Received: December 26, 2019. Revised: January 22, 2020.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Word Representations for Neural Network Based Myanmar Text-to-Speech System

Aye Mya Hlaing[1]* **Win Pa Pa**[1]

[1]*Natural Language Processing Lab., University of Computer Studies, Yangon, Myanmar*
* Corresponding author's Email: ayemyahlaing@ucsy.edu.mm

Abstract: The main objective of this paper is to improve the naturalness of Myanmar Text-to-Speech (TTS) system without using time-consuming and expensive annotation of training corpus. Recently, word embedding, which has the advantage of training directly from a large amount of raw text data, has been used as the additional input features with the conventional input features or as the replacement of that conventional input features on the acoustic modelling of TTS systems. In this paper, the effectiveness of applying word vectors as the additional input features are investigated for Myanmar speech synthesis on the three acoustic modelling techniques, Deep Neural Network (DNN), Long Short-Term Memory Recurrent Neural Network (LSTM-RNN), and a hybrid of DNN and LSTM-RNN (Hybrid-LSTM-RNN). For the purpose of achieving better TTS performance, we built our own word vectors for Myanmar language. We further explore the best modelling method and vector dimension of word embedding for Myanmar TTS systems. Both objective and subjective evaluations are done on DNN, LSTM-RNN and Hybrid-LSTM-RNN based Myanmar TTS systems with and without additional input features such as Part-of-Speech (POS) and word vectors. According to the subjective results, applying additional input features in DNN, LSTM-RNN, and Hybrid-LSTM-RNN based Myanmar TTS systems can improve the naturalness of the synthesized speeches though objective results cannot lead to the significant improvement in LSTM-RNN and Hybrid-LSTM-RNN based systems. To the best of our knowledge, this is the first attempt to apply word vector features in neural network based Myanmar TTS systems.

Keywords: Word embedding, Word representations, Myanmar text-to-speech, Myanmar speech synthesis, Long short-term memory, Recurrent neural network, Part-of-speech, Word vectors, Hybrid-LSTM-RNN.

1. Introduction

Conventional TTS system consists of text analysis (front-end) and speech waveform generation (back-end) modules. Linguistic analysis and prosody prediction are done in front-end part, and intermediate linguistic representations which includes rich segmental and suprasegmental information are generated. The back-end uses these intermediate linguistic representations for acoustic modelling and synthesizes the speech waveform. The front-end can be divided into several modules and each module is usually trained by using manually annotated corpus. Among them, some suprasegmental features for intonation prediction such as Tone and Break Indices (ToBI) can be achieved by manually annotated training corpus with high consistency among different annotators. This kind of annotation is time consuming and very expensive. Therefore, it is worthwhile to find the way of improvement in naturalness of TTS systems without using such ToBI features.

Recently, distributed word representations or word vectors which can be obtained by unsupervised learning from large amount of unstructured text data have been applied in speech synthesis. These distributed representations of words in a vector space can capture a large number of precise syntactic and sematic word relationships [1]. The continuous-valued word features have been used as surrogates for part-of-speech features in phrase-break prediction task [2] and as the additional features to phrase break prediction using bidirectional long short term memory (BLSTM) modelling [3]. A Vector Space Model-based approach for linguistic specification [4] was combined with Deep Neural Networks (DNN) for TTS synthesis [5]. In [6], four different kinds of published word embedding were tested and using word embedding can improve the performance of BLSTM recurrent neural network (RNN) based TTS synthesis without POS and ToBI input features. However, it still has a gap to the upper bound system which uses manually labelled POS and ToBI as input features for training acoustic model. The embedded vectors of various linguistic units were used as the additional or alternative features in neural network based acoustic modelling and the results indicated that using that features only lead to insignificant improvement of RNN based acoustic model [7]. Enhanced word embedding in combination with acoustic parameters were utilized in acoustic model of TTS synthesis [8, 9].

For Myanmar language, only four Statistical Parametric Speech Synthesis (SPSS) work, HMM-based Myanmar TTS [10], CART-based Myanmar TTS [11], DNN-based Myanmar speech synthesis [12], and enhancing of Myanmar speech synthesis with Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) [13] are found publicly. Many linguistic features were extracted by utilizing proposed Myanmar question set and these features have been used in neural network based acoustic modelling for Myanmar speech synthesis as the conventional input linguistic features [12, 13]. However, part of speech (POS) features and intonation features such as ToBI are not included in acoustic modelling because there is no high accuracy POS tagger and manually annotated corpus with ToBI tags for Myanmar language. Therefore, in this work, we investigated the effectiveness of word vectors whether they help to promote the naturalness of synthesized speech as the additional linguistic representations to the conventional linguistic features without using ToBI features and compared the objective and subjective results with the results of previous published work in [12, 13].

