Investigating Deep Neural Network Adaptation for Generating Exclamatory and Interrogative Speech in Mandarin

Yibin Zheng[1,3]*, Ya Li*[1]*, Zhengqi Wen*[1]*, Bin Liu*[1]*, Jianhua Tao*[1,2,3]

[1]National Laboratory of Pattern Recognition,
[2]CAS Center for Excellence in Brain Science and Intelligence Technology,
Institute of Automation, Chinese Academy of Sciences, 100190, Beijing, China
[3]School of Computer and Control Engineering, University of Chinese Academy of Sciences
{yibin.zheng, yli, zqwen, liubin, jhtao}@nlpr.ia.ac.cn

Abstract

Currently, most speech synthesis systems generate speech only in a reading style, which greatly affects the expressiveness of synthetized speech. To improve the expressiveness of synthetized speech, this paper focuses on the generation of exclamatory and interrogative speech for Mandarin spoken language. We propose a multi-style (exclamatory and interrogative) deep neural network-based acoustic model with a style-specific layer (can have multiple layers) while other layers are shared. The style-specific layer is used to model the distinct style specific patterns. The shared layers allow maximum knowledge sharing between the declarative and multi-style speech. Such method is also validated on different neural networks. Besides, models that adapted on both spectral and F0 parameter is compared with models while only F0 parameter is adapted. Both subjective and objective evaluations show that this method is superior to prior work (which is trained by the combination of constrained Maximum likelihood linear regression (CMLLR) and structural maximum a posterior (SMAP)), and the BLSTM where top 1 layer is style-specific layer that adapted on both spectral and F0 parameter can achieve the best results.

Index Terms: Speech synthesis, F0 contour, deep neural network adaptation, exclamatory speech, interrogative speech

1. Introduction

Recently unit-selection and concatenative approaches [1-4] produce high-quality synthetic speech, but require large-scale speech corpora if the speech is to sound natural. Using these approaches to develop a humanlike speech synthesizer, which could control many kinds of emotional expressions and speaking styles, we would have to prepare many corpora corresponding to the different styles. These approaches are thus unsuitable to quick addition of several new emotional expressions and speaking styles to a speech synthesizer [5] and are particularly impractical when we need to reproduce intermediate degrees of emotional expressions and speaking styles. Moreover, the recording training data is often at great expense, the available training data is always very limited, especially for some of the speaking styles, such as exclamatory and interrogative. To deal with the lacking training data problem, a statistical parametric speech synthesis (SPSS) system [6]-[8] is proposed, which can easily and flexibly generate natural sounding synthetic speech with varying speaking styles and/or emotional expressions.

For Hidden Markov model (HMM) SPSS based speech synthesis, it has been already shown that the emotional expressions and speaking styles of synthetic speech can be easily reproduced, controlled, and transformed by using style modeling [9], model adaptation [10], model interpolation and model morphing [11], or multiple-regression HMMs [12].

Employment of deep neural networks (DNNs) has led the research of SPSS to a new stage [13-16]. For DNN-based SPSS, the concept of speaker and language factorization has been introduced in the multi-speaker and language DNN [17], in which all speakers and languages share the same hidden layers, each speaker has a speaker-specific output layer and each language has a language-specific layer. In multi-speaker and multi-language DNN, network is decomposed into three parts: language-specific layers; shared layers and specific-specific layers. Language-specific layers exploiting with multiple speakers' data in specific languages and speaker-specific layers populated with multiple languages' data from specific speakers, thus model becomes more robust than separated and independent modelling. Shared layers serve as a bridge to connect the language and speaker-specific layers. Also, the DNN model has been replaced with bidirectional long short term memory recurrent neural network, Yu et al. proposes a multilingual BLSTM [18], in which the hidden layers are shared across different languages while the input and output layers are language-dependent.

Motived by the use of multi-speaker DNN and multi-language DNN [17], we propose a multi-style (exclamatory and interrogative) deep neural networks acoustic model with a style-specific layer while other layers are shared. The style-specific layer is used to model the distinct style specific patterns. The shared layers allow maximum knowledge sharing between the declarative and multi-style sentences. In the multi-style deep neural networks acoustic model, different neural networks types are tested, such as DNN, RNNs and its varieties BLSTM; Then adaptation on both spectral and F0 parameter is compared with models while only F0 parameter is adapted; Besides, we also compare the effect of the number of style-specific layers in the network.

2. Previous works

During the past several years, speech synthesis technologies have achieved great improvement, but they still fall short of expressiveness in their ability to convey information via different speaking styles. To make synthesized speech more expressive, some methods have been tried, including the methods mentioned above and some other methods like rule-based [19] and CART model [20]. However, most of current investigation about expressive speech still focuses on emotional speech. Some spontaneous styles like exclamatory and interrogative speech have seldom been touched.

