Design and Implementation of Burmese Speech Synthesis System Based on HMM-DNN

Mengyuan Liu
School of Information Science and Engineering
Yunnan University
Kunming, China
e-mail: liumeyu@qq.com

Jian Yang
School of Information Science and Engineering
Yunnan University
Kunming, China
e-mail: jianyang@ynu.edu.cn

***Abstract*—The research and application of speech synthesis in Chinese and English are widely used. However, most non-universal languages have relatively few electronic language resources, and speech synthesis research is lagging behind. Burmese is a type of alphabetic writing, and Burmese belongs to Tibetan-Burmese branch of the Sino-Tibetan language. In order to develop the Burmese speech synthesis application system, this paper studies the Burmese speech waveform synthesis method, designs and implements a HMM-based Burmese speech synthesis baseline system, and based on this, introduces a deep neural network (DNN) to replace the decision tree model of HMM speech synthesis system, thereby improving the acoustic model to improve the quality of speech synthesis. The experimental results show that the baseline system is feasible, and the introduction of DNN speech synthesis system can effectively improve the quality of speech synthesis.**

Keywords-speech synthesis; HMM; acoustic model; decision tree; deep neural network

I. Introduction

With the development of information technology, the research of speech synthesis technology in Chinese and English has been relatively mature. However, due to the lack of linguistic resources in Burmese, the research on speech synthesis is lagging behind. Burmese is the official language of Myanmar and has a history of more than a thousand years, spoken by about 54 million people[1]. Similar to Chinese, Burmese has four tones.

The initial phase of the Burmese speech synthesis system is mainly focused on waveform concatenation techniques, such as Myanmar text-to-speech system with rule-based tone synthesis proposed by Kyawt in 2011[2]; Diphone-Concatenation speech synthesis for Myanmar Language proposed by Soe in 2013[3]. In 2017, Hlaing proposed using phoneme concatenation method Myanmar speech synthesis[4], which is the speech synthesis through waveform concatenation. This waveform concatenation technology based on large-scale corpus is extremely costly and cannot be effectively put into practice application. In 2015, Ye first proposed a Myanmar statistical parameters speech synthesis method based on HMM, but there are still many pronunciation errors in the synthesized speech[5], The HMM-based Burmese speech synthesis still needs to be further studied. In recent years, speech synthesis based on HMM-DNN and end-to-end speech synthesis based on DNN have gradually become the mainstream technology for developing application systems. Due to the end-to-end speech synthesis based on DNN requires high training sample size and computational ability[6], this method is not involved in this paper.

In order to develop the Burmese speech synthesis application system, this paper studies the Burmese speech synthesis method, designs and implements a HMM-based Burmese speech synthesis baseline system, completes Grapheme-to-Phoneme transcription, automatic phoneme segmentation, context attributes and question set design, speech synthesis and other work, and based on this, the DNN acoustic model is introduced to replace the decision tree model in the HMM speech synthesis system, which solves the limitations of some traditional acoustic models, thereby improving the quality of speech synthesis.

The content of this paper is organized as follows: The second part introduces the Burmese speech synthesis system based on HMM and DNN; the third part mainly introduces the design and implementation of speech synthesis system; the fourth part is the analysis of experimental results; the last part is a summary.

II. Speech Synthesis System based on HMM-DNN

A. HMM-based Speech Synthesis System

The typical speech synthesis system has two main parts, front-end text analysis and back-end speech synthesis. The front-end text analysis is mainly to get text, and performing text analysis to obtain the back-end training information. Text analysis mainly including the normalization, word segmentation and Grapheme-to-Phoneme(G2P), etc. The back-end speech synthesis is based on the information obtained by the front-end, through the training of the data model, the parameters are predicted, and then conduct speech synthesis. This paper mainly focuses on the back-end part of the speech synthesis system.

The framework of statistical parameter speech synthesis system based on HMM is shown as Fig. 1[7], the complete HMM speech synthesis system can be divided into two parts: training and synthesis. The training part mainly extracts acoustic feature parameters from the corpus, and then performs HMM modeling, and the model clustering and training are conducted based on labels information, context attributes and question set. The synthesis part is to perform text analysis of the text that need to be synthesized, predict the parameters according to the model of the training part, and then perform speech synthesis through the speech synthesizer.