For this purpose, we collected monolingual Myanmar text corpus for building word embedding for Myanmar language. As far as we know, none of the previously Myanmar TTS has been used word embedding features for acoustic modelling. To find the best dimension and modelling of word vectors for Myanmar language, Continuous Bag-of-Words (CBOW) and Skip-gram models with different dimensions are investigated on DNN based Myanmar speech synthesis, and the best word vector is applied in other speech synthesis models. The effectiveness of the word embedding features in DNN, LSTM-RNN, and Hybrid-LSTM-RNN based Myanmar speech synthesis were examined by both objective and subjective evaluations in this work.

The rest of this paper is organized as follows. Section 2 shows existing front-end processing flow of Myanmar TTS. The proposed neural network based Myanmar TTS system with different input features are presented in section 3. Section 4 describes the detail process of preparing data sources for Myanmar language and building word vectors for Myanmar language. Section 5 shows the experiments and the results of objective and subjective evaluations on the experiments. The presented work is summarized in section 6.

2. Text analysis for Myanmar TTS

Fig. 1 shows the existing process flow of text analysis part in Myanmar TTS. Word segmentation is the essential preprocessing step for text analysis because Myanmar text generally lacks white space between words although space is sometimes included between phrases. We applied word segmentation using conditional random fields (CRFs) [14]. In text normalization module, semiotic classes for Myanmar language were identified and Weighted Finite State Transducers (WFST) based Myanmar number normalization was implemented as the separate module [15]. This WFST-based module was integrated into the pipeline of text analysis. For grapheme to phoneme conversion, we built large Myanmar pronunciation dictionary [16] and syllable information was also included in it. Phoneme features including consonant features such as consonant type, place of articulation, consonant voicing, and lip rounding and vowels features such as vowel frontness,

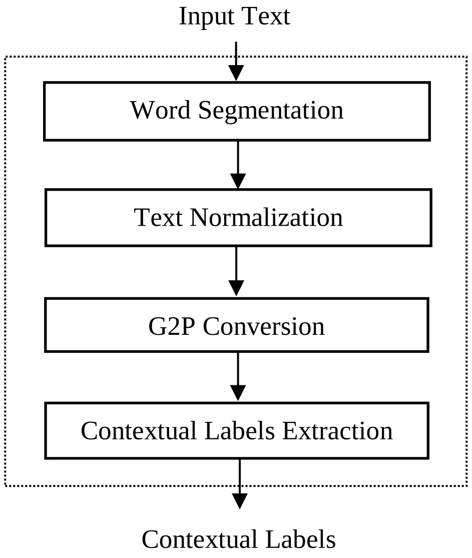


Figure. 1 Text analysis for Myanmar language

vowel height, tone, and nasality has been prepared [11]. These phoneme features and large pronunciation dictionary with syllable information have been used in Festival [17] to extract contextual labels for utterances formatted as HTS-style label [18].

However, these contextual labels are not included POS information and ToBI information [12, 13], and they are referred to as conventional input linguistic features in this paper.

3. Neural network based Myanmar TTS systems with different input features

Fig. 2 shows the structure of neural network based Myanmar speech synthesis with different input features. Input text is converted into contextual labels by text analysis phase. Input linguistic features, L_t and LP_t at frame t are extracted from contextual labels by using a question set for Myanmar language. L_t is the conventional input linguistic features used in our previous work [12, 13] and it includes the binary features for categorical contexts (e.g. phoneme identity) and numerical features (e.g. the number of syllable in the word). LP_t refers L_t with POS features of the preceding, current, and succeeding words.

Input utterances are segmented into words and getting word vector features for the current word w_t is done by using the Myanmar word embedding model. $C(w_t)$ is word vector representation for the current word w_t at frame t.

The input feature vectors, I1 and I2 include L_t and LP_t respectively. I3 is achieved by cascading of L_t and $C(w_t)$, and I4 is achieved by cascading of LP_t and $C(w_t)$.

Different input feature vectors, I1, I2, I3, and I4 are applied in acoustic modelling of all DNN, LSTM-RNN, and Hybrid-LSTM-RNN based Myanmar speech synthesis. Using these different input feature vectors in the state-of-the-art acoustic modelling techniques is our contribution of this work.

