbol{o}_5$ … • • … $\boldsymbol{o}_T$
State sequence $\boldsymbol{q}$
1 1 1 1 2 … 2 3 … 3

Figure 6.1 Example of Three-State, Left-to-Right HMM [57]

Figure 6.1 shows an example of a three-state left-to-right HMM. An *N*-state HMM λ is characterized by sets of initial-state probabilities$\{\pi_i\}_{i=1}^{N}$, state-transition probabilities $\{a_{ij}\}_{i,j=1}^{N}$, and state-output probability distributions $\{b_i(\cdot)\}_{i=1}^{N}$. The $\{b_i(\cdot)\}_{i=1}^{N}$ are typically assumed to be single multivariate Gaussian distributions

$$b_i(o_t) = N(o_t; \mu_i; \Sigma_i) \tag{6.1}$$

$$b_i(o_t) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_i|}} exp\left\{-\frac{1}{2}(o_t - \mu_i)^{\mathrm{T}} \Sigma_i^{-1} (o_t - \mu_i)\right\} \tag{6.2}$$

where μ_i and Σ_i are a d-by-1 mean vector and a d-by-d covariance matrix, respectively; d is the dimension of the acoustic parameters; and o_t is an observation vector, which consists of the vocoder parameters at frame t [57].

Let $O = \left[O_1^{\mathrm{T}}, O_2^{\mathrm{T}}, \dots, O_{\mathrm{T}}^{\mathrm{T}}\right]^{\mathrm{T}}$, and W be a set of speech parameters and corresponding linguistic specifications to be applied for training HMMs, respectively. The training of a HMMs is simply written as follows:

$$\lambda_{max} = \arg\max_{\lambda} p(O|\lambda, W) \tag{6.3}$$

$$p(O|\lambda, W) = \sum_{\forall q} \pi_{q_0} \prod_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(O_t) \tag{6.4}$$

where $q = \{q_1, q_2, \dots, q_T\}$ is a state sequence.

Let $o = \left[o_1^{\mathrm{T}}, o_2^{\mathrm{T}}, \dots, o_{T'}^{\mathrm{T}}\right]^{\mathrm{T}}$ and w be speech parameters and corresponding linguistic specifications to be generated at synthesis time. The synthesis from HMMs can be written as follows:

$$o_{max} = \arg\max_{o} p(o|\lambda_{max}, w) \tag{6.5}$$

6.2 Speech Parameter Generation for HMM

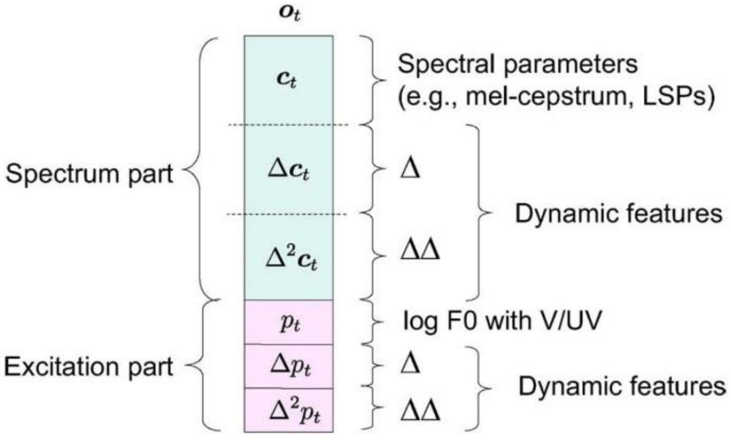


Figure 6.2 Example of an Observation Vector at Each Frame [57]

Generating the optimal speech parameter vector sequence given a set of HMMs and the input linguistic specification is to get a vector sequence of speech parameters, $o = \left[o_1^{\mathrm{T}}, o_2^{\mathrm{T}}, \dots, o_{T'}^{\mathrm{T}}\right]^{\mathrm{T}}$ which maximize $p(o|\lambda_{max}, w)$ with respect to o,

$$o_{max} = \arg\max_{o} p(o|\lambda_{max}, w) \tag{6.6}$$

$$o_{max} = \arg\max_{o} \sum_{\forall q} p(o, q|\lambda_{max}, w) \tag{6.7}$$

To approximate this problem, the most likely sequence is employed in the same way as Viterbi algorithm,

$$o_{max} \approx \arg\max_{o,q}(o, q|\lambda_{max}, w) \tag{6.8}$$

The joint probability of o and q can be simply written as

$$o_{max} = \arg\max_{o,q} p(o|q, \lambda_{max})P(q|\lambda_{max}, w) \tag{6.9}$$

The optimization problem of the probability of the observation sequence o is divided into the following two optimization problems:

$$q_{max} = \arg\max_{q} P(q|\lambda_{max}, w) \tag{6.10}$$

$$o_{max} \approx \arg\max_{o} p(o|q_{max}, \lambda_{max}) \tag{6.11}$$

The maximization problem of Eq. (6.10) can be solved by state-duration probability distributions and the maximization problem of Eq. (6.11) is maximizing $p(o|q, \lambda)$ with respect to *o* given the predetermined state sequence q_{max} [57].

If the parameter vector at frame *t* is determined independently of preceding and succeeding frames, this will cause discontinuity in the generated spectral sequence at transitions of states that degrade the quality of synthesized speech. To avoid this case, it is assumed that the speech parameter vector o_t consists of the *M*-dimensional static feature vector c_t (e.g., cepstral coefficients) and the *M*-dimensional dynamic feature vectors $\Delta c_t, \Delta^2 c_t$ (e.g., delta and delta-delta cepstral coefficients) as

$$o_t = [c_t^{\mathrm{T}}, \Delta c_t^{\mathrm{T}}, \Delta^2 c_t^{\mathrm{T}}]$$

The dynamic feature vectors are determined by linear combination of the static feature vectors of several frames around the current frame [70]. Let $\Delta^{(0)} c_t = c_t$, $\Delta^{(1)} c_t = \Delta c_t$, and $\Delta^{(2)} c_t = \Delta^2 c_t$, the general form $\Delta^{(n)} c_t$ is defined as

$$\Delta^{(n)} c_t = \sum_{\tau=-L_-^{(n)}}^{L_+^{(n)}} w_{t+\tau}^{(n)} c_t \qquad , 0 \le n \le 2 \tag{6.12}$$

Where $L_{-}^{(0)} = L_{+}^{(0)} = 0$ and $w_{0}^{(0)} = 1$. Then, the optimization problem of the observation sequence o is considered to be maximizing $p(o|q_{max}, \lambda_{max})$ with respect to c under the constraints Eq. (6.12).

A typical form of the observation vector, which includes both static and dynamic features, is shown in Figure 6.2.

6.3 HMM based Speech Synthesis

Figure 6.3 shows the overview of HMM-based speech synthesis system. The detail description of training part and synthesis part is as follows:

6.3.1 Training Part

In the training part, both spectral parameters such as mel-cepstral coefficients and excitation parameters such as *F₀* are modeled at the same time. Sequences of mel-cepstral coefficient vector extracted from speech database using a mel-cepstral analysis technique [15] are modeled by continuous density HMMs. The conventional discrete or continuous HMMs cannot be applied for F_0 modeling because the observation sequence of an F_0 pattern is comprised of one-dimensional continuous values and a discrete symbol representing “unvoiced”. Therefore, multi-space probability distribution HMM (MSD-HMM) is applied to F_0 pattern modeling and generation [54]. Voiced and unvoiced observations of F_0 can be modeled in a unified model without any heuristic assumption [55]. To keep synchronization between spectral parameters and excitation parameters, they are modeled simultaneously by separate streams in a multistream HMM [73]. For duration modeling, HMM-based speech synthesis typically uses a semi-Markov structure to model the temporal structure of speech [78].

In HMM-based speech synthesis, various phonetic, prosodic, and linguistic contexts are used for the context-dependent modeling of HMMs to model variations in spectrum, pitch, and duration. However, it is impossible to prepare speech database which covers all combinations of contextual factors. To alleviate this problem, decision-tree-based context clustering technique [46, 74] is applied to cluster similar states and to tie model parameters among several context-dependent HMMs. Since each of spectrum, pitch, and duration has its own influential contextual factors, the distributions for the spectral parameter, pitch parameter, and state duration are clustered independently.

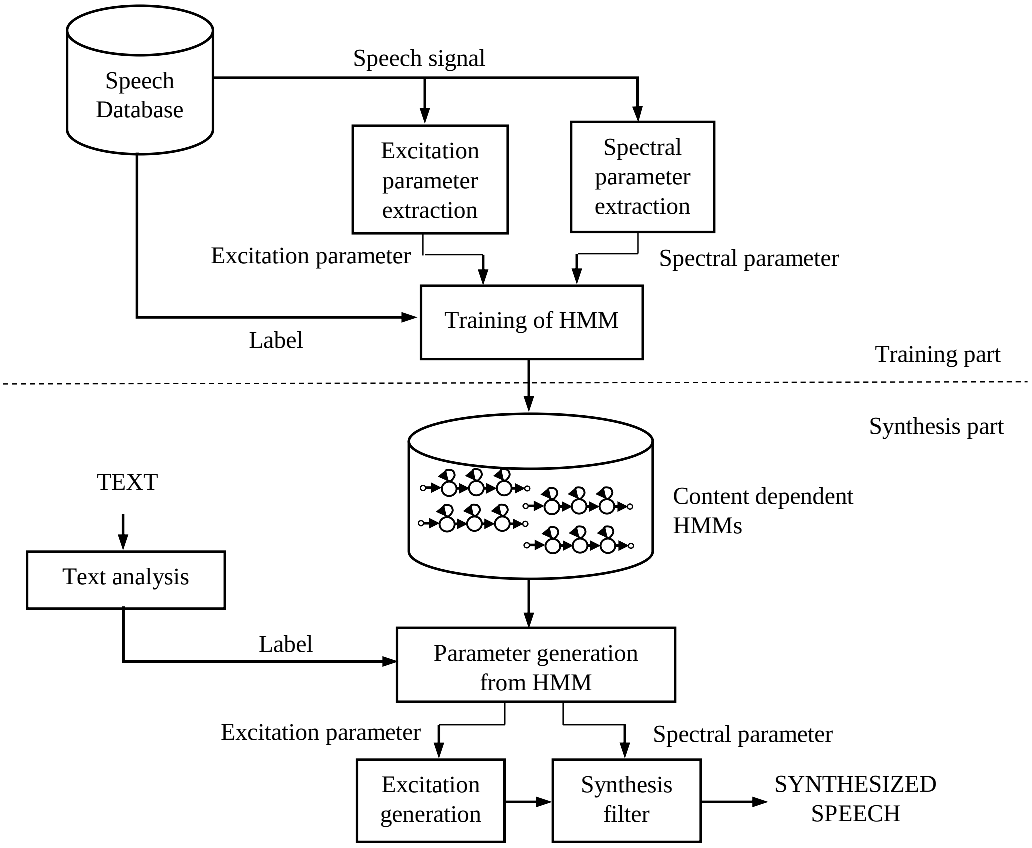


Figure 6.3 Overview of HMM-based Speech Synthesis System [53]

6.3.2 Synthesis Part

In the synthesis part of HMM-based system, an arbitrarily given text is converted into a sequence of context-dependent phoneme labels. According to these labels, a sentence-level HMM is constructed by concatenating context-dependent phoneme HMMs. Then, spectral and F_0 parameter sequences are determined so as to maximize their output probabilities using the speech parameter generation algorithm described in Section 6.2. Finally, speech is synthesized from the generated mel-cepstral and F_0 parameter sequences by using mel-log spectral approximation filter (MLSA) [15].

6.4 HMM based Myanmar Speech Synthesis

HMM-based speech synthesis for Myanmar language was implemented and used as the baseline system for this research. Figure 6.4 shows the overview of the synthesis procedures in HMM based Myanmar speech synthesis. In the synthesis time, a given text is converted to a contextual label sequence according to the text analysis module. According to label sequence, the speech parameters such as mel-cepstral and F_0, i.e. the statistics of static and dynamic feature sequence are generated from decision tree-clustered context dependent HMMs. In order to generate smooth speech parameter trajectories, they are generated considering the relation between static and dynamic features by maximum likelihood parameter generation (MLPG) algorithm [56]. Finally, speech waveform is synthesized from the generated mel-cepstral and F_0 parameter sequences by using MLSA filter.

The speech corpus used in training and the detailed training processes of the system are described in the following sections.