In HMM-based systems, HMM modeling is required for each context attribute, but limited data cannot completely cover all context combinations, so clustering

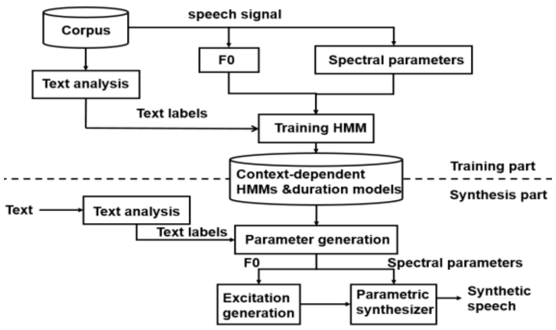

Figure 1. Speech synthesis system framework.

training based on decision tree is used in HMM speech synthesis systems. Although context clustering based on decision trees enables HMM speech synthesis systems to synthesize better quality speech, there are still some limitations[8]: First of all, decision tree clustering cannot solve complex context feature problems well. Solving this problem can also be represented by a large amount of data, but it violates the design purpose of small data training speech synthesis based on HMM system. Then there is the problem that the decision tree divides the training samples, each leaf node is independent of each other and cannot share parameters, the data of each leaf node will be very rare, and it cannot be used for other related models. This will make some rare linguistic features to be directly ignored, thus affecting the speech quality of speech synthesis system.

An effective way to solve the above problem is to replace the decision tree model with a DNN acoustic model. This method can not only solve complex context feature with high-dimensional data input[8], but also train all data to get the weight of each feature, which can effectively solve the problem of poor generalization ability of data.

B. Speech Synthesis System Based on HMM-DNN

DNN is an artificial neural network with many hidden layers between the input and output layers. The DNN simulates human speech generation through a hierarchical structure, and transforms language text information into the final speech output[9]. Due to the complexity of speech data, shallow model structures such as HMM have limited modeling capabilities and cannot capture high-order correlations between data features very well. The powerful modeling ability of DNN is more suitable for modeling complex speech data, and its deep structure can compactly represent large-span, highly complex features. At the same time, DNN considers context dependent speech data features, and can improve the accuracy of modeling by using high-dimensional feature vectors with high discrimination. Therefore, we use the DNN acoustic model to replace the original acoustic model structure, which can better improve the quality of synthesized speech.

The Burmese speech synthesis framework based on the DNN acoustic model is shown in Fig. 2[8]. First, the input text is converted into a context labeling, and then the system converts into input sequence according to the context labeling. The input sequence is a binary sequence answered by the corresponding question set. The input sequence is trained and mapped by the network forward algorithm to obtain the output sequence. The output sequence contains various acoustic features and their dynamic parameter[10], where the weights are trained by the training data. The input and output are frame-to-frame forced alignment by the trained HMM model and then put into the network for training. Through the training of the network, the best matching model of input and output is obtained. Like the HMM, the DNN acoustic model also generates speech parameters. By setting the predicted output feature from the DNN as mean vector and setting the pre-calculated output feature variance from all training data as covariance matrix, the speech parameter generation algorithm can generate the acoustic parameters, and then uses the generated parameters for speech synthesis.

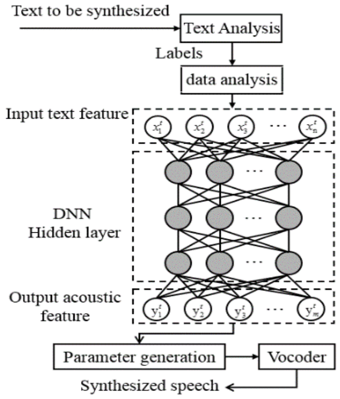

Figure 2. Speech synthesis framework based on DNN acoustic model.

It can be seen that the text analysis, speech parameter generation and waveform synthesis modules of the DNN-based system can be shared with the HMM-based system, and the DNN-based system only needs to replace the part from the context labeling to the decision tree clustering.