The acoustic features O_t is the output acoustic features at frame t and the features are 60-

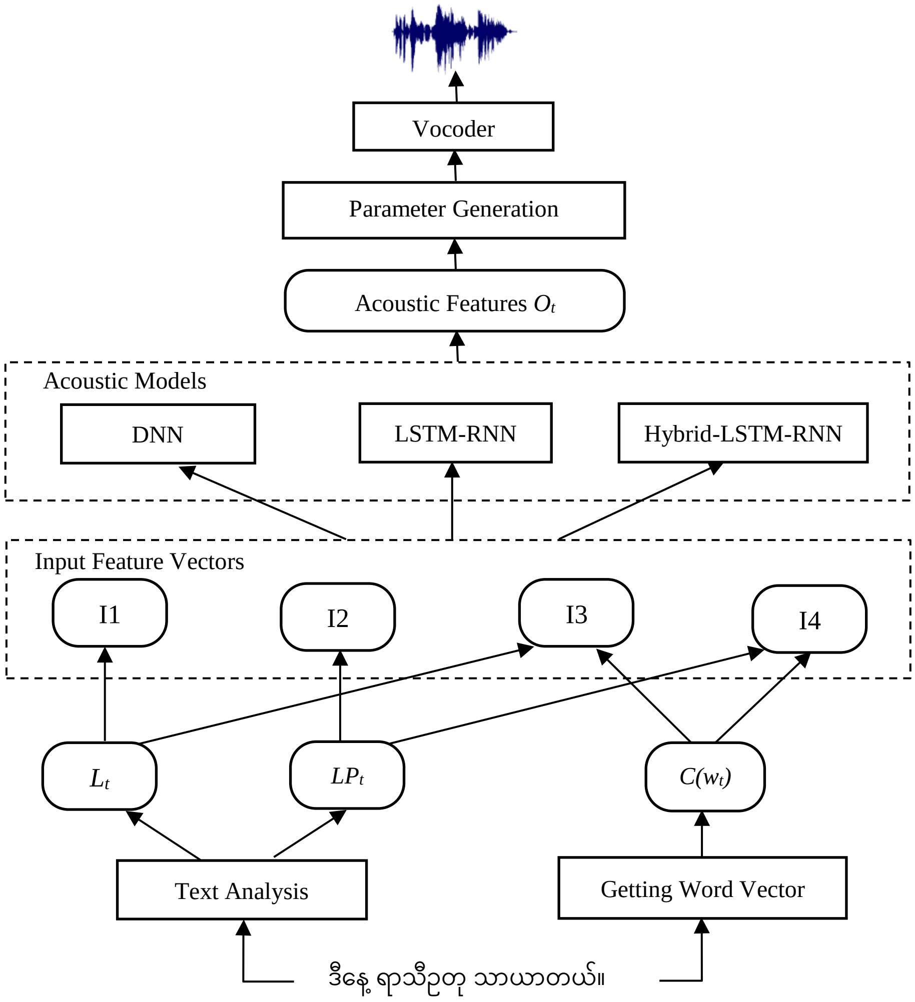


Figure. 2 Structure of neural network based Myanmar speech synthesis with different input features

dimensional Mel-Cepstral Coefficient (MCCs), 5-dimentional band aperiodicities (BAPs), and logarithmic fundamental frequencies (log F_0). A binary voiced/unvoiced feature is used for voicing information.

In the training stage, input linguistic features and output acoustic features are force aligned frame-by-frame by Hidden Markov Models (HMMs) in advance. The weights of neural network are initialized randomly and then they are updated to minimize mean squared error (mse) between target and predicted output features.

At the synthesis time, input features are mapped to output acoustic features by a trained acoustic model. The speech parameter generation step can generate smooth trajectories of speech parameter features. Finally, a synthesized waveform is generated by the vocoder according to the given speech parameters.

4. Building word embedding for Myanmar language

Distributed word representations or word vectors have been applied to natural language processing tasks and achieved the state-of-the-art performance [19, 20]. Recently, word vectors have been applied in speech synthesis [6-9]. 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 [21, 22]. In [21], the size of word vectors is small, and it contains about 55K entries for Myanmar language and can be downloaded from the link [23]. In [22], 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 [24]. 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 we have built our own word vectors for Myanmar language with standard encoding Unicode for more coverage and better performance.

4.1 Data collection

Firstly, we collected a large monolingual Myanmar corpus for the purpose of building high quality word vectors with wide coverage. Myanmar data from Asian Language Treebank (ALT) parallel corpus [25] 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 [26]. Another one is Myanmar data of ASEAN-MT parallel data [27], 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 our 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 1.

4.2 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 [14] was done on the data collected from websites and blogs. This corpus was used in modelling word vectors for Myanmar language.

4.3 Modelling word vector

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) [28, 29]. Two particular models for learning word representations that can be efficiently trained on large amounts of text data are CBOW and Skip-gram

Table 1. 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%

Received: December 26, 2019. Revised: January 22, 2020.

