Development of a Vietnamese Speech Synthesis System for VLSP 2019

Van Thinh Nguyen, Tien Thanh Nguyen, My Linh Nguyen Thi and Van Hai Do
Viettel CyberSpace Center
{thinhnv20, thanhnt116, linhntm18, haidv21}@viettel.com.vn

***Abstract*—This paper describes our deep neural network-based speech synthesis system for high quality of Vietnamese speech synthesis task. The system takes text as input, extract linguistic features and employs neural network to predict acoustic features, which are then passed to a vocoder to produce the speech wave form.**

***Index Terms*—Speech Synthesis, Deepneural Network, Vocoder, Text Normalization.**

I. Introduction

The fifth International Workshop on Vietnamese Language and Speech Processing (VLSP 2018) organizes the shared task of Named Entity Recognition, Sentiment Analysis, Speech Recognition and Speech Synthesis at the first time for Vietnamese language processing. The goal of this workshop series is to attempt a synthesis of research in Vietnamese language and speech processing and to bring together researchers and professionals working in this domain.

In this paper we describe the speech synthesis system which we participated in the TTS track of the 2018 VLSP evaluation campaign.

II. System Architecture and Implementation

A. System Architecture

With the aim of improving the naturalness and intelligibility of speech synthesis system, we propose apply new technologies of speech synthesis for acoustic modeling and waveform generation such as using deep neural network combined with new type of vocoder. To archive to this goal, a new architecture of speech synthesis system is proposed in figure 1. Our proposed system takes text as an input and normalize it into standard text which is readable. Linguistic feature extraction is then applied to extract text's linguistic features as an input for acoustic model. For vocoder parameter generation, Acoustic model used to take given input linguistic feature and generate predicted vocoder parameter. The speech waveform is generated by vocoder, which is new type.

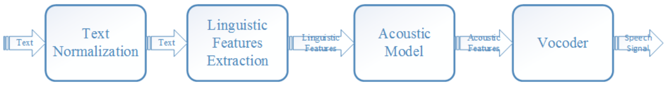

Fig. 1. The proposed speech synthesis system [1]

B. Front End

1) Text Normalization: Text Normalization plays an important role in a Text-To-Speech (TTS) system. It is a process to decide how to read Non Standard Words (NSWs) which can't be spoken by applying letter-to-sound rules such as CSGT (cnh st giao thng), keangnam (cang nam). The process decides the quality of a TTS system. The module implemented and based on using regular expression and using abbreviation dictionary . Regular expression is a direct and powerfull technique to clasify NSWs. We build expressions that describe the date, time, scrore, currency and mesuarment. An abbreviation dictionary containing foregin proper names, acronyms...[2]

2) Linguistics Features Extraction: Linguitics Feature, which was used as input features for the system, had been extracted by generating a label file from linguistics properties of the text (Part-of-Speech tag, word segmentation, and text chunking) and mapping the corresponding information to binary codes presented in a question file. Each piece of information was encoded into an one-hot vector, which was later concatenated horizontally to form a single one-hot vector presenting the text.

C. Acoustic Modeling

Acoustic model is based on deep neural network, specially it is feedforward neural network with enough layers, a simplest type of network. The architecture of network is shown in figure 2. follow this network, The input linguistic features used to predict the output parameter via several layers of hidden units [1]. Each node of the network is called perceptron and each perceptron perform a nonlinear function, as follow:

$$h_t = H(W^{xh}x_t + b^h)$$

$$y_t = W^{hy}h_t + b^y$$

Where H(.) is a nonlinear activation function in a perceptron (in this system, we use TANH function for each unit) [3], W^{xh} and W^{hy} are the weight matrices, b^x and b^y are bias vector.

D. Vocoder

Currently, The speech synthesis system use many type of vocoder and most of vocoder are based on source filter model [4]. In our system we used a vocoder-based speech synthesis system, named WORLD, which was developed in an effort to improve sound quality of real-time application using speech [5]. WORLD vocoder consist of three algorithm for obtaining