III. DESIGN AND IMPLEMENTATION

A. Grapheme-to-Phoneme Conversion

Burmese is a complex language, and its writing is very special. In the process of speech synthesis, in order for the computer to correctly read the Burmese language, Burmese text needs to be translated into Roman text[11]. The G2P transcription is to translate the Burmese text into a Roman alphabet text according to their pronunciation in Burmese. The transcription of this paper is based on the syllable structure of Burmese. The Burmese syllables can be divided into initials and finals. The initials are composed of consonants and complex consonants, the complex consonants are composed of consonants and medial. The finals are composed of vowels.

There are stacked words and pronunciation change in Burmese, after consulting Burmese experts and reading a large number of documents, we summarized the rules of the stacked words and pronunciation change. According to these rules, the corresponding program is designed to deal with these phenomena to improve the accuracy of the transcription.

The transcription scheme used in this paper is mainly designed with reference to the international phonetic of the Burmese alphabet and some transcription schemes proposed by Burmese scholars. The transcription process is carried out in conjunction with a Burmese dictionary (26600 words) published by the Myanmar Language Association.

The whole transcription process is as follows: The sentence is divided into words. Look for the word in the dictionary, if there is, use the transcription in the dictionary directly. If there is no such word in the dictionary, then the transcription will be performed according to the rules. First, the stacked words are processed according to the rules, then the words are divided into syllables, and the syllables are further parsed into consonant, vowel and medial, and then judged whether they belong to four kinds of pronunciation change. If so, the syllables are changed according to the rules. Then, according to the transcription scheme, the consonant, vowel and medial are transcribed separately, then combined them according to the spelling rules. The generated Roman alphabet text is the G2P transcription text. The transcription correctness rate of the 4000 sentences of Burmese text used in this paper reached 93.6%, which met the conditions for developing the Burmese speech synthesis system. This is a G2P transcription example:

Burmese text: ရာသီဥတု (weather)
G2P transcription text: ja3dhi3u1du1

B. *Selection of Synthesis Unit*

In the process of speech synthesis, it is important to choose the appropriate synthesis unit. Through reading a large number of documents and consulting Burmese experts, Burmese syllable structure is similar to Chinese. Therefore, referring to the structure of the Chinese syllable structure, the Burmese syllables are divided into initials and finals, and the initials and finals are chosen as the synthesis unit of Burmese. Burmese is a tonal language, there are 50 finals and 66 initials in total.

After the synthesis unit is determined, the prosody text can be generated. The prosody text includes all the work of the front-end text analysis, which is very important for the back-end speech synthesis. The prosody text contains prosody information such as word boundaries, syllable boundaries, and phoneme boundaries.

C. *Automatic Phoneme Segmentation*

The automatic segmentation of the phoneme is a process of dividing the initialized monophone label file into more accurate label files with phoneme duration information. The automatic segmentation of phonemes can be divided into three stages:

The first stage: mainly preparing training data and training models. This paper selects 4000 Burmese sentences, including audio and text. Initialized monophone file is generated combine audio and prosody text. Acoustic feature parameters are extracted from the audio using the HTK tool. Then, according to the initial HMM model of each phoneme, the global mean and variance are obtained by HCompV[7]. HInit reads in all the initialization training data, and the mean and variance are re-estimated. Finally, HRest re-estimates the parameters of HInit estimation.

The second stage: HHEd modifies the HMM model corresponding to each phoneme and the sentence according to the parameters estimated in the first stage, and then the model is re-estimated as training set by HERest. After a total of 5 times training, the second stage of parameter revaluation and model building was completed.

The third stage: using HVite to force alignment to get the segmentation label file. Finally, a monophone labeling file with accurate duration information can be obtained.

D. *Context Attributes and Question Set Design*

In continuous speech, each phoneme interacts with each other and there is a common phenomenon of coarticulation. The context attribute set is the coarticulation attribute labeling set of each central phoneme and the front and back phonemes. In the process of speech synthesis, context dependent model training is performed in the training stage to obtain the optimal parameters. In the synthesis stage, the state prediction is carried out through the model and the context attribute label of the input text. Therefore, the context attribute label is important. The label data include the positions of speech units, tones, prosody information, etc. This paper designs a context attributes set based on the language characteristics and pronunciation style of Burmese language, realize the process of automatically converting the Burmese text into the corresponding label file.