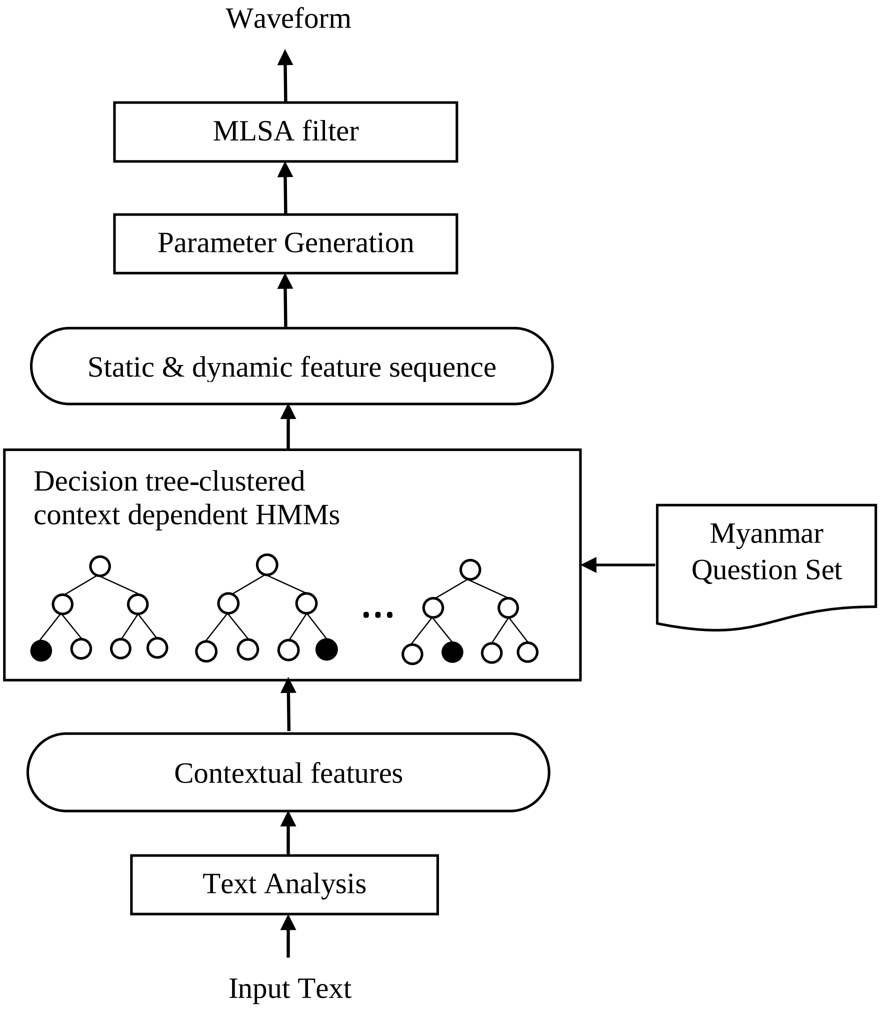


Figure 6.4 An Overview of Synthesis Procedures in HMM-based Myanmar Speech Synthesis

6.5 Myanmar Speech Synthesis with CLUSTERGEN

As our preliminary experiments, word information is applied to Myanmar speech synthesis using CLUSTERGEN [6], a statistical parametric synthesizer that has been created within the widely used Festival/FestVox voice building suite [7]. In Myanmar language, word information is important for the naturalness of Myanmar speech because pronunciation changes depend on word information. Figure 6.5 depicts the workflow of Myanmar TTS using CLUSTERGEN method.

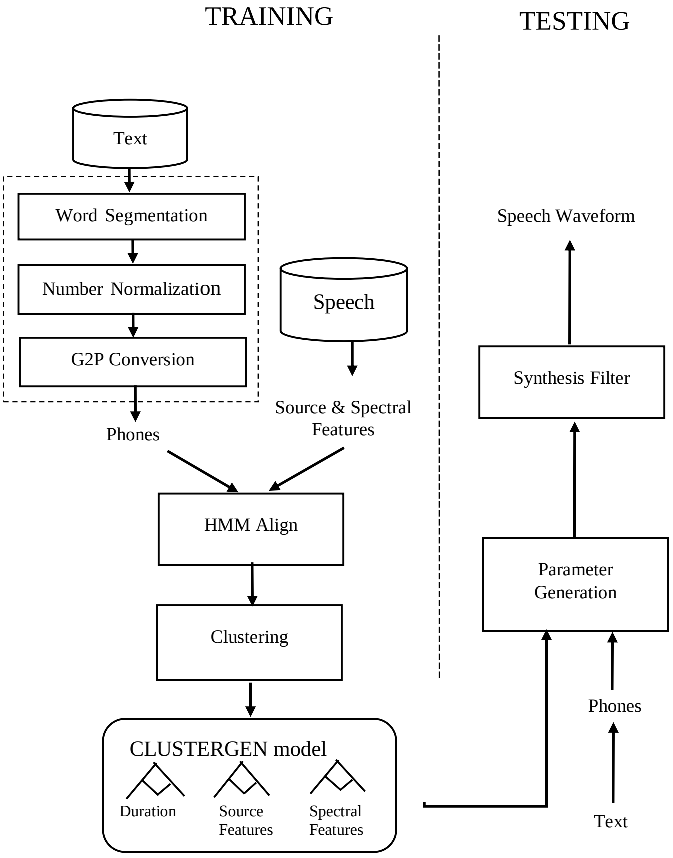


Figure 6.5 The Workflow of Myanmar TTS Using CLUSTERGEN

In the training phase, source feature and spectral features are extracted from speech database. As the preprocessing step, word segmentation, number normalization, and G2P conversion are applied on the transcription. F_0 and 24 MFCCs were extracted from speech database and combined 25 feature vector for every 5ms. No delta features are used in modelling CLUSTERGEN. Ergodic Hidden Markov Model (EHMM) [39] labeler is used to force align the phonemes generated from the transcriptions with the audio. The speech features are then clustered using available phonetic and high level features at the phoneme state level. Clustering is done for building source and spectral CART trees, and duration CART tree. CART trees are built to find questions that split the data to minimize impurity. The impurity is calculated as

$$N * \left(\sum_{i=1}^{24} \sigma_i \right) \tag{6.13}$$

Where N is the number of samples in the cluster and σ_i is the standard deviation for MFCC feature *i* over all samples in the cluster.

In the testing phase, the phone string is generated from the input text. CLUSTERGEN model is used to generate the feature parameters and duration. Smoothing is done by a simple 3-point moving average to each track of coefficient as in the following Equation:

$$s_t' = (s_{t-1} + s_t + s_{t+1})/3.0 \tag{6.14}$$

where s_n is the sample at time point n. CLUSTERGEN generates the utterance frame by frame, rather than by state, it allows more detailed modeling [2]. Then the speech is reconstructed from predicted parameters using MLSA filter [15].

6.6 Speech Corpus

The construction of robust TTS system needs to use the phonetically balanced corpus (PBC) for modeling. Myanmar phonetically balanced corpus (PBC) [51] built from Basic Travel Expression Corpus (BTEC) [27] recorded by a female speaker was used as the training corpus. 4,000 utterances were recorded in a sound-proof room at National Institute of Information and Communications Technology (NICT) with reading style. For the CLUSTERGEN and the HMM based Myanmar speech synthesis, 3,900 utterances were used as the training set, 100 utterances as the test set. These speeches were down-sampled from 48kHz to 16kHz. This speech corpus was applied in all experiments of Myanmar speech synthesis in this research.

6.7 Experiments

Speech corpus described in Section 6.6 was used in modelling CLUSTERGEN voice for Myanmar language. To evaluate the naturalness of syllable and word based speech synthesis, subjective evaluation has been done on two models: HMM-based model using syllable information [51] and CLUSTERGEN model using word information. HMM-based Myanmar speech synthesis is also conducted by utilizing various linguistic features.

6.7.1 Evaluation on the Effect of Word Information

Comparison Mean Opinion Score (CMOS) described in Section 2.2.2 was used for subjective evaluation of this experiment. Listeners were presented with a pair of synthesized speech samples generated by these two systems. On the trials, the synthesized speech of HMM-based model with syllable information was followed by the synthesized one of CLUSTERGEN model with word information. Listeners were asked to judge the quality of the second sample relative to the quality of the first sample. The evaluation was based on the naturalness for segmenting the correct form of words in the synthesized speech. The judgment was made on 5-point CMOS scores. The test samples included 20 random sentences and there were 21 native listeners in this evaluation. They could play samples as many times as they liked.

The results of subjective evaluation are shown in Figure 6.6. According to the chart, CMOS scores of most synthesized speeches are greater than 3 which means that the naturalness of the synthesized speech of word-based model is better than that of the syllable-based model. The CLUSTERGEN model with word information got average 4.009 score of CMOS and it shows that word information is important for promoting the naturalness of Myanmar synthesized speech.

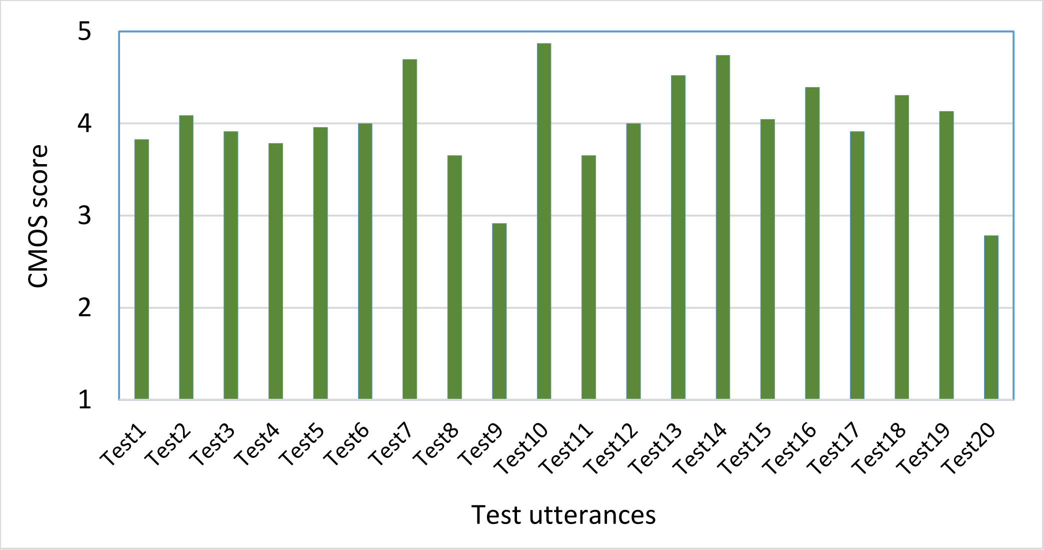


Figure 6.6 Comparison Mean Opinion Score (CMOS) on Test Utterances

6.7.2 Findings

Some issues on the synthesized speeches are found by inspecting many synthesized speeches of CLUSTERGEN model. In the pronunciation of foreign words ending with “လ်” (L) such as “အီးမေးလ်” (email), “လီဗာပူးလ်” (Liverpool), the duration of the pronunciation of “လ်” (L) in synthesized speech is longer than usual pronunciation. There are few samples pronounced with the wrong tone between normal tone (Tone1) and breathy tone (Tone3). As an example, “ပို” (/pou/) with normal tone in some utterances is synthesized wrongly as “ပိုး” (/pou:/) with breathy tone and “မုန်း” (/moun:/) with breathy tone as “မုန်” (/moun/) with normal tone.

6.7.3 Evaluation on HMM based Myanmar Speech Synthesis

According to our preliminary result of CLUSTERGEN voice, the importance of word information can be seen clearly. Therefore, HMM-based Myanmar speech synthesis with many contextual linguistic features including word information was implemented and used as the baseline system for this research. Standard five-state left-to-right Hidden Semi-Markov Models (HSMM) with no skip was used for training HMM-based system. Contextual linguistic labels extracted by text analysis part implemented in Section 5.3 were applied in modelling and a question set for Myanmar language was applied in decision tree based context clustering. Spectral envelope, fundamental frequency F_0, and duration were modeled simultaneously by the corresponding HMMs. Minimum description length (MDL) factor of 1.0 was used in decision tree state clustering. Global variance (GV) enhancement and modulation spectrum-based postfilter were applied on training HMM-based system. The publicly available HTS toolkit[1] was used to implement the HMM-based speech synthesis for Myanmar language. It achieves the objective results of **5.015** MCD in dB, **39.693** F_0 RMSE in Hz and **8.316** Voice/Unvoiced swapping error in percentage. These objective results will be used in comparison with other experimental results in next chapter.

[1] http://hts.sp.nitech.ac.jp

6.8 Summary

In this chapter, the basic structure and algorithms of HMM-based TTS system and the overview of CLUSTERGEN model are presented. Word information is used in Myanmar TTS using CLUSTERGEN and the results of subjective evaluation are reported. To sum up, word information is important for the naturalness of synthesized speeches in Myanmar language according to our preliminary results of CLUSTERGEN model. Therefore, more contextual linguistic information including word information is applied in HMM-based Myanmar speech synthesis and it will be used as the baseline system for comparing neural network based Myanmar TTS systems in next chapter.

CHAPTER 7

DEEP NEURAL NETWORK BASED MYANMAR SPEECH SYNTHESIS

In recent years, artificial neural network based acoustic model has become the state-of-the-art modelling in speech synthesis area. HMM-based speech synthesis is effective because of its flexibility in changing speaker identities, emotions, and speaking styles. However, decision tree clustered context dependent HMMs has some limitations [75]. Decision trees are inefficient to model complex context dependencies. Therefore, Deep Neural Network (DNN) has been applied in modelling the relationship between linguistic features and acoustic features and it can give better synthesized speech than HMM [40, 75].

With the purpose of achieving better synthesized speech, DNN was applied in acoustic modelling of Myanmar TTS system. In this chapter, the implementation of DNN based Myanmar speech synthesis was described in detail. The introduction of DNN with the perspective of speech synthesis was presented in the chapter. Objective and subjective evaluations are done on DNN based model by comparing the HMM based model described in Chapter 6. The effectiveness of precise state boundaries and coarse phone boundaries on aligning input linguistic features and output acoustic features for training DNN are investigated. With the purpose of exploring the more suitable method and dimension of word vectors for neural network based Myanmar TTS systems, the DNN based models are experimented with all trained word vectors (W1, W2, W3, W4, W5, and W6) reported in Section 5.4.1.3. The performance of DNN based Myanmar TTS system with different input features has also been evaluated in this chapter.

7.1 Deep Neural Network

A feed-forward artificial neural network that has more than one layer of hidden units between its input and output layers is usually called a DNN. At each hidden layer, each hidden unit typically maps the weighted sum of its inputs from the layer below using a nonlinear activation function and passes it to the layer above [30]. Figure 7.1 is the illustration of DNN with two hidden layers. If a hyperbolic tangent function *h*(.) is used as an activation function, its output is calculated as

$$h_j^{(l)} = h\left(b_j^{(l)} + \sum_i h_i^{(l-1)} w_{ij}^{(l)}\right) \qquad (7.1)$$

where $h_j^{(l)}$ is the jth hidden unit at the lth layer. $b_j^{(l)}$is the bias of the jth unit at the lth layer, and $w_{ij}^{(l)}$ is the weight associated with the link from $h_i^{(l-1)}$ to $h_j^{(l)}$. For speech synthesis task, to predict target features from the activations in the hidden layer below, a linear activation function is used

$$\tilde{y}_j = b_j^{(L+1)} + \sum_i h_i^{(L)} w_{ij}^{(L+1)} \qquad (7.2)$$

where L is the number of hidden layers. The set of parameters of an L hidden layer DNN can be optimized in a supervised way by minimizing a loss function that measures the different between reference data and predicted output using the back-propagation algorithm [45]. For regression tasks such as speech synthesis, the mean square error is commonly adopted as the loss function

$$L(y, \tilde{y}; \lambda) = \sum_j (y_j - \tilde{y}_j)^2 \qquad (7.3)$$

where y_j and $\tilde{y}_j$ are the jth dimension of the correct and predicted outputs, respectively. Minimizing the mean square error between y_j and $\tilde{y}_j$ with respect to λ is equivalent to the maximum likelihood estimation of λ.