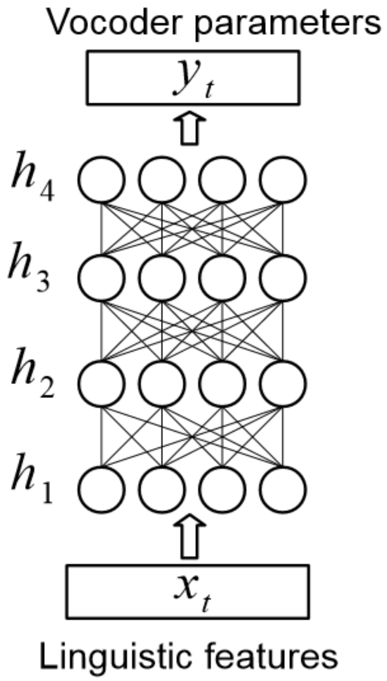

Fig. 2. The feedforward neural network for acoustic modeling

three speech parameters, which are F0 contour estimated with DIO [6], spectral envelop is estimated with CheapTrick [7] and excitation signal is estimated with PLATINUM used as an aperiodic parameter [8], and a synthesis algorithms for obtaining three parameter as an input. With WORLD vocoder, speech parameter predicted from acoustic model which correspond to input text sentence, will be used for produce speech waveform.

III. Experimental Setup and Results

A. Data Preparation

In this section, we describe our effort to process VLSP dataset with include two difference corpus. Big corpus have more than 22 hours of recording audio which have much more noise type, and Small corpus have 45 minute of high quality of recording audio. The noise in the silience of signal is removed, the silience is added into start and end of signal. With both corpus, the transcript is corrected by our phonetic dictionary.

B. Experimental Setup

To demonstrate how we archive high quality of speech synthesis, we report experimental setup for this architecture. we used speech corpus collected in previous section. With Big corpus, 13,703 utterances were used for training, 614 as a development set, and 614 as the evaluation set. With Small corpus 740 utterances were used for training, 80 for test set.

The input features for neural network, is extracted by front-end, consisted 632 features. 623 of these derived from linguistic context, including phoneme identity, part of speech and positional information within a syllable, word, phrase, etc. The remain 9 features are within phoneme positional information. The speech acoustic features extracted by WORLD vocoder for both training and decoding. Each speech feature vecor contain 60 dimensional Mel Cepstral Coefficients (MFCCs), 5 band aperiodicities (BAPs) and fundamental frequency on log scale (logF0) at 5 milliseconds frame intervals.

Deep neural network is configured with 6 feedforward hidden layers and each layer has 1024 hyerbolic tangent unit.

C. Results

The objective result of the system trained with Big corpus and Small corpus are presented in table 1. it shown that, MCD: Mel cepstral distortion [9], BAP: distortion of band aperiodicities and V/UV: voice/unvoice error are quite low, that mean we has traned good acoustic model which return the best result. F0 RMSE is caculated on linear scale.

TABLE I
The objective result of speech synthesis system

	MCD (dB)	BAP (dB)	F0 RMSE (Hz)	V/UV(%)
Big corpus system	5.379	0.154	34.497	25.212
Small corpus system	5.661	0.431	27.480	10.978

The subjective results presented in table 2. this table show mean opinion score of two system and score of natural audio.

TABLE II
The subjective results

	Big corpus	Small corpus
MOS of TTS system	2.3	3.3
MOS of Natural speech	4.3	4.5

IV. Conclusion

In this paper our speech synthesis system based deep neural network is shown, the deep nearal network applied in noise data shows a worse quality than clean data even the amount of data is much larger. the result the text to speech system based on deep neural network still poor compared to natural speech, therefore in the future we will implement better technology like End to End speech systhesis based sequence to sequence model to archive better quality.

V.

References

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

[2] D. A. Tuan, P. T. Lam, and P. D. Hung, “A study of text normalization in vietnamese for text-to-speech system.”

[3] J. Jantzen, “Introduction to perceptron networks,” *Technical University of Denmark, Lyngby, Denmark, Technical Report*, 1998.

[4] F. Espic, C. Valentini-Botinhao, and S. King, “Direct modelling of magnitude and phase spectra for statistical parametric speech synthesis,” *Proc. Interspeech, Stochohlm, Sweden*, 2017.

[5] 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.

[6] M. Morise, H. Kawahara, and H. Katayose, “Fast and reliable f0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech,” in *Audio Engineering Society Conference: 35th International Conference: Audio for Games*. Audio Engineering Society, 2009.

[7] M. Morise, “Cheaptrick, a spectral envelope estimator for high-quality speech synthesis,” *Speech Communication*, vol. 67, pp. 1–7, 2015.

[8] ——, “Platinum: A method to extract excitation signals for voice synthesis system,” *Acoustical Science and Technology*, vol. 33, no. 2, pp. 123–125, 2012.

[9] J. Kominek, T. Schultz, and A. W. Black, “Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion,” in *Spoken Languages Technologies for Under-Resourced Languages*, 2008.