Because the number of the phoneme model in the training data is limited, and the number of context dependent models is large, there will be overfitting problems in the training process. In addition, in the synthesis stage, the context attributes of the text that need to be synthesized have not appeared in the training data, the synthesis cannot be completed. In order to solve these problems, the system uses the decision tree clustering algorithm to carry out model clustering, and combines the question set to perform the contextual model clustering training[12]. Decision tree cluster training based on question in the question set. So, the question sets are designed to combine the pronunciation features of Burmese.

E. *Training and Synthesis*

1) HMM speech synthesis system: In the study of this paper, HMM speech synthesis main modeling parameters are configured as follows: Mel-cepstrum parameters, fundamental frequency parameters as acoustic parameters; initials and finals as the basic modeling unit; 5-state HMM model with no-skip structure from left-to-right is adopted.

The overall design of the Burmese speech synthesis system based on HMM is divided into three stages:

In the first stage, the HTK tool is used to extract the speech acoustic parameters, including the spectral parameters and the fundamental frequency parameters. Then, the monophone list, the monophone time label, the context attribute label and question sets are stored in the specified folder of the system.

In the second stage, the acoustic model training is mainly carried out, and the system mainly calls the HTK toolkit to complete the training. The decision tree model clustering of Burmese speech synthesis system is based on context attributes and question sets, and the acoustic parameters and triphone context dependent HMM model are obtained.

The third stage is to synthesis. First, text analysis is performed on the Burmese sentence that needs to be synthesized, and the text analysis result is converted into a corresponding context dependent labels, then putting them into the synthesis system. Finally, according to the parameters model, the corresponding acoustic parameters are generated by the parameter generation algorithm, and the speech is synthesized by the parameter synthesizer.

2) HMM-DNN speech synthesis system: Based on the HMM-DNN Burmese speech synthesis system, the DNN acoustic model is used to replace the traditional decision tree model to train the acoustic model. According to the features of the input and output sequence setting, and the study of the DNN acoustic model, the DNN training network parameters configuration in this paper is shown in Table I.

TABLE I. DNN ACOUSTIC MODEL NETWORK SETTINGS FOR BURMESE SPEECH SYNTHESIS SYSTEM

DNN acoustic model network parameters	Parameter value (function)
Batch size	256
Hidden layer activation function	Sigmoid
Learning rate	0.001
Number of hidden layers	3
Number of hidden layer units	2048
Number of input layer units	483
Number of output layer units	109
Optimizer	Adam

In the DNN acoustic model training, the composition of the input and output data is related to the quality of the entire acoustic model. The input sequence is a 483-bit vector sequence transformed from the Burmese question set and the context attribute set. The DNN training model is a frame-to-frame correspondence, each sequence represents the corresponding text feature of a frame[13]. Mainly the various feature information is extracted from the context attribute labeling corresponding to the frame, including phonemes and part of speech and other information. The sequence value includes two types. The first type is the binary value of answering questions in question set, for example, is the current phoneme `k'? The second type is a specific value, such as the position of the phoneme in the current syllable.

Then the input sequence is mapped to the output sequence through the forward algorithm training. The output sequence of the training data is a 109-bit vector sequence, which mainly includes spectral parameters, fundamental frequency features, and corresponding one-order and second-order differences.

In the synthesis stage of the DNN acoustic model, the corresponding context label is generated by the text to be synthesized, and then converted into a binary sequence, the best parameter sequence is predicted by trained DNN model. The predicted output feature sequence is treated as a mean vector, and the variance of the output features constitutes a covariance matrix, thereby generating parameters, and then using the generated parameters for speech synthesis.

IV. EXPERIMENTAL

A. Data Preparation

In this paper, the audio is recorded by a professional female broadcaster whose native language is Burmese. The Burmese pronunciation corpus used in this experiment is a total of 4000 sentences, the duration is about 6.6 hours, the total size of wav audio files is about 2 GB, the Burmese text has a corpus size of 628 KB, the audio sample rate is 48 KHz.