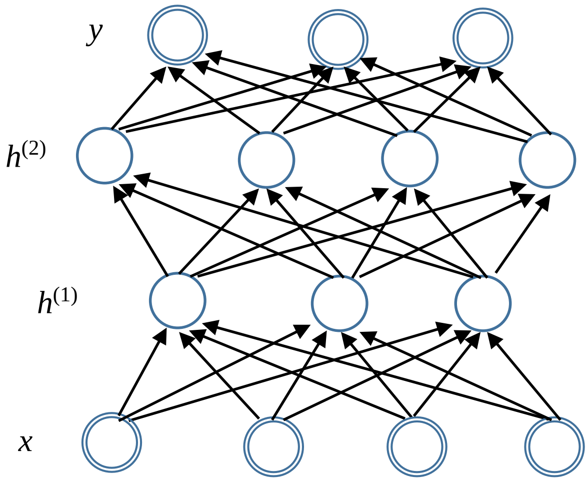


Figure 7.1 An Illustration of DNN with Two Hidden Layers

7.2 DNN based Myanmar Speech Synthesis

Figure 7.2 illustrates DNN-based speech synthesis framework with 3 hidden layers. The given input text is transformed to the contextual labels by our text-analysis part described in Section 5.3. A sequence of input features x_t is obtained by applying a question set for Myanmar language presented in Section 5.2.4. Input features are binary features for categorical contexts (e.g. is the current phoneme kh?) and numerical features (e.g. the number of syllables in the current word) for numerical contexts. The output features y_t at frame t are spectral and excitation parameters, and their dynamic features. The weights of DNN are trained by using pairs of input and output features extracted from training data.

In synthesis time, input features are extracted from the input text and then these are mapped to output features (mean and variances of speech parameter vector sequence) by the trained DNN. The speech parameter generation algorithm can output the smooth trajectories of speech parameters in accordance with the statistics of static and dynamic features. Finally, a synthesized speech waveform according to the given speech parameters is generated by the vocoder.

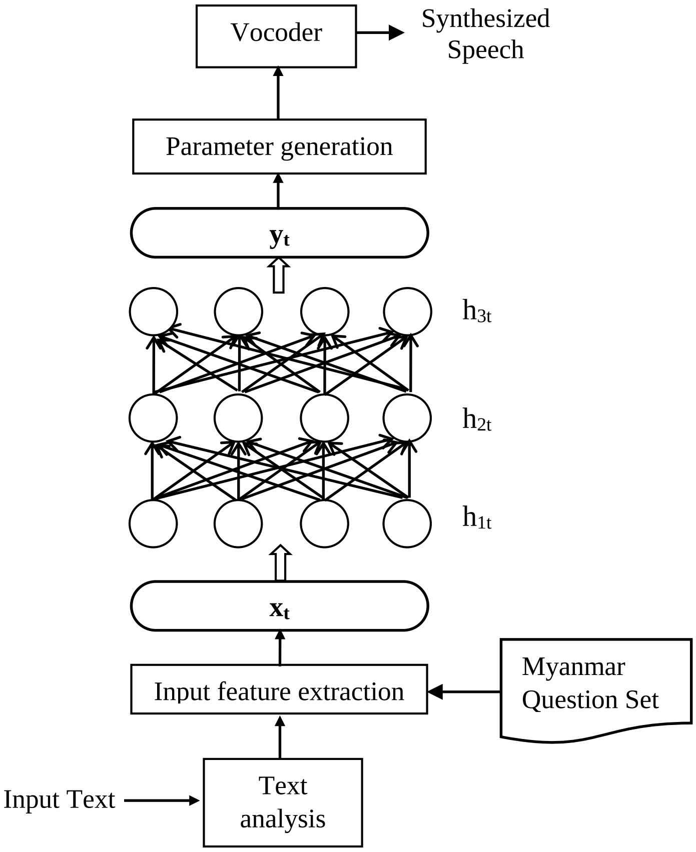


Figure 7.2 A DNN-based Speech Synthesis Framework with Three Hidden Layers, h_{ij} denotes activation at i^{th} layer at j^{th} frame

7.3 Using Word Embedding in Acoustic Modelling

Figure 7.3 illustrates the structure of neural network based acoustic model using word embedding. Getting word vector features for current word w_t at frame t is done by using the Myanmar word embedding model.

L_t is linguistic feature vector at frame t generated from the front-end of Myanmar TTS system. $C(w_t)$ is word vector representation for the word w_t. L_t and $C(w_t)$ at frame t are cascaded as input feature vector and use it in training neural network based acoustic model. O_t is the output acoustic features at frame t.

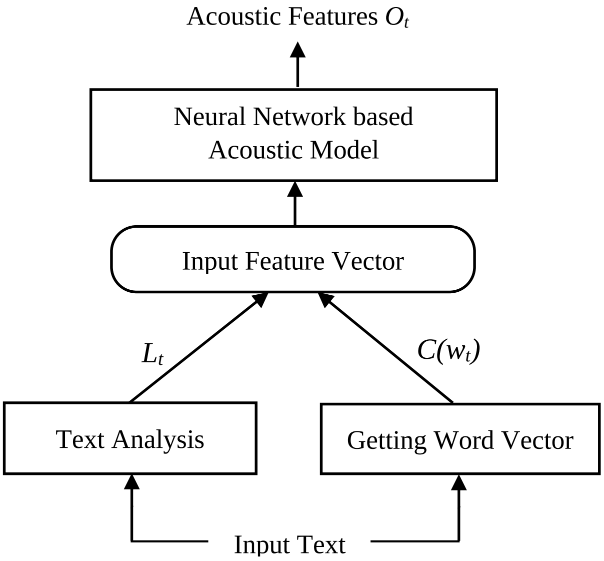


Figure 7.3 Structure of Neural Network based Acoustic Model with Word Embedding

7.4 Experiments

Experiments are done on DNN based Myanmar speech synthesis and both objective and subjective evaluations have been conducted.

7.4.1 Experimental Setups

The architecture of the DNNs was 6-hidden layers, 1024 units per layer. The tangent or tanh function was used as the hidden activation function, and a linear activation function was used at the output layer. 3,800 utterances from Myanmar speech corpus in Section 6.6 were used as the training set, 100 utterances as the development set, and 100 utterances as the evaluation set. These all sets are disjoint sets.

The input features for DNN-based systems consisted of 635 features including 619 binary features for categorical linguistic contexts (*e.g.*, phoneme identities, tone types) and 16 numeric features for numerical linguistic contexts (*e.g.* the number of syllables in the current word, the number of words in the utterance). 9 numeric features for frame related features were used in all experiments. WORLD [35] was used to extract 60-dimensional MCCs, 5-dimensional band aperiodicities (BAPs) and logarithmic fundamental frequency (log F_0) at 5 msec frame intervals. Input features were normalized using min-max to the range of [0.01, 0.99] and output features were normalized to zero mean and unit variance. Maximum likelihood parameter generation (MLPG) was applied to generate smooth parameter trajectories from DNN outputs and spectral enhancement post-filtering was applied to MCCs. Merlin speech synthesis toolkit [69] was used for modeling DNNs and training was done on Nvidia K80 GPU.

7.4.2 Evaluations of DNN-based Systems with State Level and Phone Level Alignments

The effect of alignment using state boundaries and phone boundaries between input linguistic features and output acoustic features has been investigated for training DNN. Input linguistic features and output acoustic features are generally needed to be force aligned by HMMs in advance for training DNNs. DNN using state level alignment (DNN_{st}) and DNN using phone level alignment (DNN_{ph}) were compared in this experiment. In DNN_{st}, the input linguistic features and the output acoustic features were aligned at the precise state level. For training DNN_{st}, HVite from HTK tools[1] was used to do forced alignment. For training DNN_{ph}, the input and output features were aligned at phone level and used input features to indicate the coarse boundaries in a given phone. Ergodic Hidden Markov Model (EHMM) in CLUSTERGEN [6] setup was applied for doing this forced alignment using phone level.

7.4.2.1 Objective Evaluation

Mel-cepstral distortion (MCD) in dB, F_0 distortion in root mean squared error (RMSE) and voiced/unvoiced (V/U) swapping error in percentage are used as the objective measures.

[1] http://htk.eng.cam.ac.uk/download.shtml

Table 7.1 shows the results of objective measures of HMM-based system and DNN based ones. By comparing the objective results of DNN based systems with those of HMM-based system, the DNN-based systems outperform the HMM-based one in log F_0 prediction and V/U swapping. In particular, the RMSE of $F0$ is reduced from 39.693 Hz to 31.233 Hz and V/U error rate is also reduced from 8.316% to 5.470%. On the other hand, the HMM-based system achieves a better performance in Mel-cepstrum prediction. As the comparison of DNN_{st} and DNN_{ph}, the DNN_{st} has better prediction than DNN_{ph} across all acoustic parameters. It shows that training DNN with aligned state boundaries is more efficient for generating better synthesized speech than training DNN with aligned coarse boundaries.

Table 7.1 Comparison of Objective Results on HMM-based and DNN-based Systems

Models	**MCD (dB)**	**F_0 RMSE (Hz)**	**V/U(%)**
HMM	**5.015**	39.693	8.316
DNN_{st}	5.355	**31.233**	**5.470**
DNN_{ph}	5.564	32.472	6.548

7.4.2.2 Subjective Evaluation

The performance of HMM-based system and DNN-based ones are further evaluated by subjective listening tests. Two AB preference tests were conducted to compare the performance of these systems. Twenty utterances were randomly selected for these tests and synthesized by all systems. Twenty-four native Myanmar people were participated in these preference tests. Each subject evaluated 20 pairs. The subjects can choose one of their preferences or “Neutral” if the difference between speeches generated by both systems cannot be perceived or difficult to judge which one is better. Figure 7.4 shows the preference scores of first AB listening test. This shows the speech synthesized by the DNN-based system is significantly preferred than the HMM-based system. The preference score (87%) of the DNN-based system is higher than the HMM-based system (4%). According to the second AB listening test shown in Figure 7.5, the perception difference between the DNN_{st} and DNN_{ph} is not significant. Their preference scores are 19% and 21% respectively. Synthesized speech samples can be accessed on the link[2].

[2] http://www.nlpresearch-ucsy.edu.mm/subtest.html

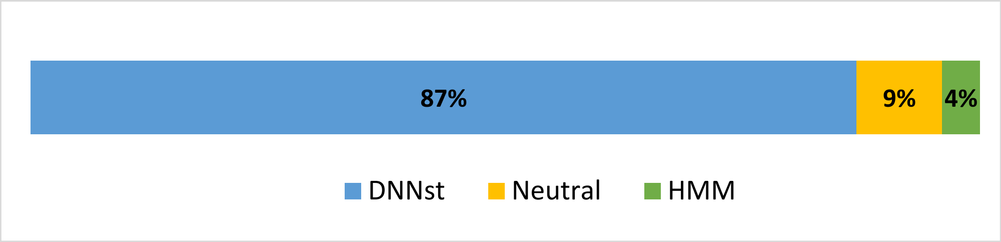


Figure 7.4 Preference Scores of HMM-based System and DNN-based System

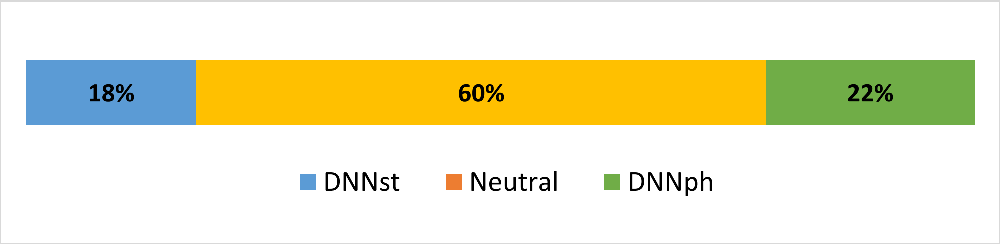


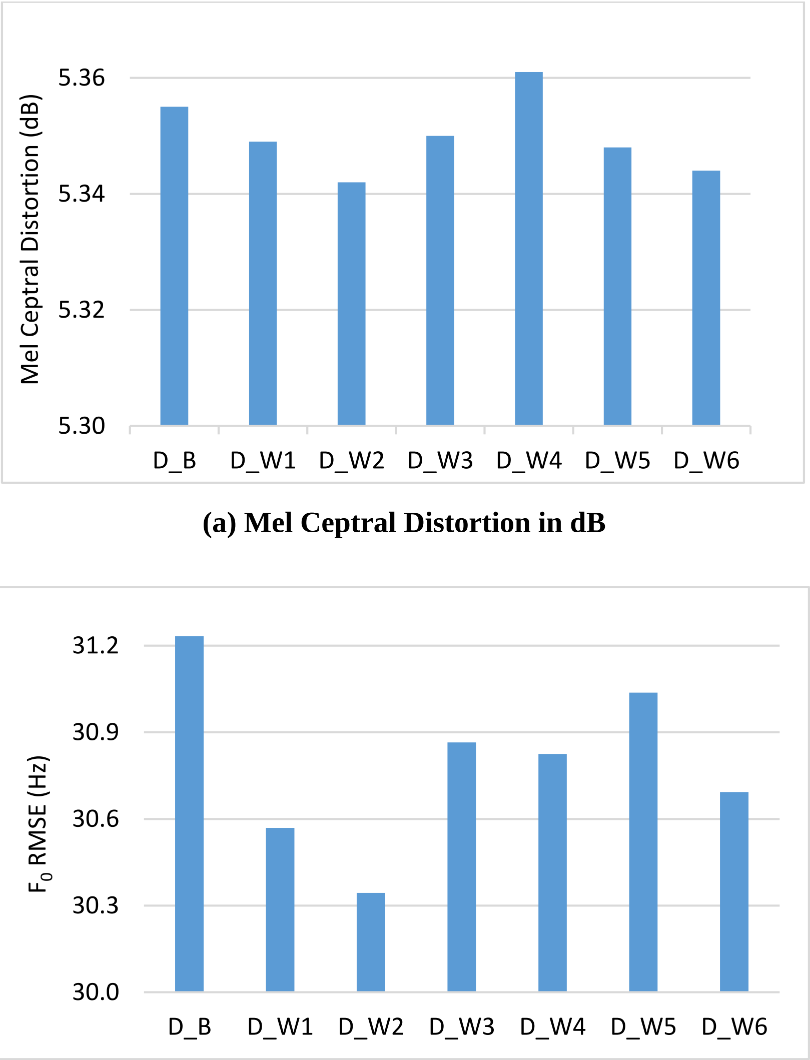
Figure 7.5 Preference Scores of Two DNN-based Systems (DNN$_{st}$ and DNN$_{ph}$)