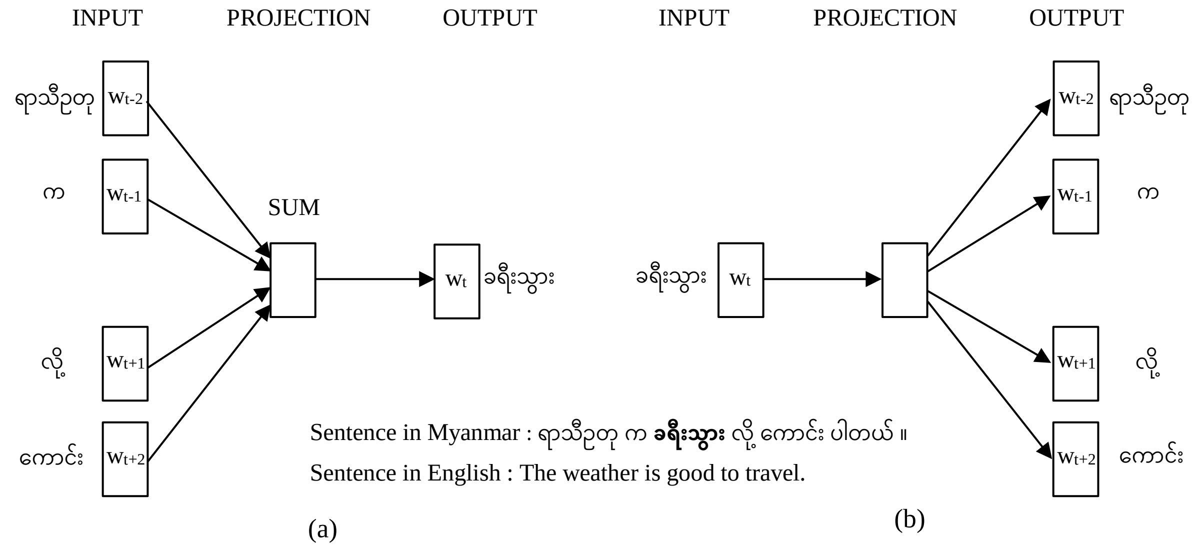


Figure. 3 Model architectures of CBOW and Skip-gram: (a) CBOW and (b) Skip-gram

models [1, 19]. 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 neighbouring words within a sentence. Distributed representation of current word is used to predict words within a certain range before and after the current word. Fig. 3 shows the architectures of CBOW and Skip-gram models for an example Myanmar sentence.

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

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

where c is the size of training context. The basic Skip-gram defines $p(w_{t+1}|w_t)$ using softmax function and the formulation is expensive. Therefore, an efficient Negative sampling (NEG) is used to replace over $log\, p(w_{t+1}|w_t)$ term in Skip-gram formulation [1]. This NEG is also used for training the CBOW model.

Using the monolingual corpus for Myanmar language described in Section 4.1, we trained word vectors with word2vec [30], a tool for computing continuous distributed representations of words.

Table 2. Examples of the nearest five neighbors of 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

CBOW and Skip-ram models with different sizes of word vectors are trained. The number of total words in train file 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 2 and input words are shown in the bold style.

5. Experiments and results

Objective and subjective evaluations are done to explore the effect of additional word vector features on neural network based Myanmar TTS systems. The objective and subjective results are compared to the previous best results of DNN based Myanmar speech synthesis published in [12] and LSTM-RNN based Myanmar speech synthesis published in [13]. All systems used the same training data and test data.

5.1 Objective evaluation

Objective results are used to measure the quality of synthesized speech in terms of distortions between the synthesized speech and natural speech. The objective measures used in this work are Mel Ceptral Distortion (MCD) in dB and *F₀* distortion in root mean square error (RMSE). The lower value means the better performance. 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 by the following equation.

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

In Eq. (2), $v_d(t)$ 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), and T' is the number of non-silence frames.

RMSE of F_0 distortion is calculated by Eq. (3).

$$RMSE = \sqrt{\frac{\sum_{t\in V}(f_{syn}(t) - f_{ref}(t))^2}{\#V}} \quad (3)$$

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. These objective measures were used for evaluating the performance of TTS systems in our experiments.

5.2 Experimental setup for neural network based Myanmar TTS systems

The data source for training all neural network based TTS was Myanmar phonetically balanced corpus (PBC) [10] built from Basic Travel Expression Corpus (BTEC) [31] recorded by a female speaker. 3,800 utterances were used as the training set, 100 utterances as the development set, and 100 utterances as the test set. We used 16kHz sampling rate for speech data.

Conventional input linguistic features were generated from our front-end shown in Section 2 as the basic input features which consists of current context, and preceding and succeeding two contexts at phoneme, syllable, word, and utterance levels. Additional 9 numeric frame related features were also used in all our experiments.