In the past few years, some researches have been conducted on the generation of exclamatory and interrogative speech based on HMM SPSS, in which “average style voice” is created from a large corpus of declarative sentences and adapted with a small amount of speech data from a target speaking style [21-24]. This research started by transforming the spectral parameters of speech [17][25] by using several adaptation techniques developed for automatic speech recognition such as maximum-likelihood linear regression (MLLR) [26] or MAP-VFS, which is an algorithm combining maximum a posteriori (MAP) adaptation and vector field smoothing (VFS). Then, to simultaneously model and adapt the excitation parameters of speech as well as spectral parameters, the multi-space probability distribution (MSD) HMM and its MLLR adaptation algorithm [22], [27] have been used. Furthermore, to simultaneously model and adapt duration parameters for the spectral and excitation parameters, the MSD hidden semi-Markov model (MSD-HSMM) [28] and its MLLR adaptation algorithm have been used. We have also replaced the MLLR adaptation algorithm with constrained structural maximum a posteriori linear regression (CSMAPLR) [29, 30] and achieved superior performance than other adaptation algorithms. Such system that adapted by using CSMAPLR is adopted as one of the baseline system, which is the combination of constrained MLLR (CMLLR) and structural MAP (SMAP).

3. Proposed approaches

Neural networks have re-emerged as a potential powerful acoustic model for SPSS. In [31], feed-forward neural networks (FNN) are employed to map a linguistic representation derived from input text directly to acoustic features. However, the temporal sequence nature of speech is not explicitly modeled in the FNN architectures. In [14], a BLSTM was employed to map a sequence of linguistic features corresponding sequence of acoustic features and achieved the state-of-the-art performance. In this paper, FNN and BLSTM would both be investigated on our proposed models. The overall architecture between FNN and BLSTM is the same except the different neural networks types they employed. For convenience, we only present our proposed models based on BLSTM.

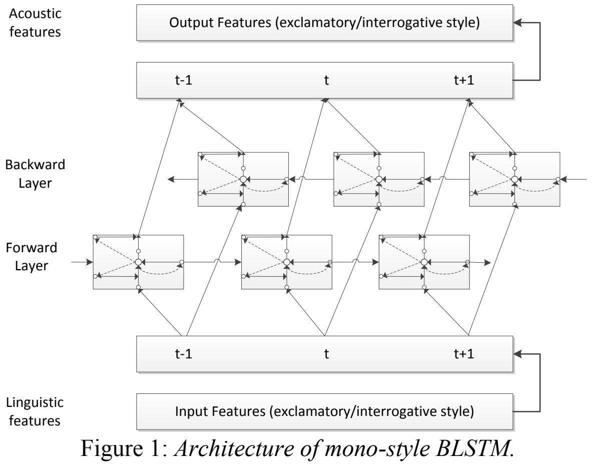
Acoustic features
Output Features (exclamatory/interrogative style)
t-1
t
t+1
Backward Layer
Forward Layer
t-1
t
t+1
Linguistic features
Input Features (exclamatory/interrogative style)

Figure 1: *Architecture of mono-style BLSTM.*

3.1. Mono-style BLSTM

The BLSTM model can be viewed as a sophisticated transformation model that jointly learns the relationships between linguistic features and output acoustic features. Figure 1 shows mono-style BLSTM models of speaking styles of exclamatory and interrogative, respectively. We call each model as the mono-style BLSTM, where the input linguistic features and the corresponding output acoustic features of a BLSTM come from single speaking style.

3.2. Multi-style BLSTM

3.2.1. Model structure

Inspired by the successful employment of multi-speaker and multi-language DNN, we keep the hypothesis that DNN-based SPSS can benefit from a large of declarative sentences and solve the adaptation problem by sharing the hidden layers among different speaking styles. With this assumption, we propose a multi-style BLSTM that shares hidden layers across different speaking styles.

Figure 2 shows the architecture of the proposed multi-style BLSTM where different speaking styles have their own output layers and related weight matrices from shared layers. In this framework, BLSTM is structured in two major layers: shared layers and style-specific layers. Each layer can have multiple hidden layers. The shared layers are style-independent and are shared across all speaking styles in the training corpus, as a global linguistic feature transformation universal to all the specking styles. Conversely, style layers are built to have a specific transformation for each speaking style. The linguistic features will first go through the shared layers and then predict the style-specific acoustic features with corresponding style-specific layers.

Formally, the multi-style DNN for speaking style l, denoted as $F_l(\cdot)$, can be decomposed to

$$y = F_l(x) = S_l(H(x)) \quad (1)$$

where x is the linguistic input feature vector; y is the acoustic output feature vector; and $H(\cdot)$, $S_l(\cdot)$ are shared layers and the style-specific layers for speaking style l respectively.

In this framework, multi-style BLSTM is trained with speech data from both low-resource speaking style and resource-rich speaking style. The cross-style information captured by the shared layers of multi-style BLSTM leads to better performance than the mono-style BLSTM.