7.4.2.3 Findings

The preference score of DNN-based synthesized speech was 83% more than that of HMM-based speech in contrary to mel-cepstral distortion of DNN based system was 0.34 higher than that of HMM-based one as the scores can be seen in Table 7.1 and Figure 7.4. There are some small noises in HMM-based synthesized speech and it might be the cause of listeners did not prefer HMM-based synthesized speech. 270 synthesized speeches of 100 from development set, 100 from test set and 70 from open internet data were inspected on all three systems, HMM, DNN$_{st}$ and DNN$_{ph}$. Types of tone and vowel errors found in syllable-based HMM are discussed in [51], but it is found that there are only about 0.45% incorrect pronunciation of normal voice (Tone 1) to breathy voice (Tone 2). And there are also 3 skipped (missing) phonemes they are occurred in the synthesized speech of DNN$_{ph}$ but it did not occur in DNN$_{st}$. The listening tests in the DNN based synthesized speech did not find incorrect vowels pronunciation, while they are occurred in HMM-based synthesized speech.

7.4.3 The Effect of Training Method and Vector Dimension of Word Vectors on DNN based Myanmar TTS System

With the purpose of exploring the more suitable method and dimension of word vectors for neural network based Myanmar TTS systems, all trained word vectors (W1, W2, W3, W4, W5, and W6) reported in Section 5.4.1.3 are used as the additional input features to the system. The term “D” is denoted as the DNN based TTS system, “D_B” refers the DNN based acoustic model with state level alignments, and “D_W#” as the DNN based TTS system with additional input word vector W#, the alias of trained word vector. Figure 7.6 depicts the objective results of DNN based Myanmar TTS system with different word vectors. As shown in Figure 7.6 (a) and (b), the better prediction of Mel Ceptrum and F_0 can be achieved by applying word vector “W2” as the additional input features to the DNN based system. According to the objective results of this experiment, the word vector “W2”, the CBOW model with dimension 200, is selected to utilize as the additional input to all neural network based Myanmar TTS systems for further experiments.



(a) Mel Ceptral Distortion in dB

(b) F_0 RMSE in Hz

Figure 7.6 Objective Results of DNN based Systems with Different Word Vectors

7.4.4 Evaluations of DNN based Myanmar TTS Systems with Different Input Features

The effectiveness of the word embedding features in DNN based Myanmar TTS systems are investigated. Besides, the contribution of word vector features are compared with POS features in DNN based systems. The effectiveness of word vector features, POS features, and combination of these two features in acoustic modelling of Myanmar TTS systems was analysed by applying these features as the additional input features to the linguistic features extracted by text analysis part.

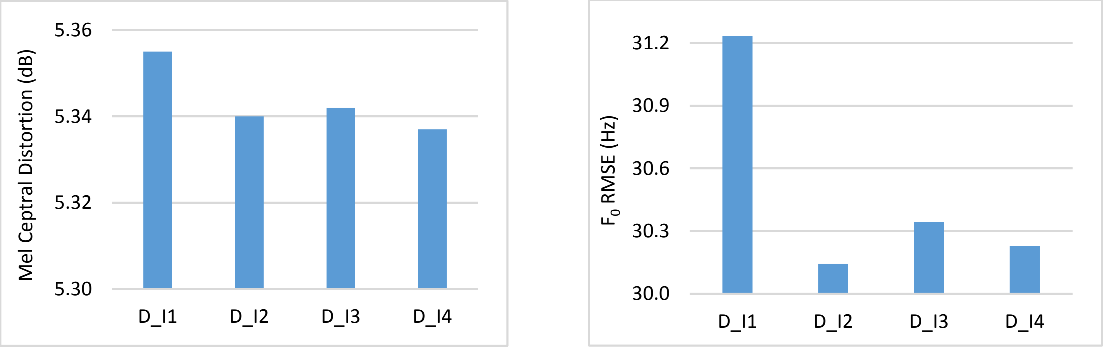
Myanmar pronunciation dictionary [21] was used in Festival for extracting the POS information of current, two preceding and succeeding words. The POS tags of some words in the training corpus are prepared manually because some are not included in the pronunciation dictionary.

Four sets of different input feature vectors defined as I1, I2, I3, and I4 are used in training DNN based acoustic models and the detail information can be seen as follows:

1) I1 : contextual linguistic features generated by text analysis part
2) I2 : I1 and POS features
3) I3 : I1 and word vector features obtained from "W2"
4) I4 : I1 and both POS features, and word vector features obtained from "W2"

The numbers of input features in I1, I2, I3, and I4 are 635, 674, 835, and 874 respectively.

7.4.4.1 Objective Evaluation



(a) Mel Ceptral Distortion in dB **(b) *F_0* RMSE in Hz**

Figure 7.7 Objective Results of DNN based Myanmar TTS Systems with Different Input Features

"D_I#" is denoted as the short forms of DNN based Myanmar TTS systems with input features I# and the objective results of these systems are shown in Figure 7.7. According to the results shown in Figure 7.7 (a) and (b), D_I2, D_I3, and D_I4 give the better results than D_I1. As the comparison of MCD in Figure 7.7 (a), D_I4 gets the best result among all. Using additional POS features and/or word vector features give the better prediction on the DNN-based system. It means that word vector can encode useful information for acoustic modelling in DNN based system.

7.4.4.2 Subjective Evaluation

MUSHRA listening tests were conducted to subjectively evaluate the effectiveness of word vectors on DNN based Myanmar TTS systems. For each test, 22 non-expert native Myanmar speakers of age range from 25 to 45 years were participated. The subjects were instructed to listen the speech samples generated by four systems with different input features and rate them using 0-100 scale on their naturalness. 20 synthetic speeches were randomly ordered and presented without labels in the tests.

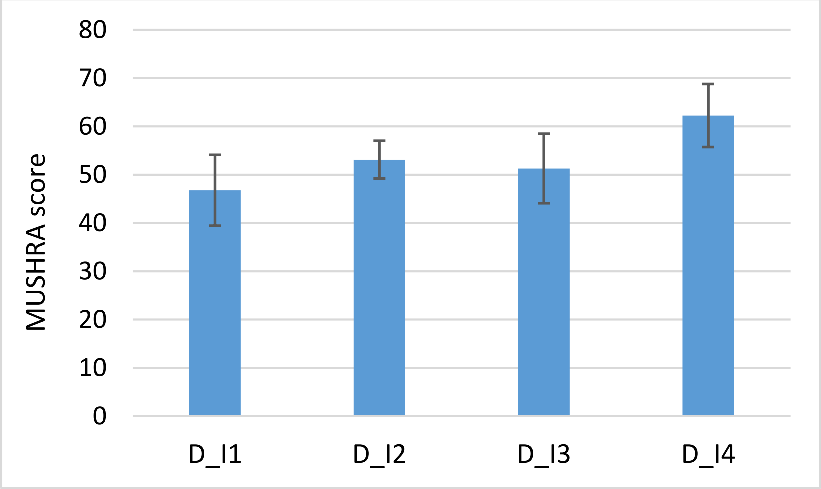


Figure 7.8 MUSHRA Scores for DNN based Myanmar TTS Systems with Different Input Features

Figure 7.8 depicts the MUSHRA results for DNN based system with different input features. It shows that word embedding or POS features can improve the naturalness of the synthesized speeches on DNN based system because the scores of D_I2, D_I3, and D_I4 are higher than that of D_I1. The system using both word embedding and POS features in addition to conventional input features achieves the highest score among all systems. The subjective results are consistent with the objective results of DNN based systems shown in Figure 7.7. It can be concluded that the effectiveness of word vector can be seen in DNN based systems.

7.5 Summary

According to the experimental results, DNN based Myanmar TTS system outperforms the HMM based Myanmar TTS system in terms of naturalness and it has been used as the baseline in the next chapter. The state level alignment is more effective than the phone level alignment in training neural network based speech synthesis. The word vector “W2”, the most suitable word vector for Myanmar speech synthesis according to our experiments, will be utilized as the additional input to all neural network based Myanmar TTS systems for further experiments. The objective and subjective results show that word embedding or POS features can improve the naturalness of the synthesized speeches on DNN based systems.

CHAPTER 8

LONG SHORT-TERM MEMORY RECURRENT NEURAL NETWORK BASED MYANMAR SPEECH SYNTHESIS

Recently, Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) has become the state-of-the-art modelling technique for a variety of machine learning problems. LSTM-RNN architecture capable of learning long-term temporal dependencies [18]. In SPSS, DNN-based acoustic models give an efficient and distributed representation of complex dependencies between linguistic and acoustic features [40, 75]. For Myanmar speech synthesis, the experimental results in Chapter 7 shows that DNN-based acoustic model achieved better synthesized speeches than HMM-based one. However, DNN-based acoustic modelling assumes that each frame is sampled independently, although certainly there are correlations between consecutive frames in speech [75]. Recurrent Neural Networks (RNNs) were applied for modeling sequential data which have correlations between consecutive frames in speech. Unfortunately, standard RNN has a limitation of the range of context. The problem is that the influence of a given input on the hidden layer either decays or blows up exponentially around the network's recurrent connections [17]. To overcome this vanishing gradient problem, LSTM-RNN architecture is designed to model temporal sequences and their long-term dependencies [22].

In recent years, LSTM-RNN has been applied to acoustic modelling for SPSS [13, 67, 76] and these studies demonstrated that LSTM-RNN can achieve significantly better performance on SPSS than DNN. Therefore, analyzing and experimenting LSTM-RNN are done whether it can get more natural synthetic speech for Myanmar TTS system. Applying LSTM-RNN in acoustic modelling of Myanmar speech synthesis with linguistic features described in Chapter 5 is the proposed system in this research. To the best of our knowledge, this is the first work to apply LSTM-RNN architecture in Myanmar speech synthesis.

This chapter presents a brief introduction of LSTM-RNN architecture used in this thesis. The comparisons of LSTM-RNN architectures for Myanmar speech synthesis were experimented. DNN-based acoustic model implemented in Chapter 7 was used as the baseline system. The importance of contextual linguistic features and the effect of applying explicit tone information in different architectures of LSTM-RNN was examined using the proposed Myanmar question set described in Section 5.2.4. Moreover, the effectiveness of applying word vectors as the additional input features are investigated on LSTM-RNN based Myanmar speech synthesis in this chapter.

8.1 Long Short-Term Memory

The LSTM-RNN is an architecture well-suited to speech processing tasks requiring the use of long range contextual information. In a DNN-based speech synthesis, the sequential nature of speech is ignored at mapping between the input contextual linguistic features and the output acoustic features. LSTM-RNN is designed to model learning long time-dependencies. Figure 8.1 shows an illustration of a standard LSTM unit. It features input gate, forget gate, output gate, block input, a memory cell (the Constant Error Carousel), an output activation function, and peephole connections [18]. The input, output, and forget gates provide continuous analogues of write, read and reset operations for the cell. The output of the block is recurrently connected back to the block input and all the gates [19]. The main idea in LSTM architecture is a memory cell that can maintain its state over time, and non-linear gating units that regulate the information flow into and out of the cell. Peephole connections provide feedback from the cell to the gates, allowing the gates to carry out their operations as a function for both the incoming inputs and the previous state of the cell. The vector formulas for a vanilla LSTM layer forward pass is formulated as follows:

Input Gate:

$$i_t = \delta(W^i x_t + R^i h_{t-1} + p^i \odot c_{t-1} + b^i) \tag{8.1}$$

Forget Gate:

$$f_t = \delta(W^f x_t + R^f h_{t-1} + p^f \odot c_{t-1} + b^f) \tag{8.2}$$

Memory Cell:

$$c_t = f_t \odot c_{t-1} + i_t \odot g(W^c x_t + R^c h_{t-1} + b^c) \tag{8.3}$$

Output Gate:

$$o_t = \delta(W^o x_t + R^o h_{t-1} + p^o \odot c_t + b_o) \tag{8.4}$$

Cell Output:

$$h_t = o_t \odot g(c_t) \tag{8.5}$$

where h_t is the hidden activation at time t and x_t is the input vector at time t. $W^*$and R^* are weight matrices from input to gate and from hidden to hidden, respectively. p^* and b^* are peephole connections and biases, respectively. $\delta(.)$ and $g(.)$ are point-wise non-linear activation functions. The logistic sigmoid ($\delta(\mathrm{x}) = \frac{1}{1+e^{-x}}$) is used as gate activation function and the hyperbolic tangent ($\mathrm{g(x)} = \tanh(\mathrm{x})$) is usually used as output activation function. $\odot$ means point-wise multiplication of two vectors.