For the acoustic features, 60-dimensional MCCs, 5-dimentional BAPs, and log F_0 at 5 msec frame step were extracted by using WORLD [32] vocoder. 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 used to generate smooth parameter trajectories at synthesis time.

The model structure of DNN based systems has six feedforward hidden layers of 1024 hyperbolic tangent units each as the same configuration in [12]. The LSTM-RNN and Hybrid-LSTM-RNN based systems follow the configurations of LSTM-RNN based systems in [13]. The LSTM-RNN system has two hidden layers with 512 LSTM units each and Hybrid-LSTM-RNN system is a structure of four feedforward hidden layers of 1024 hyperbolic tangent units each, followed by two LSTM-RNN layers with 512 units. Stochastic gradient descent (sgd) based learning rate scheduling was used for DNN and Hybrid-LSTM-RNN based systems and Adam optimizer [33] was used for LTSM-RNN based systems. Merlin speech synthesis toolkit [34] was used for modelling all systems on K80 GPU. Keras [35] python library was used with Merlin for training LSTM-RNN and Hybird-LSTM-RNN based system.

5.3 Experimental setup of training word vectors for Myanmar language

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 [1]. 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 3.

5.4 Effect of training method and vector dimension of word vectors on DNN based Myanmar TTS system

We evaluated the performance of each system by

Table 3. Trained Word Vectors for Myanmar language

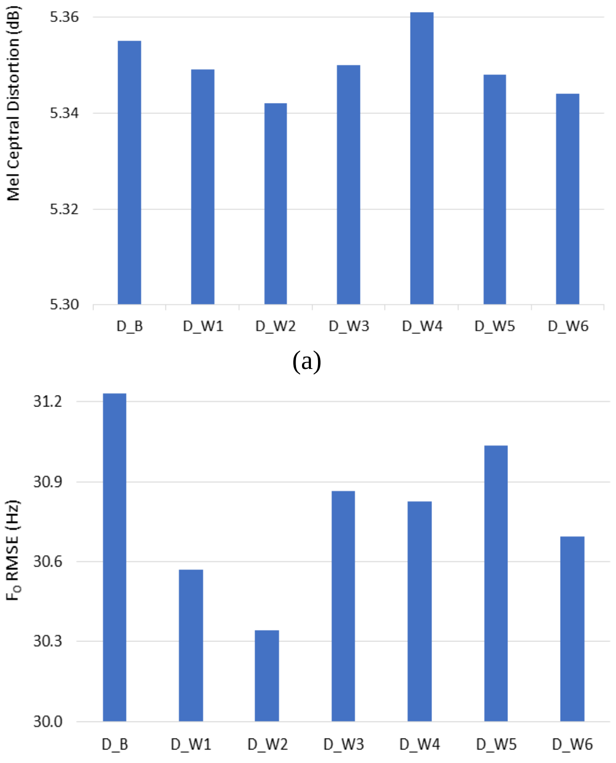
Alias	**Training Method**	**Dimension**
W1	CBOW	100
W2	CBOW	200
W3	CBOW	300
W4	Skip-gram	100
W5	Skip-gram	200
W6	Skip-gram	300

using two types of objective measures: MCD in dB and F_0 distortion in RMSE in terms of Hz. 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 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 best DNN based acoustic model presented in [12], and "D_W#" as the DNN based TTS system with additional input word vector W#, the alias of trained word vector shown in Table 3. Fig. 4 depicts the objective results of DNN based Myanmar TTS system with different word vectors. As shown in Fig. 4(a) and 4(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. Acoording 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 systems for futher experimetns.

5.5 Performance of neural network based Myanmar TTS systems with different input features

The effectiveness of the word embedding features in DNN, LSTM-RNN, and Hybrid-LSTM-RNN based Myanmar TTS systems are investigated. Besides, we compared the contribution of word vector features with POS features in all systems. We analysed the effectiveness of word vector features, POS features, and combination of these two features in Myanmar TTS systems by comparing the previous best results of DNN based system in [12] and LSTM-RNN based systems in [13] which used conventional input linguistic features.

Myanmar pronunciation dictionary [16] 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.