B. Phoneme Segmentation Experimental Results and Analysis

According to the knowledge of the spectrogram, the vowel has a clear first and second formant. The red line in the below figure is the vowel formant. It can be seen that the vowel position is basically correct, and occasionally there is a deviation in the position of the consonant, and the result of the phoneme segmentation is generally correct. We randomly selected 100 sentences and obtained the correct rate of segmentation is 83.6%, which basically meets the requirements of the speech synthesis system.

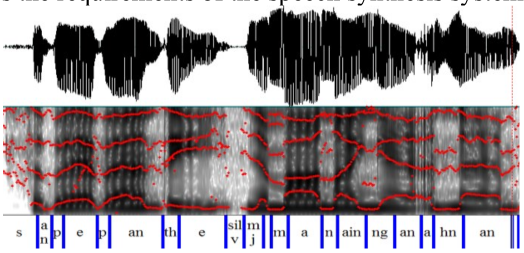

Figure 3. Phoneme segmentation result.

C. Speech Synthesis Experiment Results and Evaluation

1) Objective evaluation: The following is a waveform of a sentence (Burmese: စံပယ်ပန်းသည် မြန်မာနိုင်ငံအနှံ့ အပြားတွင် ပေါက်ရောက်သည်. Translation: Osmanthus fragrans is grown and opened in many places in Myanmar), the first is the original speech waveform, the second is the synthetic speech waveform based on HMM, and the third is the synthetic speech waveform based on HMM-DNN. It can be seen that although the shape of the original waveform is basically restored, the waveform of based on HMM-DNN is closer to the original speech waveform.

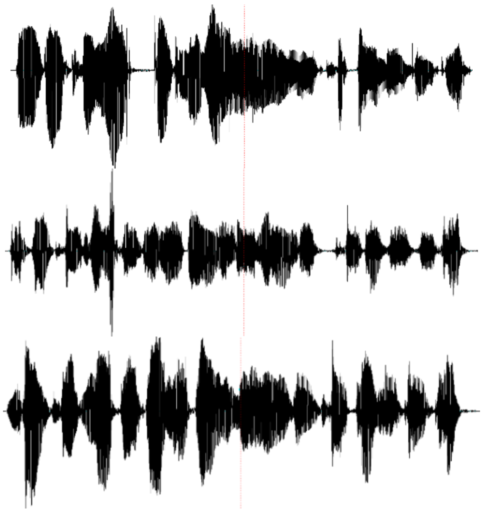

Figure 4. Speech synthesis comparison of speech waveforms.

The following is the spectrogram, the first is the synthetic speech spectrogram based on HMM, the second is the synthetic speech spectrogram based on HMM-DNN, it can be seen that the phoneme boundary of the synthetic speech based on HMM-DNN is clearer, the transition between phonemes is more natural.

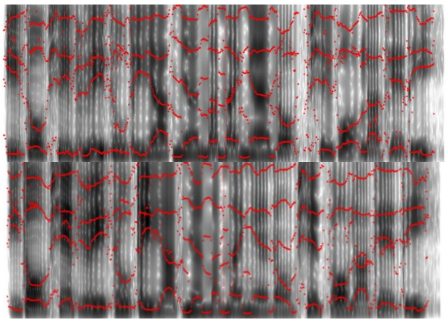

Figure 5. Speech synthesis comparison of spectrograms.

2) *Subjective evaluation:* The speech synthesis system is a key part of human-computer communication. The quality and naturalness of synthetic speech, that is, the subjective feelings that people hear, are critical. Five students studying speech synthesis are invited to make a comparative analysis of the synthesized and original speech (20 sentences), and subjectively describe the naturalness perception of different speeches according to the naturalness level shown in the table below:

TABLE II. NATURALNESS LEVEL

Level	Speech naturalness
5	Very natural
4	More natural
3	Acceptable
2	Less natural
1	Unacceptable

In the naturalness test, the synthesized speech and the original speech are subjectively described, and the subjective description results are statistically analyzed. The results are shown in the Fig 6:

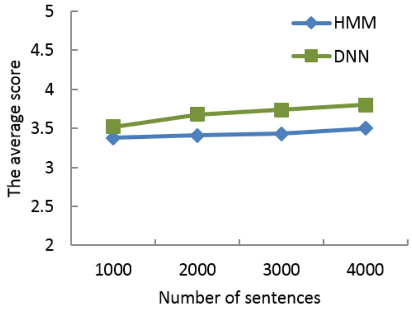

Figure 6. Evaluation results.