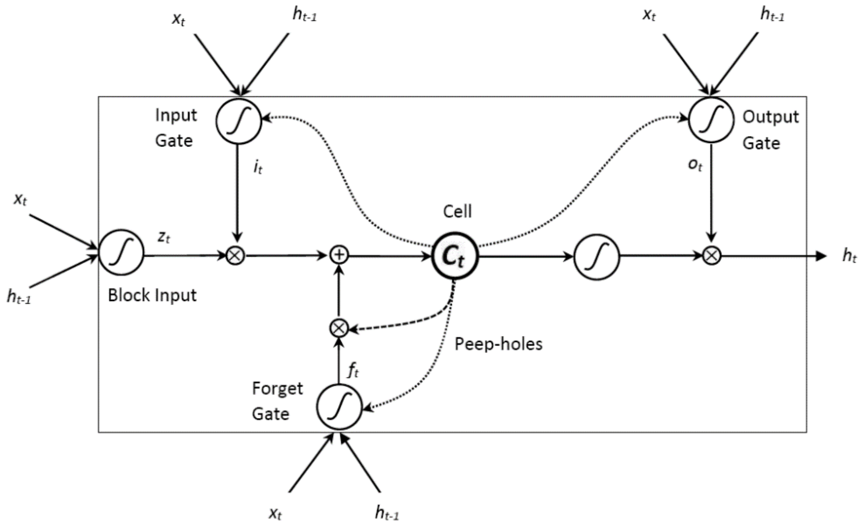


Figure 8.1 An Illustration of a Long Short-Term Memory Unit [69]

8.2 LSTM-RNN based Myanmar Speech Synthesis

Figure 8.2 illustrates the proposed architecture of LSTM-RNN based Myanmar TTS system. At the training time, such contextual labels depicted in Section 5.2.3 are extracted from the text corpus by text analysis part described in Section 5.3. Linguistic features are extracted from these contextual labels by applying our proposed Myanmar question set. Input features include binary features for categorical contexts (e.g. phoneme identity, tone type of the syllable) and numerical features for numerical contexts such as the number of syllables within the current word in forward and backward directions. Acoustic features like Mel-Cepstral Coefficient (MCCs), logarithmic fundamental frequencies (log F_0) are extracted from speech corpus. For training LSTM-RNNs, input features and output acoustic features can be force aligned frame-by-frame by HMMs in advance. The weights of LSTM-RNN are initialized randomly and then they are updated to minimize the mean squared error between the target features and predicted output features.

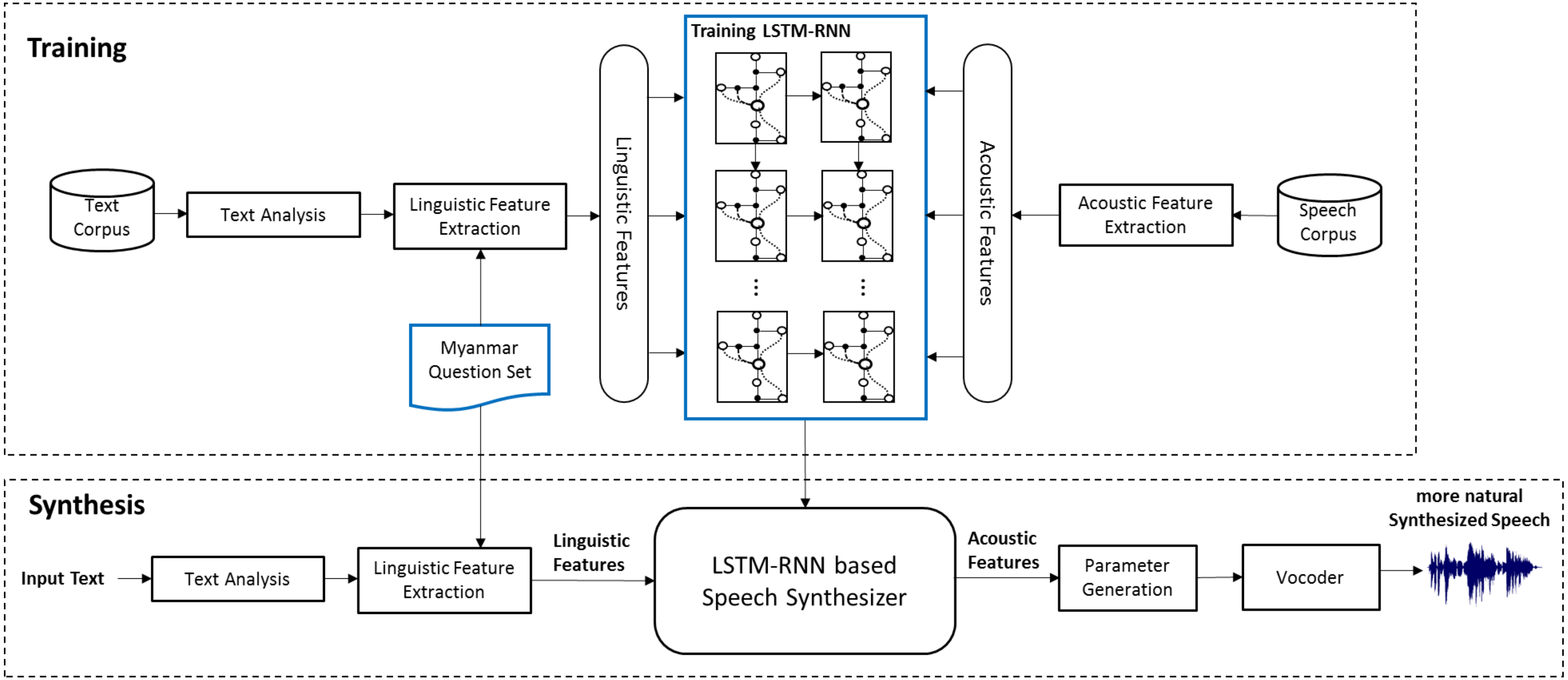


Figure 8.2 The Proposed Architecture of LSTM-RNN based Myanmar Text-to-Speech System

At the synthesis time, input Myanmar text is converted into the contextual labels and linguistic features are extracted from these labels by using the same Myanmar question set. These linguistic features will be mapped to output acoustic features by the trained LSTM-RNN based speech synthesizer. The output acoustic features are used with the speech parameter generation algorithm. Finally, the vocoder outputs a synthesized waveform according to the given speech parameters.

8.3 Experiments

Experiments are conducted to evaluate the performance of LSTM-RNN based Myanmar speech synthesis and to find the more suitable network architecture for Myanmar TTS system.

8.3.1 Experimental Setups

The same training, development and test sets in the experiment of DNN-based acoustic modelling in Chapter 7 were used in this LSTM-RNN based acoustic modelling for Myanmar language. The proposed question set was used for extracting input linguistic features for Myanmar language. WORLD [35] vocoder was used to extract 60-dimensional MCCs, 5-dimentional BAPs, and log F_0 at 5 msec frame step. A binary voiced/unvoiced feature was used for voicing information. Input linguistic features were min-max normalized to the range of [0.01, 0.99], and acoustic features were mean-variance normalized before training. MLPG was applied to generate smooth parameter trajectories at the synthesis time. Merlin speech synthesis toolkit [69] with Keras [9] python library was applied for modeling all systems on K80 GPU for training. DNN-based speech synthesis was used as the baseline in the experiments. The following network architectures of speech synthesis systems were used in our experiments:

1) DNN : a baseline system with six feedforward hidden layers of 1024 hyperbolic tangent units each
2) LSTM-1L : a single hidden layer with LSTM-RNN (512 units)
3) LSTM-2L : two hidden layers with LSTM-RNN (512 units each)

4) Hybrid-LSTM-1L : a hybrid of DNN and LSTM-RNN, five feedforward hidden layers of 1024 hyperbolic tangent units each, followed by a single LSTM-RNN layer with 512 units
5) Hybrid-LSTM-2L : a hybrid of DNN and LSTM-RNN, four feedforward hidden layers of 1024 hyperbolic tangent units each, followed by two LSTM-RNN layers with 512 units each

According to the preliminary results, it is found that LSTM-RNN hidden layers with 512 units gave better objective results than that with 256 and 1024 units. Therefore, LSTM-RNN hidden layers with 512 units have been used in all experiments. Silence frames were not taken into account in training for avoiding overlearning silence labels in acoustic modeling. The weights of all LSTM-RNNs were initialized randomly and then they were updated to minimize mean squared error (mse) between target and predicted output features. Stochastic gradient descent (sgd) based learning rate scheduling was used for all hybrid systems and Adam optimizer [28] was used for LTSM-1L and LSTM-2L. Exact LSTM gradient with untrancated Backpropagation Through Time (BPTT) [17] was applied for training LSTM-RNNs. All systems were trained with batch size of 25 sentences. Hyperparameters for each system were optimized on the development set. Fixed momentum was used and learning rates were tuned in these systems. A linear activation function was applied at the output layer for all systems.

8.3.2 Evaluation on the Effect of Contextual Linguistic Information

The effect of contextual linguistic information on all LSTM-RNN architectures were analyzed. As the LSTM-RNNs can get the past contextual information through their recurrent connections, the effect of preceding two contextual information on modeling all LSTM-RNN based Myanmar speech synthesis systems was experimented. Figure 8.3 and 8.4 depict the comparisons of MCD and F_0 RMSE using C_635 and C_423 on all LSTM-RNN architectures for Myanmar speech synthesis respectively. C_635 refers 635 input linguistic features including current context, and preceding and succeeding two contexts at phoneme, syllable, word, and utterance levels. C_423 refers 423 input linguistic features including only current context, and succeeding two contexts at these levels. In this case, tone information is also included in contextual linguistic features of C_635 and C_423. 9 numeric features for frame related features are also used for all experiments. C_635 and C_423 are extracted by applying the proposed Myanmar question set. Figure 8.3 shows that applying C_635 on all architectures gets better prediction on Mel-Cepstrum than applying C_423. In Figure 8.4, all architectures applied C_635 except LSTM-1L get better F_0 RMSE than that applied C_423. These objective results confirm that preceding contextual information is still important for modeling LSTM-RNN based speech synthesis.

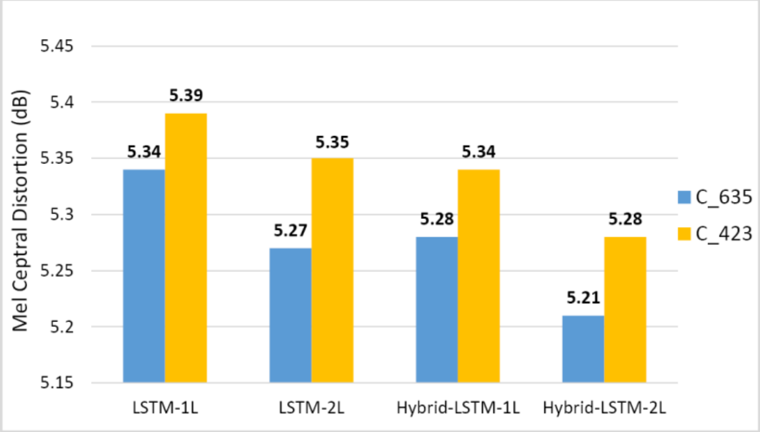


Figure 8.3 Effect of Left Contextual Information on MCD

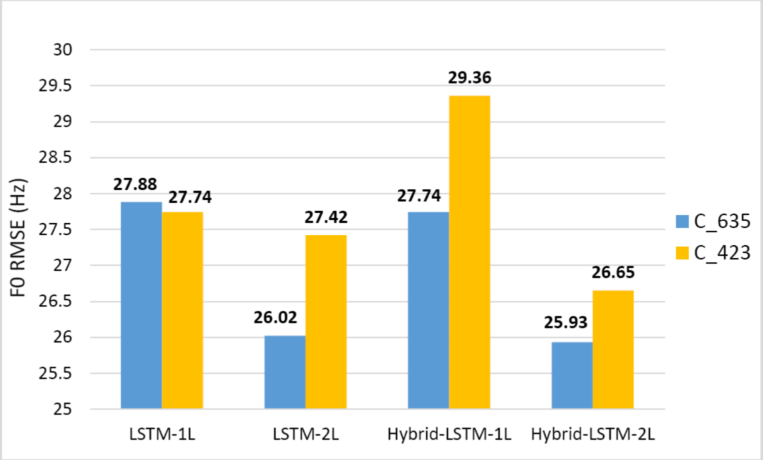


Figure 8.4 Effect of Left Contextual Information on F_0

8.3.3 Evaluation on the Effect of Explicit Tone Questions in Myanmar Question Set

Though tone information is already included in the grapheme of vowels in Myanmar language, explicit tone information was added in the input linguistic features by applying questions about tone types of vowels in Myanmar question set. Comparisons of tone information and no tone information on modeling LSTM-RNN based speech synthesis were experimented. In this experiment, all the systems with tone information use C_635 input features. Using explicit tone information in modeling Myanmar speech synthesis gives better MCD on all network architectures in the experiments according to Figure 8.5. As shown in Figure 8.6, all architectures modeling with explicit tone information except LSTM-1L get better F_0 RMSE than no explicit tone information. In general, it can be concluded that explicit tone questions in Myanmar question set are useful for modeling Myanmar speech synthesis.