(a)

(b)

Figure. 4 Objective results of DNN based systems with different word vectors: (a) Mel Ceptral Distortion in dB and (b) F_0 RMSE in Hz

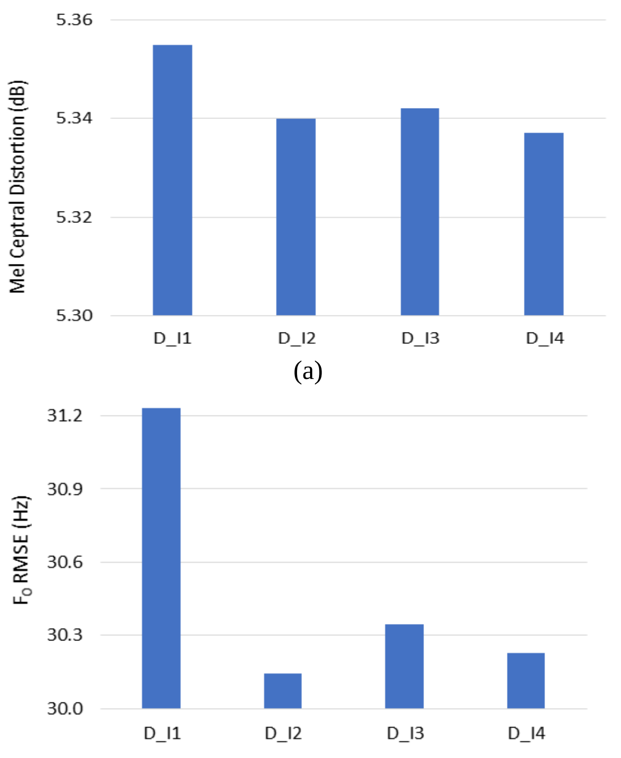
Four sets of different input feature vectors, I1, I2, I3, and I4 presented in Fig. 2 are used in training DNN, LSTM-RNN, and Hybrid-LSTM-RNN based acoustic models and the detail information can be seen as follows:

1) I1 : conventional input linguistic features used in previous published work [12, 13]
2) I2 : I1 + POS features
3) I3 : I1 + word vector features acquired from "W2"
4) I4 : I1 + POS features and word vector features acquired from "W2"

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

We denoted "D_I#", "L_I#", and "HL_I#" as the short forms of DNN, LSTM-RNN and Hybrid-LSTM-RNN based Myanmar TTS systems with input features I# respectively and the objective results of these systems are shown in Fig. 5, 6, and 7. The term D_I1 refers the best DNN based acoustic model presented in [12]. The L_I1 and HL_I1 are the best LSTM-RNN and Hybrid-LSTM-RNN based Myanmar speech synthesis models reported in [13].

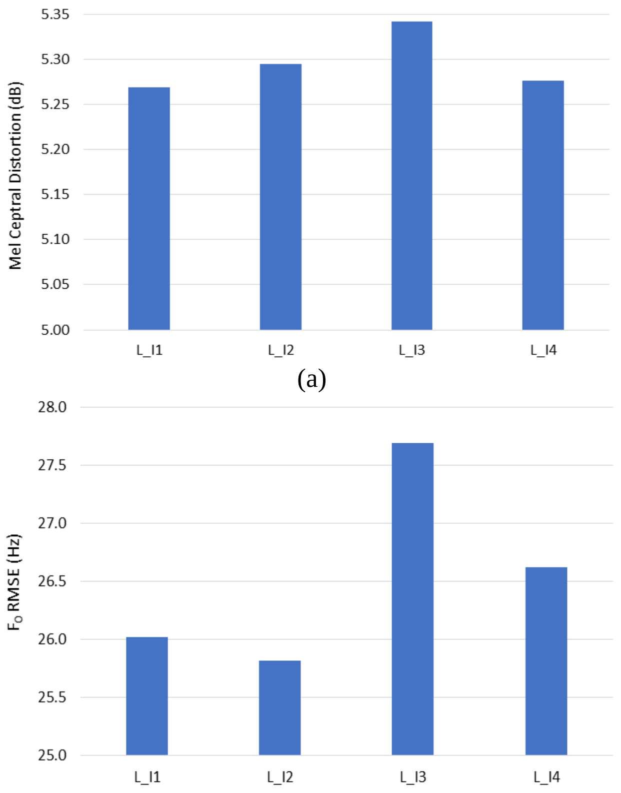
According to the results shown in Figs. 5 (a) and



(a)

(b)

Figure. 5 Objective results of DNN based Myanmar TTS systems with different input features: (a) Mel Ceptral Distortion in dB and (b) F_0 RMSE in Hz



(a)

(b)

Figure. 6 Objective results of LSTM-RNN based Myanmar TTS systems with different input features: (a) Mel Ceptral Distortion in dB and (b) F_0 RMSE in Hz

(b), D_I2, D_I3, and D_I4 give the better results than D_I1. As the comparison of MCD in Fig. 5 (a), D_I4 gets the best result among all. 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.

As we can see in Fig. 6, 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-RNN based systems.