From the above subjective evaluation, we can see that the Burmese speech synthesis baseline system based on HMM has certain feasibility. The score of the speech synthesis based on HMM increases with the increase of the data amount, but the amplitude is not large. The score of the speech synthesis based on HMM-DNN increases more with the increase of the data amount. And the overall score is higher than that of HMM-based speech synthesis. It can be said that the introduction of DNN speech synthesis system effectively improves the quality of speech synthesis.

V. CONCLUSION

This paper focuses on the design and implementation of the Burmese speech synthesis baseline system based on HMM, and completes the Grapheme-to-Phoneme transcription, automatic phoneme segmentation, context attributes and question set design, and the speech synthesis. Based on this, the DNN acoustic model is introduced to replace the decision tree clustering model in HMM speech synthesis, which solves some limitations of traditional decision tree clustering and improves the quality of speech synthesis. The experimental results show that the quality of synthesized speech based on HMM has reached a general level, and the quality of speech synthesis introduced into the DNN acoustic model is improved on the basis of HMM. However, the synthesized speech is still lacking in rhythm. Later, we can consider adding more rhythm information to make the synthesized speech more natural. In addition, it is also possible to consider further improve the acoustic model by using a more advanced network structure to better improve the quality of speech synthesis.

ACKNOWLEDGMENT

This work is supported by the Natural Science Foundation of China (No.61961043).

REFERENCES

[1] Wang. D, “Burmese Language Tutorial (No.1),” Beijing University Publishing House, 2012. (in Chinese).

[2] Kyawt Y, and Tomio T, “Myanmar text-to-speech system with rule-based tone synthesis,” Acoustical Science and Technology, vol. 32, no. 5, 2011, pp. 174–181.

[3] Ei.P, Aye T, “Diphone-Concatenation speech synthesis for Myanmar Language,” International Journal of Scientific, Engineering and Technology Research, vol. 2, No. 4, 2013, pp. 1078–1087.

[4] Hlaing C, Thida A, “Myanmar speech synthesis system by using phoneme concatenation method,” International Conference on Signal Processing and Communication. 2017.

[5] Ye K, Win P, Jinfu N, Yoshinori S,Andrew F, Chiori H, Hisashi K, Eiichiro S, “HMM Based Myanmar Text to Speech System,” INTERSPEECH 2015.

[6] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, et al, “Tacotron: Towards end-to-end speech synthesis,” 2017.

[7] Tokuda K, Nankaku Y, Toda T, et al, “Speech Synthesis Based on Hidden Markov Models,” Proceedings of the IEEE, 2013, 101(5):1234-1252.

[8] Zen H, Senior A, “Schuster M. Statistical parametric speech synthesis using deep neural networks,” IEEE International Conference on Acoustics. IEEE, 2013.

[9] Zhen-Hua Ling, Shi-yin Kang, Heiga Zen, et al, “Deep Learning for Acoustic Modeling in Parametric Speech Generation: A systematic review of existing techniques and future trends,” J. IEEE Signal Processing Magazine, 2015, 32(3):35-52.

[10] Qian Y, Fan Y, Hu W, et al, “On the training aspects of Deep Neural Network (DNN) for parametric TTS synthesis,” 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) ,2014:3829-3833.

[11] Thu Y, Ye K, Win P, Jinfu N, “Syllable Pronunciation Features for Myanmar Grapheme to Phoneme Conversion,” the 13th International Conference on Computer Applications. 2015:161-167.

[12] Tokuda K, Zen H, Black A W, “An HMM-based speech synthesis system applied to English”. IEEE Workshop on Speech Synthesis. 2013.

[13] Hashimoto K, Oura K, Nankaku Y, et al, “The effect of neural networks in statistical parametric speech synthesis,” 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),2015:4455-4459.