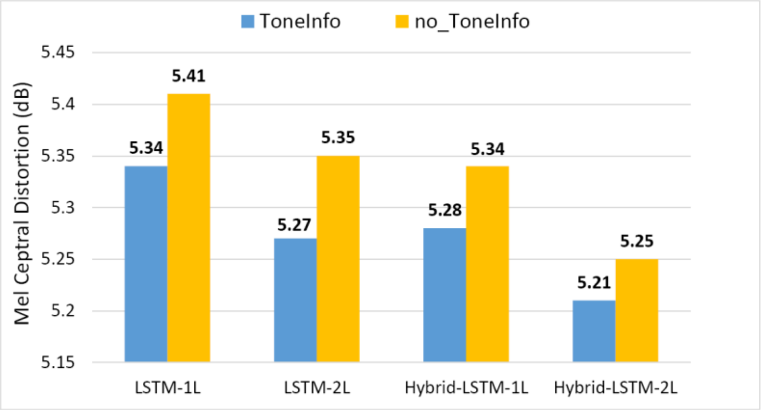


Figure 8.5 Effect of Explicit Tone Information on MCD

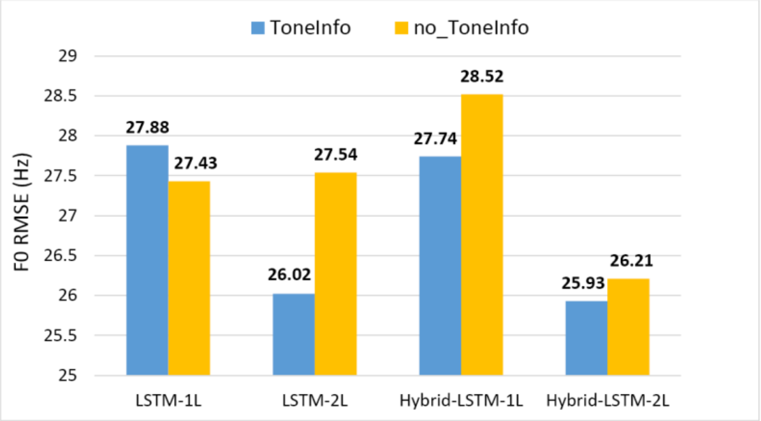


Figure 8.6 Effect of Explicit Tone Information on F_0

8.3.4 Performance of Myanmar Speech Synthesis with Different Network Architectures

Experiments of Myanmar speech synthesis with different network architectures are conducted by applying C_635 for contextual linguistic features and 9 numerical features for frame related features. The objective and subjective evaluations are done for finding the most suitable network architecture for Myanmar speech synthesis.

8.3.4.1 Objective Evaluation

Table 8.1 presents the objective results of different network architectures for Myanmar speech synthesis. It is observed that all LSTM-RNN based speech synthesis systems achieve better objective results than the baseline DNN except BAP distortion of LSTM-1L and Hybrid-LSTM-1L. It shows that LSTM-2L objectively outperforms LSTM-1L across all objective measures, and Hybrid-LSTM-2L gets better objective results than Hybrid-LSTM-1L in terms of MCD, BAP, and F_0 RMSE. These results confirm that two hidden layers of LSTM-RNNs can give better performance over single hidden layer of LSTM-RNN. In particular, MCD of Hybrid-LSTM-2L architecture decreases 0.15(dB) from that of the baseline DNN, and F_0 RMSE of Hybrid-LSTM-2L 25.93(Hz) is significantly better than that of DNN 31.23(Hz). Hybrid-LSTM-2L is the best network architecture for Myanmar speech synthesis in our experiments.

Table 8.1: Objective Evaluation Results of Different Network Architectures for Myanmar Speech Synthesis

Models	**MCD (dB)**	**BAP (dB)**	**F_0 RMSE (Hz)**	**V/U (%)**
DNN (baseline)	5.36	0.21	31.23	5.47
LSTM-1L	5.34	0.21	27.88	5.31
LSTM-2L	5.27	0.20	26.02	5.26
Hybrid-LSTM-1L	5.28	0.21	27.74	**5.06**
Hybrid-LSTM-2L	**5.21**	**0.20**	**25.93**	5.16

8.3.4.2 Subjective Evaluation

The performance of DNN, LSTM-2L, and Hybrid-LSTM-2L systems was subjectively evaluated by perceptual tests. Thirty utterances were randomly selected from the evaluation set and open domain, internet data. These utterances were synthesized by the baseline DNN, LSTM-2L, and Hybrid-LSTM-2L systems. Three AB preference tests (DNN vs. LSTM-2L, DNN vs. Hybrid-LSTM-2L, and LSTM-2L vs. Hybrid-LSTM-2L) were participated by 20 non-expert native speakers of age range from 20 to 40 years. The synthetic speeches were presented in random order in each pair of all three tests. Subjects were given 30 pairs of synthesized speeches and asked to choose the more natural one in each pair or "Neutral" if the difference between two speech samples cannot be perceived. The scores of three AB preference tests with 95%

confidence intervals are presented in Figure 8.7, 8.8, and 8.9. The higher preference scores on LSTM-2L and Hybrid-LSTM-2L over the baseline DNN can also be seen clearly in the Figure 8.7 and 8.8. They confirm that LSTM-RNN based systems can generate more natural synthesized speech than DNN based system. Again, the two LSTM-RNN based systems are compared in Figure 8.9 by the preference score and here, the performance of Hybrid-LSTM-2L is obviously preferred over LSTM-2L by the native listeners. According to the three preference tests, it can be concluded that the naturalness of Hybrid-LSTM-2L system is highly preferred than that of DNN and LSTM-2L.

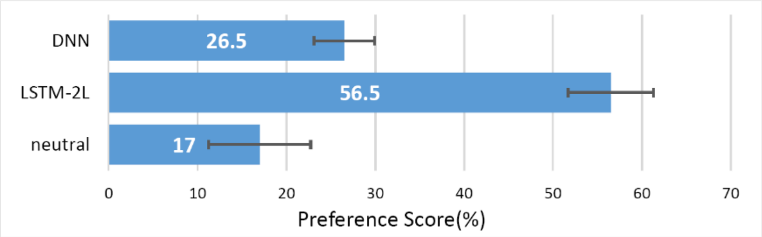


Figure 8.7 Preference Scores of DNN vs. LSTM-2L with 95% Confidence Intervals

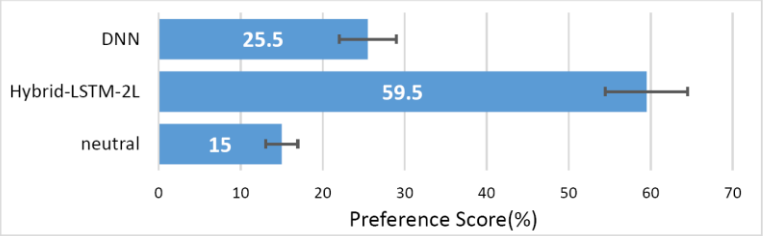


Figure 8.8 Preference Scores of DNN vs. Hybrid-LSTM-2L with 95% Confidence Intervals

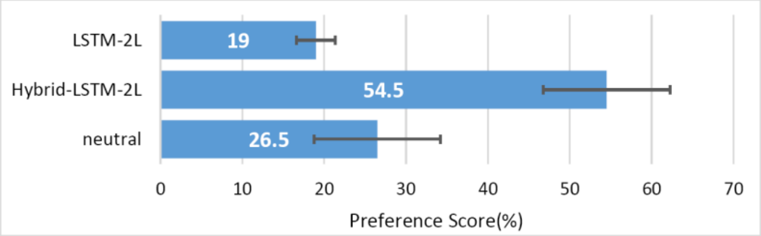


Figure 8.9 Preference Scores of LSTM-2L vs. Hybrid-LSTM-2L with 95% Confidence Intervals

The naturalness of the synthesized speeches generated by DNN, LSTM-2L, and Hybrid-LSTM-2L systems were further evaluated in terms of Mean Opinion Score

(MOS). The same 20 subjects from AB preference tests were also used in the MOS test. It is the subject to rate the naturalness of synthesized speeches on a scale from 1 to 5 where 1 is bad and 5 is excellent. The scores of DNN, LSTM-2L, and Hybrid-LSTM-2L are shown in Figure 8.10 with 95% confidence intervals of MOS results by the error bars. The LSTM-RNN based systems give higher MOS scores than the baseline DNN, and the Hybrid-LSTM-2L has the best result among all. Some samples of synthesized speeches generated by these systems are available for listening on here[1]. All AB preference tests and MOS test confirmed that LSTM-RNN based systems offer better performance than the baseline DNN, and furthermore, Hybrid-LSTM-2L outperform both DNN and LSTM-2L in terms of naturalness. It can be observed that the preference on Hybrid-LSTM-2L achieved the highest score not only in terms of objective but also subjective evaluation.

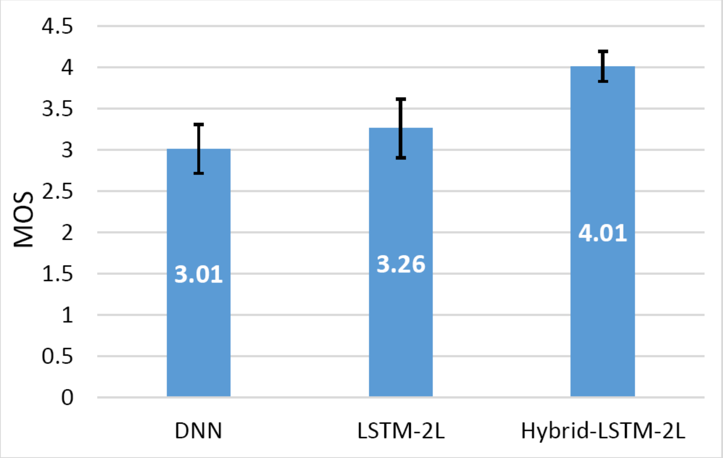


Figure 8.10 Mean Opinion Scores (MOS) with 95% Confidence Intervals of DNN, LSTM-2L, and Hybrid-LSTM-2L

8.3.4.3 Findings

It can be noticed that though LSTM-2L and Hybrid-LSTM-2L have only a slight difference in objective results, their subjective scores are notably different. In particular, the difference of MCD between two systems is only 0.06(dB) and the difference of F_0 RMSE is only 0.09(Hz). However, the difference of MOS results between two systems (0.75) is relatively high. The occurrence of breath pauses insertion in wrong places in LSTM-2L is more than that of DNN and Hybrid-LSTM-2L, made LSTM-2L to be less

[1] http://www.nlpresearch-ucsy.edu.mm/subeval.html

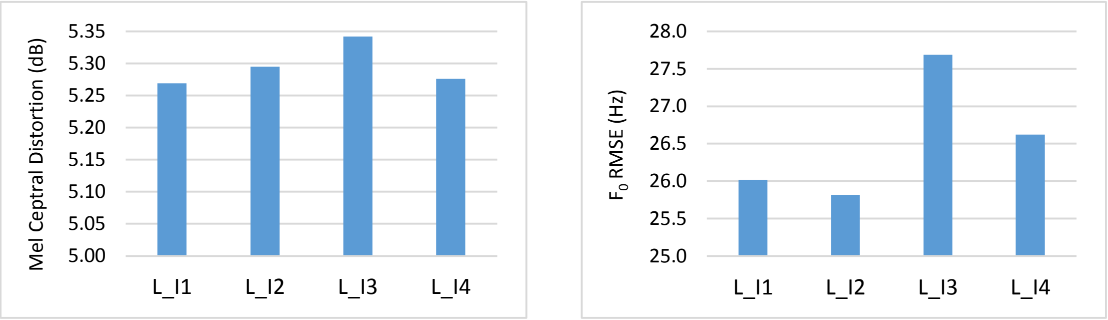
preferred by the listeners. 270 synthesized speeches (100 each from development and evaluation sets, and 70 from open internet data) were inspected on DNN, LSTM-2L, and Hybrid-LSTM-2L systems. It is found that LSTM-RNN based speech synthesis can reduce half of incorrect pronunciation of tones over DNN based speech synthesis. Better prediction of F_0 by LSTM-RNN contributed to the more natural synthesized speech of Myanmar speech synthesis in addition to better prediction of other factors (MCD, BAP, V/UV).

8.3.5 Performance of LSTM-RNN based Myanmar TTS Systems with Different Input Features

The effect of word embedding features in LSTM-RNN based Myanmar TTS systems are examined by applying four sets of different input feature vectors, I1, I2, I3, and I4 used in DNN based systems reported in Section 7.4.4. Both objective and subjective evaluations are done on LSTM-2L and Hybrid-LSTM-2L models which have better performance for Myanmar speech synthesis according to our experiments in Section 8.3.4. The terms "L_I#", and "HL_I#" were used as the short forms of LSTM-2L and Hybrid-LSTM-2L based Myanmar TTS systems with input features I# respectively.

8.3.5.1 Objective Evaluation

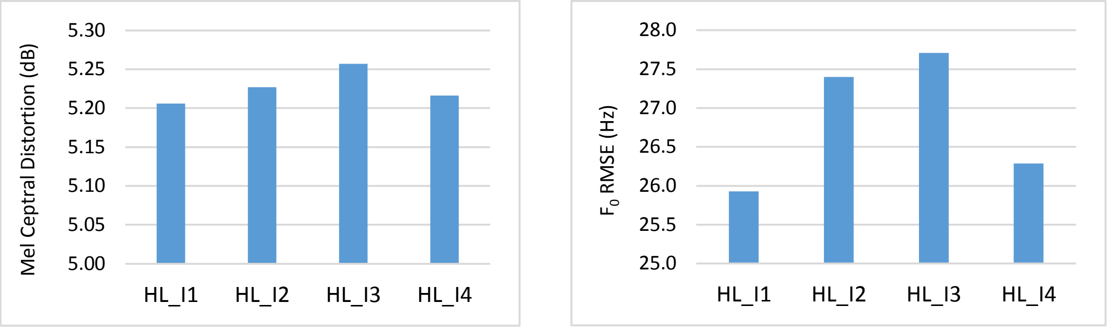
The objective results of LSTM-2L and Hybrid-LSTM-2L speech synthesis models with different input features are shown in Figure 8.11 and 8.12, respectively. The L_I1 and HL_I1 models are same as LSTM-2L and Hybrid-LSTM-2L models shown in Table 8.1.



(a) Mel Ceptral Distortion in dB **(b) F_0 RMSE in Hz**

Figure 8.11 Objective Results of LSTM-2L Systems with Different Input Features

As we can see in Figure 8.11, there is no improvement in prediction of Mel Spectrum by using POS and/or word vector features, though little improvement can be seen on F_0 prediction by using POS information in LSTM-2L systems.



(a) Mel Ceptral Distortion in dB **(b) F_0 RMSE in Hz**

Figure 8.12 Objective Results of Hybrid-LSTM-2L Systems with Different Input Features

In the case of Hybrid-LSTM-2L based systems, POS or word embedding features cannot give further improvement as shown in Figure 8.12. According to the results of HL_I4 in Figure 8.12 over that of HL_I2 and HL_I3 systems, using both features in acoustic modelling can slightly improve the performance of the system. Meanwhile, HL_I1 achieves the best performance in prediction of Mel Cepstrum and F_0 among all systems, it means linguistic features generated by text analysis are good enough for hybrid systems.