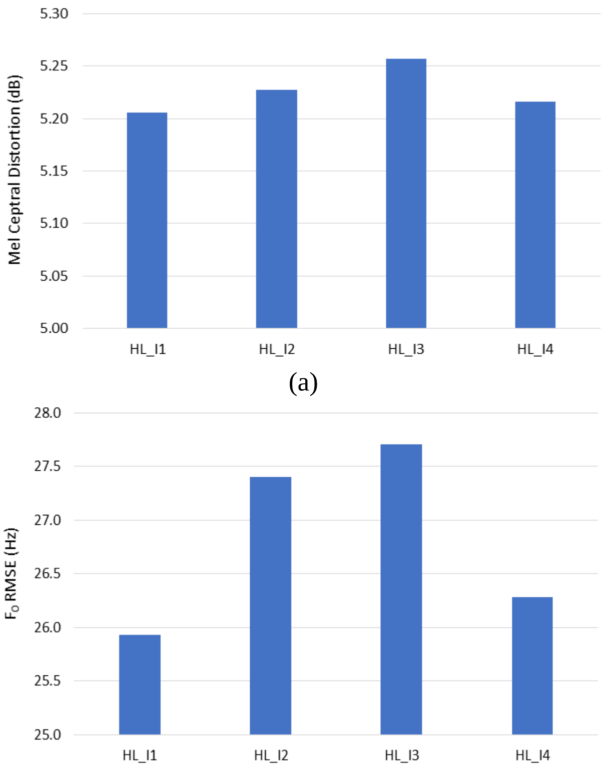
In the case of Hybrid-LSTM-RNN based systems, POS or word embedding features cannot give further improvement as shown in Fig. 7. According to the results of HL_I4 in both Fig. 7(a) and 7(b) over that of HL_I2 and HL_I3 systems, using both features in acoustic modelling can slightly improve the performance of the system. Meanwhile, HL_I1 achieves the best performance in prediction of Mel Cepstrum and F_0 among all systems, it means conventional linguistic features 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 and any significant improvement cannot be found in LSTM-RNN and Hybrid-LSTM-RNN based systems. This is consistent with the results in [7] though this is different with [6].

In comparing the results of Fig. 5, 6, and 7, we can conclude that both LSTM-RNN and Hybrid-LSTM-RNN based systems 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-RNN based systems is 5.206 and the best F_0 RMSE is 25.929.

Word vector features may not be effective in acoustic modelling for LSTM-RNN and Hybrid-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 [7].

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 other linguistic features (such as some 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.



(a)

(b)

Figure. 7 Objective results of Hybrid-LSTM-RNN based Myanmar TTS systems with different input features: (a) Mel Ceptral Distortion in dB and (b) F_0 RMSE in Hz

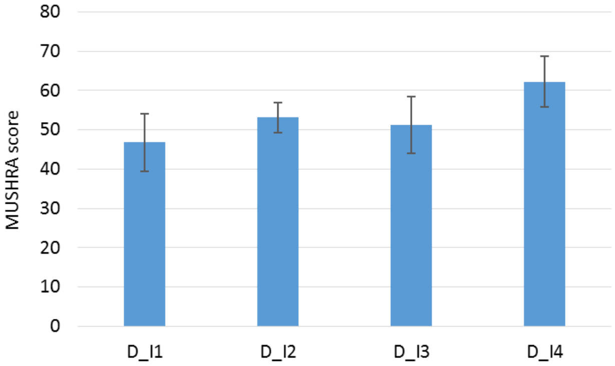


Figure. 8 MUSHRA scores for DNN based Myanmar TTS systems with different input features

5.6 Subjective Evaluation

Two MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor) listening tests [36] were conducted to subjectively evaluate the effectiveness of word vectors on DNN based and Hybrid-LSTM-RNN based Myanmar TTS systems presented in Section 5.5. 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 in each test and rate them using 0-100 scale on their naturalness. The rating scale are 0-20 (bad), 21-40 (poor), 41-60 (fair), 61-80 (good) and 81-100 (excellent). 20 synthetic speeches were randomly ordered and presented without labels in the tests.

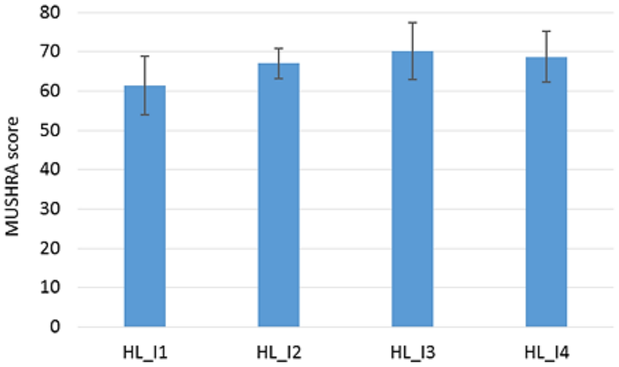


Figure. 9 MUSHRA scores for Hybrid-LSTM-RNN based Myanmar TTS systems with different input features

Fig. 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 Fig. 5. It can be concluded that the effectiveness of word vector can be seen in DNN based systems.