According to the objective results, we can conclude that the effect of word vectors can be seen clearly in DNN based systems in Chapter 7 and any significant improvement cannot be found in LSTM-RNN based systems, particularly LSTM-2L and Hybrid-LSTM-2L systems. This is consistent with the results in [63] though this is different with [65].

Word vector features may not be effective in acoustic modelling for LSTM-RNN based systems. The reason may be that the word vectors were trained without any acoustic clues or prosodic knowledge and the better hidden representations computed by the recurrent connections in LSTM-RNN [63].

It can be observed that POS information was less useful for the LSTM-RNN based system. It might be the fact that suprasegmental information is already got from contextual linguistic features (such as positional features in contextual labels) and it might be the noise in POS tags due to the lack of high accuracy POS tagger for Myanmar language.

In comparing the results of the best MCD and F_0 of DNN based system reported in Figure 7.7 and the best MCD and F_0 of Hybrid-LSTM-2L system reported in Figure 8.12, we can conclude that LSRM-RNN based system without any additional features such as POS features or word vector features perform better than the DNN based system with POS and word vector features. The best MCD of DNN based systems is 5.337 and the best F_0 RMSE is 30.143 while the best MCD of Hybrid-LSTM-2L based systems is 5.206 and the best F_0 RMSE is 25.929. It can be observed that linguistic features generated by text analysis are good enough for Myanmar TTS system with LSTM-RNN based acoustic modelling.

8.3.5.2 Subjective Evaluation

To subjectively evaluate the effectiveness of word vectors on Hybrid-LSTM-2L systems, MUSHRA listening test with the same setting in Section 7.4.4.2 was conducted. The MUSHRA scores of Hybrid-LSTM-2L systems applying different input feature vectors are illustrated in Figure 8.13. According to the results, the scores of the systems including word embedding and/or POS features are slightly higher than the system with conventional input features.

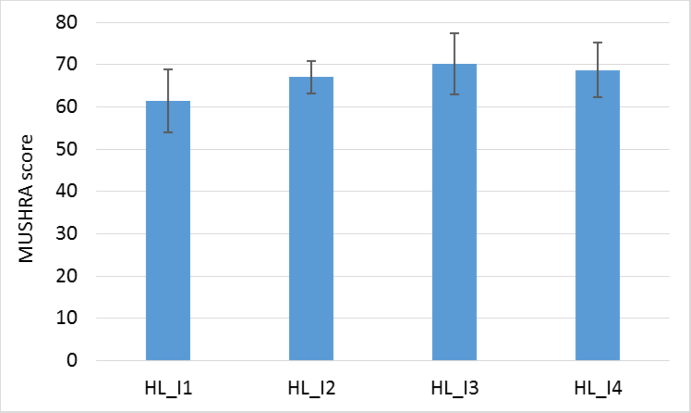


Figure 8.13 MUSHRA Scores for Hybrid-LSTM-2L Systems with Different Input Features

In particular, the user preferences on HL_I2, HL_I3 and HL_I4 systems are slightly higher than the HL_I1. It can be seen that word embedding features and/or POS features give small improvement in the perception of native listeners though any improvement is not found in objective results of Hybrid-LSTM-2L systems in Figure 8.12. Some samples of synthesized speeches generated by DNN and Hybrid-LSTM-2L based Myanmar speech synthesis applying different input features are given on the link[2].

8.4 Summary

In this chapter, our proposed architecture of LSTM-RNN based Myanmar TTS system was presented in detail. The effect of contextual linguistic features extracted by using proposed Myanmar question set on LSTM-RNN based speech synthesis was explored and it shows that the preceding contextual information and explicit tone information are still important for modelling LSTM-RNN based speech synthesis though it has the ability of accessing past information through their recurrent connections. Both objective and subjective results confirm that Hybrid-LSTM-2L (the hybrid of DNN and LSTM-RNN) system offers more suitable network architecture for Myanmar speech synthesis in naturalness. It can be observed that word embedding features and/or POS features can give little improvement in the subjective results though they cannot lead to any improvement in objective results of LSTM-RNN based systems. Therefore, it can be concluded that linguistic features generated by text analysis part and proposed Myanmar question set presented in Chapter 5 are good enough for LSTM-RNN based Myanmar TTS system. It can be proved that the naturalness of Myanmar TTS system can be promoted by applying linguistic information on LSTM-RNN based acoustic modelling.

[2] http://www.nlpresearch-ucsy.edu.mm/subeval-wv.html

CHAPTER 9
CONCLUSION AND FUTURE WORK

This thesis reports the research results of LSTM-RNN based acoustic modelling with contextual linguistic information extracted by our text analysis part for enhancing Myanmar TTS system in naturalness. All six contributions in the text analysis part and acoustic modelling part in SPSS meet the objectives of this dissertation described in Chapter 1.

To promote the quality of Myanmar TTS system, NSWs with number are firstly normalized into their standard words in the initial step of text analysis. Myanmar number normalization has been implemented by applying the identified semiotic classes and WFSTs in Chapter 3. The performance was evaluated in terms of WER and the evaluation results show that it can get acceptable results and can be used practically in text analysis part of Myanmar TTS system. The good pronunciation of Myanmar numbers have been achieved in Myanmar TTS system by applying this WFSTs based Myanmar number normalization module.

The first large amount of pronunciation dictionary for Myanmar language was built and the applicability of the dictionary for G2P conversion was explored by applying machine learning techniques such as sequence to sequence modelling in Chapter 4. This dictionary was prepared with syllable information and was used as the main entries of Myanmar pronunciation dictionary for G2P conversion in Festival. The correct phoneme representation of Myanmar text with syllable information have been achieved by this pronunciation dictionary.

Contextual linguistic information used in acoustic modelling of speech synthesis model is important for the naturalness in speech synthesis. Therefore, HTS style contextual labels are extracted by configuring and applying Festival speech synthesis architecture with our phoneme features file and Myanmar pronunciation dictionary with syllable information. The question set for Myanmar language which is the language dependent requirement, is proposed and used for context clustering of HMM-based speech synthesis and extracting linguistic features of neural network based speech synthesis. The tasks done in Chapter 3, 4 and 5 are integrated as a text analysis part for Myanmar TTS system and contextual linguistic features obtained from this text analysis part have been applied in acoustic modelling of Myanmar TTS systems.

According to the experimental results described in Chapter 8 and 9, it can be confirmed that these contextual linguistic features and proposed question set are effective in modelling DNN and LSTM-RNN based Myanmar speech synthesis in terms of naturalness.

The accuracy of acoustic model in TTS system is very important for getting better performance in naturalness. Therefore, state-of-the-art acoustic modelling techniques such as DNN and LSTM-RNN are applied in Myanmar speech synthesis to promote the quality of synthesized speech.

The importance of word information is explored by using CLUSTERGEN on Myanmar speech synthesis in Chapter 6 and it can be confirmed that word information can promote the naturalness of Myanmar TTS system although word segmentation process is still ambiguous. According to this preliminary result, many contextual linguistic information including word information are taken into account for further experiments of Myanmar TTS system. HMM-based system with many contextual linguistic labels was conducted and used as the baseline system.

Due to the limitation of decision tree clustered context dependent HMMs, DNN has been applied in acoustic modelling of Myanmar speech synthesis. Comparison was done on HMM and DNN based systems and the objective and subjective results presented in Chapter 7 shows that DNN based system surpassed the HMM based system. The effect of precise state boundaries and coarse phone boundaries on aligning input linguistic features and output acoustic features for training DNN are investigated and better performance was found in using precise state boundaries. Therefore, precise state boundary alignment was used in the further experiments.

LSTM-RNN architecture has been applied in acoustic modelling of Myanmar speech synthesis to promote the quality of synthesized speech because LSTM-RNN can model efficiently sequential data that contains correlations between consecutive frames in speech. To the best of our knowledge, this is the first attempt to apply LSTM-RNN architecture in Myanmar speech synthesis. The comparisons are done on the performance of DNN, LSTM-RNN and a hybrid of DNN and LSTM-RNN based Myanmar TTS systems with the same linguistic features extracted from the text analysis part. Both objective and subjective results reported in Chapter 8 confirm that LSTM-RNN based systems outperform DNN based system. The hybrid of DNN and LSTM-RNN denoted as Hybrid-LSTM-2L, i.e. four feedforward hidden layers followed by two LSTM-RNN layers, offers more suitable network architecture for Myanmar speech synthesis in naturalness. Furthermore, the effect of contextual linguistic features and tone information extracted by using proposed Myanmar question set on LSTM-RNN based speech synthesis was investigated and it shows that contextual information and tone information are important for modeling LSTM-RNN based Myanmar speech synthesis.

Moreover, the effectiveness of applying word embedding features as the additional input features in acoustic modelling for Myanmar speech synthesis was investigated in this work because word vectors can be trained by unsupervised learning on the large amount of unstructured text data. Word embedding features are obtained from the word vectors trained on the collected Myanmar monolingual corpus. The comparisons are done on modelling DNN, LSTM-2L, and Hybrid-LSTM-2L based Myanmar speech synthesis with and without additional input features such as word vector features and/or POS features. Both objective and subjective results show that using word vector in acoustic modelling of DNN-based systems can improve the performance of the systems. The objective evaluation results show that using word embedding features as the additional input features in acoustic modelling cannot lead to significant improvement of LSTM-2L and Hybrid-LSTM-2L systems though word embedding features and/or POS features can give little improvement in the perception of native listeners.

In comparing the results of DNN and LSTM-RNN based systems, it shows that LSTM-RNN based systems without any additional input features such as POS or word vector features perform better than the DNN-based system with POS and word vector features.

Therefore, it can be concluded that linguistic features generated by the text analysis part and proposed Myanmar question set are good enough for LSTM-RNN based Myanmar TTS system. In this work, it can be confirmed that more natural synthesized speech for Myanmar language can be achieved by using linguistic features extracted from text analysis on a hybrid of DNN and LSTM-RNN based Myanmar TTS system. The main objective of the research to enhance Myanmar TTS system by promoting the naturalness of the synthetic speech has been accomplished in many different techniques and it is proven by the experimental results that are clearly described in this thesis.

9.1 Advantages and the Limitation of the System

The more natural synthesized speech has been achieved by applying linguistic information on LSTM-RNN based speech synthesis models for Myanmar language as it has been proved in previous chapters. This system can be applied in many application areas such as voice-over functions for visually impaired person, communication aid for speech impaired person, e-books readers, automatic question and answering system, Speech-to-Speech translation system, and communicative robots.

Better pronunciation of Myanmar numbers has been achieved in this Myanmar TTS system by implementing and integrating WFST based Myanmar number normalizer.

The effective contextual linguistic information for applying acoustic modelling of Myanmar TTS system has been generated by the text analysis part and it can be applied in further research of Myanmar speech synthesis.

The large Myanmar pronunciation dictionary built and used for G2P conversion can be applied not only in Myanmar TTS system but also in Myanmar Automatic Speech Recognition (ASR) system.

As the limitation, there can be Out-of-Vocabulary problem for uncommon words in Myanmar language because lexicon based G2P conversion is currently applied in the system.

9.2 Future Work

Phrasing structure is one of the most important factors for promoting the naturalness of TTS systems, in particular for long utterances. In this work, phrase break information could not be included in contextual linguistic features because there is no phrase break prediction model and training such kind of model needs costly manual annotation of a large corpus with phrase breaks. According to our inspection of synthesized speeches, the importance of correct phrase break makes the system to be more preferred by the listeners. Therefore, making a large corpus with phrase break annotated and training a phrasing model for Myanmar language should be attempted. After that, using phrase break features obtained from that model in acoustic modelling of Myanmar speech synthesis would be the future work for better naturalness.

It can be found that word embedding features has less effective in acoustic modelling of LSTM-RNN based Myanmar speech synthesis. Therefore, in the future work, word vectors should be learned by taking counts the acoustic information related to TTS task so that they can encode sufficient prosody information for acoustic modelling. Furthermore, the usefulness of embedded syllable vectors in acoustic modelling of Myanmar TTS systems will be examined because the tones of the syllable in Myanmar language influence the acoustic properties such as F_0 and duration.

AUTHOR'S PUBLICATIONS

[P1] Aye Mya Hlaing, Win Pa Pa, Ye Kyaw Thu, “Myanmar Number Normalization for Text-to-Speech”, In Proceedings of the 15th International Conference of the Pacific Association for Computational Linguistics **(PACLING 2017)**, Yangon, Myanmar, pp. 346-356, August 16-18 2017. Communications in Computer and Information Science Book Series (Volume 781), Springer Singapore, Series ISSN : 1865-0929 **(Scimago index – Q3)**

[P2] Aye Mya Hlaing, Win Pa Pa, Ye Kyaw Thu, “Word-based Myanmar Text-to-Speech with CLUSTERGEN”, In Proceedings of the 16th International Conference on Computer Applications **(ICCA 2018)**, Yangon, Myanmar, pp. 203–208, February 22-23, 2018.

[P3] Aye Mya Hlaing, Win Pa Pa, Ye Kyaw Thu, “DNN-Based Myanmar Speech Synthesis”, In Proceedings of the 6th International Workshop on Spoken Language Technologies for Under-Resourced Languages **(SLTU 2018)**, Gurugram, India, pp. 142–146, August 29-31, 2018.

[P4] Aye Mya Hlaing, Win Pa Pa, Ye Kyaw Thu, “Enhancing Myanmar Speech Synthesis with Linguistic Information and LSTM-RNN”, In Proceedings of 10th ISCA Speech Synthesis Workshop **(SSW10)**, Vienna, Austria, pp. 189-193, September 20-22, 2019.

[P5] Aye Mya Hlaing, Win Pa Pa, “Sequence-to-Sequence Models for Grapheme to Phoneme Conversion on Large Myanmar Pronunciation Dictionary”, In Proceedings of the 22th Conference of Oriental COCOSDA **(Oriental COCODSA 2019)**, Cebu City, Philippines, pp. 149-153, October 25-27, 2019.