The MUSHRA scores of Hybrid-LSTM-RNN based Myanmar TTS systems with different input features are illustrated in Fig. 9. 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.

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-RNN based systems in Fig. 7. Some samples of synthesized speeches generated by DNN and Hybrid-LSTM-RNN based systems with different input features are given on the link [37].

6. Conclusion

In this paper, we investigated the effectiveness of word vectors on DNN, LSTM-RNN, and Hybrid-LSTM-RNN based Myanmar TTS systems. Word vectors for Myanmar language were also built by using the collected monolingual Myanmar corpus for better performance. The comparisons are done on modelling DNN, LSTM-RNN, and Hybrid-LSTM-RNN based Myanmar speech synthesis with and without additional input 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. According to the objective evaluation results, using word embedding features as the additional input features in acoustic modelling cannot lead to significant improvement of LSTM-RNN and Hybrid-LSTM-RNN based systems. However, the subjective evaluation results show that the native listeners more prefer the synthesized speeches of Hybrid-LSTM-RNN based systems using word vector and/or POS features as the additional input features than the system with conventional input features. From the results of this work, in term of naturalness, word vector features can give the effectiveness of all neural network based Myanmar TTS systems.

In our future work, word vectors will be learned by taking counts the acoustic information related to TTS task so that it can be 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 acoustic properties such as F_0 and duration.

References

[1] 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.

[2] O. Watts, J. Yamagishi, and S King, “Unsupervised continuous-valued word features for phrase-break prediction without a part-of-speech tagger”, In: *Proc. of Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[3] V. Klimkov, A. Nadolski, A. Moinet, B. Putrycz, R. Barra-Chicote, T. Merritt, and T. Drugman, “Phrase Break Prediction for Long-Form Reading TTS: Exploiting Text Structure Information”, In: *Proc. of INTERSPEECH 2017*, pp. 1064-1068, 2017.

[4] O. S. Watts, “Unsupervised learning for text-to-speech synthesis”, *Ph.D. dissertation, University of Edinburgh*, 2012.

[5] 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: *Proc. of Eighth ISCA Workshop on Speech Synthesis*, 2013.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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: *Proc. of INTERSPEECH 2015*, pp. 2237-2241, 2015.

[11] A. M. Hlaing, W. P. Pa, and Y. K. Thu, “Word-based Myanmar text-to-speech with CLUSTERGEN”, In: *Proc. of The 16th International Conference on Computer Applications*, pp. 203–208, 2018.

[12] A. M. Hlaing, W. P. Pa, and Y. K. Thu, “DNN based Myanmar speech synthesis”, In: *Proc. of the 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*, pp. 142–146, 2018.

[13] A. M. Hlaing, W. P. Pa, and Y. K. Thu, “Enhancing Myanmar Speech Synthesis with Linguistic Information and LSTM-RNN”, In: *Proc. of the 10th ISCA Speech Synthesis Workshop*, pp. 187–193, 2019.

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

[15] A. M. Hlaing, W. P. Pa, and Y. K. Thu, “Myanmar number normalization for text-to-speech”, In: *Proc. of International Conference of the Pacific Association for Computational Linguistics*, pp. 263–274, 2017.

[16] A. M. Hlaing and W. P. Pa, “Sequence-to-Sequence Models for Grapheme-to-Phoneme

Conversion on Large Myanmar Pronunciation Dictionary", In: *Proc. of Oriental COCOSDA 2019*, 2019 (*In Press*)

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

[18] http://www.cs.columbia.edu/˜ecooper/tts/lab_format.pdf

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

[20] T. Mikolov, Q. V. Le, and I. Sutskever, "Exploiting similarities among languages for machine translation", *arXiv preprint arXiv:1309.4168*, 2013.

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

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

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

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

[25] H. Riza, M. Purwoadi, T. Uliniansyah, A. A. Ti, S. M. Aljunied, L. C. Mai, V. T. Thang, N. P. Thai, V. Chea, S. Sam, S. Seng, "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*, pp. 1-6, 2016.

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

[27] 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*, 2013.

[28] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model", *Journal of Machine Learning Research*, pp. 1137-1155, 2003

[29] 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, 2009.

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

[31] 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.

[32] 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.

[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization", *arXiv preprint arXiv:1412.6980*, 2014.

[34] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system", In: *Proc. of SSW*, pp. 202-207, 2016.

[35] F. Chollet, "Keras: The python deep learning library", *Astrophysics Source Code Library*, 2018.

[36] ITU-R BS. 1534-1, "Method for the subjective assessment of intermediate quality level of coding systems", *International Telecommunication Union*, 2003.

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