[P6] Aye Mya Hlaing, Win Pa Pa, “Word Representations for Neural Network Based Myanmar Text-to-Speech System”, International Journal of Intelligent Engineering and Systems **(IJIES)**, Vol. 13, No. 2, pp. 239-249, Japan, April 2020. ISSN : 2185-3118 **(Scimago index – Q2)**

BIBLIOGRAPHY

[1] R. Al-Rfou, B. Perozzi, and S. Skiena, “Polyglot: Distributed word representations for multilingual NLP”, arXiv preprint arXiv:1307.1662, 2013.

[2] G. K. Anumanchipalli and A. Black, “Adaptation Techniques For Speech Synthesis in Under-resourced”, SLTU 2010, Penang, Malaysia, 2010.

[3] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, “A neural probabilistic language model”, Journal of machine learning research, pp. 1137-1155, 2003.

[4] B. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal, “The AT&T Next-Gen TTS system”, in Proc. Joint ASA/EAA/DAEA Meeting, pp. 15–19, 1999.

[5] M. Bisani and H. Ney, “Joint-sequence models for grapheme-to- phoneme conversion”, Speech communication, vol. 50, no. 5, pp. 434–451, 2008.

[6] A. W. Black, “CLUSTERGEN: A statistical parametric synthesizer using trajectory modeling”, in Proc. Interspeech, pp. 1762–1765, 2006.

[7] A. Black and K. Lenzo, “Building voices in the Festival Speech Synthesis System,” http://festvox.org/bsv/, 2000.

[8] A. Breen and P. Jackson, “A phonologically motivated method of selecting nonuniform units”, in Proc. Int. Conf. Spoken Lang. Process., pp. 2735–2738, 1998.

[9] F. Chollet, et al., “Keras: The python deep learning library,” Astrophysics Source Code Library, 2018.

[10] G. Coorman, J. Fackrell, P. Rutten, and B. Coile, “Segment selection in the L&H realspeak laboratory TTS system”, in Proc. Int. Conf. Spoken Lang. Process., pp. 395–398, 2000.

[11] R. E. Donovan and E. M. Eide, “The IBM trainable speech synthesis system”, in Proc. Int. Conf. Spoken Lang. Process., pp. 1703–1706, 1998.

[12] P. Ebden, R. Sproat, “The kestrel tts text normalization system”, Natural Language Engineering 21(03), pp. 333-353, 2015.

[13] Y. Fan, Y. Qian, F. L. Xie, F. K. Soong, “TTS synthesis with bidirectional lstm based recurrent neural networks”, in: Fifteenth Annual Conference of the International Speech Communication Association, 2014.

[14] J. L. Flanagan, K. Ishizaka, and K. L. Shipley, “Synthesis of speech from a dynamic model of the vocal cords and vocal tract”, The Bell System Technical Journal, 54(3), pp. 485-506, 1975.

[15] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, “An adaptive algorithm for mel-cepstral analysis of speech”, in Proc. Int. Conf. Acoust. Speech Signal Process., pp. 137–140, 1992.

[16] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov, “Learning word vectors for 157 languages”, arXiv preprint arXiv:1802.06893, 2018.

[17] A. Graves, “Supervised sequence labelling”, in Supervised sequence labelling with recurrent neural networks, Springer, pp. 5–13, 2012.

[18] A. Graves, J. Schmidhuber, “Framewise phoneme classification with bidirectional lstm and other neural network architectures”, Neural Networks Vol.18, Issues (5-6), pp. 602–610, 2005.

[19] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, J. Schmidhuber, “Lstm: A search space odyssey”, IEEE transactions on neural networks and learning systems 28 (10), pp. 2222–2232, 2017.

[20] E. T. Gunawan, D. Arifianto, “Natural Indonesian Speech Synthesis by using CLUSTERGEN”, International Conference on Information, Communication Technology and System, 2014.

[21] A. M. Hlaing, W. P. Pa, “Sequence-to-Sequence Models for Grapheme-to-Phoneme Conversion on Large Myanmar Pronunciation Dictionary”, In: Proc. of the 22th Conference of the Oriental COCOSDA, pp. 149-153, 2019.

[22] S. Hochreiter, J. Schmidhuber, “Long short-term memory”, Neural computation 9 (8), pp. 1735–1780, 1997.

[23] A. J. Hunt and A. W. Black, “Unit selection in a concatenative speech synthesis system using a large speech database”, In 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, vol. 1, pp. 373-376. IEEE, 1996.

[24] D. Jurafsky, J.H.Martin, “Speech and Language Processing”, Second edition, Pearson Prentice Hall Series in Artificial Intelligence, 2008.

[25] O. Karaali, G. Corrigan, and I. Gerson, “Speech synthesis with neural networks”, in Proc. World Congress on Neural Networks , pp. 45–50, 1996.

[26] O. Karaali, G. Corrigan, I. Gerson, and N. Massey, “Text-to-speech conversion with neural networks: A recurrent TDNN approach,” in Proc. Eurospeech, pp. 561–564, 1997.

[27] G. Kikui, E. Sumita, T. Takezawa, and S. Yamamoto, “Creating corpora for speech-to-speech translation”, In: Proc. of the Eighth European Conference on Speech Communication and Technology, pp. 381-384, 2003.

[28] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.

[29] D. H. Klatt, “Software for a cascade/parallel formant synthesizer”, J. Acoust. Soc. Amer., vol. 67, pp. 971–995, 1980.

[30] Z. H. Ling, S. Y. Kang, H. Zen, A. Senior, M. Schuster, X. J. Qian, H. M. Meng, and L. Deng, “Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends”, IEEE Signal Processing Magazine, 32(3), pp.35-52, 2015.

[31] H. Lu, S. King, and O. Watts, “Combining a vector space representation of linguistic context with a deep neural network for text-to-speech synthesis”, In Eighth ISCA Workshop on Speech Synthesis, 2013.

[32] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality”, In: Proc. of Advances in Neural Information Processing Systems, pp. 3111-3119, 2013.

[33] T. Mikolov, J. Kopecky, L. Burget, and O. Glembek, “Neural network based language models for highly inflective languages”, In: Proc. of 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4725-4728, IEEE, 2009.

[34] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space”, ICLR Workshop, 2013.

[35] M. Morise, F. Yokomori, and K. Ozawa, “World: a vocoder-based high-quality speech synthesis system for real-time applications”, IEICE

TRANSACTIONS on Information and Systems, vol. 99, no. 7, pp. 1877--1884, 2016.

[36] E. Moulines, and F. Charpentier, “Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones”, Speech communication 9, no. 5-6, pp. 453-467, 1990.

[37] W. P. Pa, Y. K. Thu, A. Finch, and E. Sumita, “Word boundary identification for Myanmar text using conditional random fields”, in International Conference on Genetic and Evolutionary Computing. Springer, pp. 447–456, 2015.

[38] B. Prachya, S. Thepchai, “Technical report for the network-based asean language translation public service project”, Online Materials of Network-based ASEAN Languages Translation Public Service for Members, NECTEC, 2013.

[39] K. Prahallad, A. Black, and R. Mosur, “Subphonetic modeling for capturing pronunciation variation in conversational speech synthesis,” in Proceedings of ICASSP 2005, Toulouse, France, 2006.

[40] Y. Qian, Y. Fan, W. Hu, and F. K. Soong, “On the training aspects of deep neural network (dnn) for parametric tts synthesis”, in Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, pp. 3829—3833, 2014.

[41] K. Rao, F. Peng, H. Sak, and F. Beaufays, “Grapheme-to- phoneme conversion using long short-term memory recurrent neural networks”, in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4225–4229, 2015.

[42] M. S. Ribeiro, O. Watts, and J. Yamagishi, “Learning Word Vector Representations Based on Acoustic Counts”, In: Proc. of INTERSPEECH 2017, pp. 799-803, 2017.

[43] H. Riza, M. Purwoadi, Gunarso, T. Uliniansyah, et al., “Introduction of the Asian Language Treebank”, in Proc. of 2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA), pp. 1-6, 2016.

[44] B. Roark, R. Sproat, C. Allauzen, M. Riley, J. Sorensen, T. Tai, “The opengrm open-source finite-state grammar software libraries”, in: Proceedings of the ACL 2012 System Demonstrations. pp. 61--66. Association for Computational Linguistics, 2012.

[45] D. Rumelhart, G. Hinton, and R. Williams, “Learning representations by backpropagating errors,” Nature, vol. 323, no. 6088, pp. 533–536, 1986.

[46] K. Shinoda and T. Watanabe, “MDL-based context-dependent subword modeling for speech recognition”, J. Acoust. Soc. Japan (E), 21:79–86, March 2000.

[47] E. P. P. Soe and A. Thida, “Diphone-concatenation speech synthesis for Myanmar language”, International Journal of Science, Engineering and Technology Research, vol. 2, no. 5, pp. 1078-1087, 2013.

[48] P. Taylor, “Text-to-Speech Synthesis”, Cambridge University Press, 2009.

[49] 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 Conference of the Pacific Association for Computational Linguistics. Springer, pp. 238–250, 2015.

[50] Y. K. Thu, W. P. Pa, F. Andrew, A. M. Hlaing, H. M. S. Naing, E. Sumita, and C. Hori, “Syllable pronunciation features for Myanmar grapheme to phoneme conversion”, in: The 13th International Conference on Computer Applications (ICCA2015), pp. 161–167, 2015.

[51] Y. K. Thu, W. P. Pa, J. Ni, Y. Shiga, A. Finch, C. Hori, H. Kawai, and E. Sumita, “HMM based Myanmar text to speech system”, in Sixteenth Annual Conference of the International Speech Communication Association, 2015.

[52] Y. K. Thu, W. P. Pa, Y. Sagisaka, and N. Iwahashi, “Comparison of grapheme-to-phoneme conversion methods on a Myanmar pronunciation dictionary”, in Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016), pp. 11–22, 2016.

[53] K. Tokuda, H. Zen, and A. W. Black, “An HMM-based speech synthesis system applied to English”, in IEEE Speech Synthesis Workshop, pp. 227--230, 2002.

[54] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, “Multi-space probability distribution HMM”, IEICE Trans. Inf. Syst., vol. E85-D, no. 3, pp. 455–464, 2002.

[55] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, “Hidden Markov models based on multi-space probability distribution for pitch pattern modeling”, In Proc. ICASSP-99, pp. 229–232, 1999.

[56] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, T. Kitamura, “Speech parameter generation algorithms for hmm-based speech synthesis”, In: Proc. of Acoustics, Speech, and Signal Processing, 2000. ICASSP’00. Proceedings. 2000 IEEE International Conference on, Vol. 3, IEEE, pp. 1315–1318, 2000.

[57] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, “Speech synthesis based on Hidden Markov Models”, Proceedings of the IEEE, vol. 101, no. 5, pp. 1234--1252, 2013.

[58] S. Toshniwal and K. Livescu, “Jointly learning to align and convert graphemes to phonemes with neural attention models”, in 2016 IEEE Spoken Language Technology Workshop (SLT). IEEE, pp. 76–82, 2016.

[59] C. Tuerk and T. Robinson, “Speech synthesis using artificial neural networks trained on cepstral coefficients,” in Proc. Eurospeech, pp. 1713–1716, 1993.

[60] T. Tun, “Acoustic phonetics and the phonology of the Myanmar language”, First Edition, Win Yadanar Press, 2007.

[61] T. Tun, “The domain of tones in Burmese”, SST 1990 Proceedings, pp.406-411, 1990.

[62] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need”, in Advances in neural information processing systems, pp. 5998–6008, 2017.

[63] X. Wang, S. Takaki, and J. Yamagishi, “Investigation of using continuous representation of various linguistic units in neural network based text-to-

speech synthesis", IEICE Transactions on Information and Systems, Vol.99, No.10, pp. 2471-2480, 2016.

[64] X. Wang, S. Takaki, and J. Yamagishi, "Enhance the Word Vector with Prosodic Information for the Recurrent Neural Network Based TTS System", In: Proc. of INTERSPEECH 2016, pp. 2856-2860, 2016.

[65] P. Wang, Y. Qian, F. K. Soong, L. He, and H. Zhao, "Word embedding for recurrent neural network based TTS synthesis", In: Proc. of 2015 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4879-4883, 2015.

[66] 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.

[67] Z. Wu, and S. King, "Investigating gated recurrent networks for speech synthesis", In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5140-5144. IEEE, 2016.

[68] Z. Wu, C. Valentini-Botinhao, O. Watts, and S. King, "Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis", in Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on. IEEE, pp. 4460—4464, 2015.

[69] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system," Proc. SSW, Sunnyvale, USA, 2016.

[70] J. Yamagishi, "An introduction to hmm-based speech synthesis", Technical Report, 2006.

[71] K. Yao and G. Zweig, "Sequence-to-sequence neural net models for grapheme-to-phoneme conversion," arXiv preprint arXiv:1506.00196, 2015.

[72] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis", in Proc. Eurospeech, pp. 2347–2350, 1999.

[73] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X.-Y. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland,

“The Hidden Markov Model Toolkit (HTK) Version 3.4”, 2006. [Online] Available: http://htk.eng.cam.ac.uk/.

[74] S. J. Young, J. J. Odell, and P. C. Woodland, “Tree-based state tying for high accuracy acoustic modeling”, In Proc. ARPA Human Language Technology Workshop, pages 307–312, March 1994.

[75] H. Zen, A. Senior, and M. Schuster, “Statistical parametric speech synthesis using deep neural networks”, in Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, pp. 7962--7966, 2013.

[76] H. Zen, H. Sak, “Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis”, in: Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on, IEEE, pp. 4470–4474, 2015.

[77] H. Zen, K. Tokuda, and A. W. Black, “Statistical parametric speech synthesis”, speech communication 51, no. 11, 1039-1064, 2009.

[78] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Hidden semi-Markov model based speech synthesis system”, IEICE Trans. Inf. Syst., vol. E90-D, no. 5, pp. 825–834, 2007.

[79] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, “Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005”, IEICE Trans. Inf. Syst., vol. E90-D, no. 1, pp. 325–333, 2007.

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