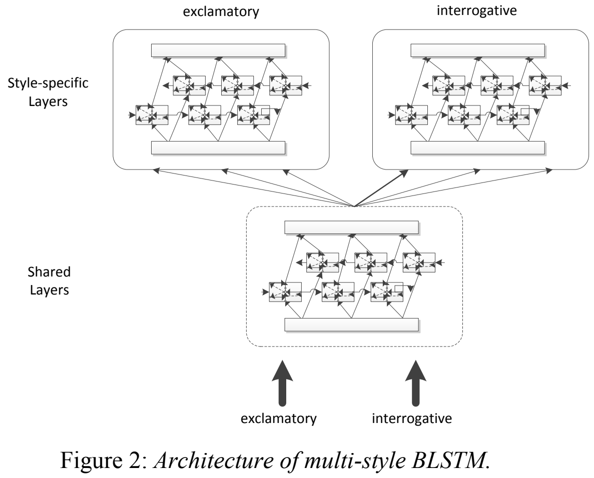

Figure 2: *Architecture of multi-style BLSTM.*

3.2.2. Model training

Training of the multi-style BLSTM is still based on the mini-batched stochastic gradient decent (SGD) algorithm as the mono-style BLSTM. For each linguistic and acoustic feature pair of speaking style l, only the related parts' gradient in the network will be computed for update.

4. Experiments

4.1. Experiment setup

A database which contains three speaking styles (declarative, exclamatory and interrogative) is adopted in this paper. This database is all recorded by a professional female broadcaster. Table 1 shows the statistics of the utterances numbers for each speaking styles. The declarative corpus contains 6,000 sentences of speech utterances, with each utterance having around 13 words. The exclamatory and interrogative corpora both contain 600 utterances, with average 15 words for each utterance. For exclamatory and interrogative utterances, 550 utterances are used for as the training set for SPSS, while the remaining 50 utterances serve as the test set. The speaking styles of exclamatory and interrogative are used for adaptation.

Table 1. Number of *utterance in each corpus of different speaking styles.*

Style	*Numbers*
declarative	6000
exclamatory	600
interrogative	600

All speech recordings are sampled at 16 kHz, windowed by a 25 ms windows, and shifted every 5 ms. 40th-order line spectral pair (LSP) coefficients, the fundamental frequency (F0) in log scale together with their delta and delta-delta deviation, and voiced/unvoiced (V/UV) flag are extracted, which serve as the acoustic features parameters for SPSS. As for the linguistic features, the phonetic and prosodic contexts of Mandarin include, the phone identity, the position of a phone, syllable and word in phrase and sentence, POS of word, prosodic phrase, intonational phrase and sentence, the length of prosodic word, prosodic phrase, intonational phrase and sentence, etc. totally 695 dimension.

To conduct the experiments, 4 major kinds of SPSS systems are to be compared, including mono-style BLSTM, multi-style BLSTM, multi-style BLSTM-F0 (only excitation parameters are adapted) and the HMM-based system that simultaneously model and adapt the excitation parameters as well as spectral parameters by using CSMAPLR [29, 30].

In HMM-based system, each HMM phone model is seven-state, left-to-right topology with single Gaussian, diagonal covariance distributions. Totally 6,000 declarative utterances are used for training the "average style model". In the style-adaptive training, CSMAPLR adaptation is employed and the estimation of multiple transforms is based on the shared decision trees constructed in the training stage of "average style model".

In mono-style BLSTM based SPSS, the exclamatory and interrogative corpus is used to train their own mono-style BLSTM directly without any adaptation technology. The input feature dimension is 695 and the output features contains a voiced/unvoiced (U/UV) flag, log-gain, log-F0, LSP and their delta and delta-delta deviation, totally 127 dimensions. An exponential decay function is used to interpolate log-F0 in unvoiced region. 80% of the science frames are removed from the training data to balance the training data and reduce computational cost. Both input and output features of training data are normalized to zero mean and unit variance.

For multi-style BLSTM based SPSS, the 6,000 utterances of declarative corpus are used as the training data to get the "average style voice", while 550 utterances in other two speaking styles corpus are used as the training data for style adaptation. In multilingual BLSTM, we use two BSTM layers with 512 units for each layer to capture the long time span contextual effect of the training data.

For multi-style BLSTM-F0 based SPSS, the difference between multi-style BLSTM is that such system models spectral and excitation parameters separately and only the excitation parameters is adapted.

We also substitute the BLSTM for FNN in every model that built on BLSTM and compare their performance in this paper. They can be called as mono-style FNN, multi-style FNN, multi-style FNN-F0 respectively. FNN is set with 4 hidden layers and 1024 nodes for each layer.

For testing, the outputs of all the systems are fed into a parameter generation module to generate smooth feature parameters with the dynamic constraints. Then formant sharping based on LSP frequencies is used to reduce the over-smoothing problem in modeling. Finally speech waveforms are synthesized by LPC synthesizer with generated speech parameters.

We use theano [32] as the implementation of our FNN and BLSTM training, and HTS [33] for the HMM-based speech synthesis.

4.2. Evaluation results and analysis

4.2.1. Network type and topology

The neural network types and the number of layers in style-specific layer affect the network topology and corresponding performance. Thus we would first evaluate different combinations of neural networks types and the style-specific layers' number for the proposed methods. In this section, both spectral and excitation parameters are adapted.

Table 2 shows the average objective test results of the proposed multi-style DNN for the generation of exclamatory speech in different network configurations. In Table 2, top1, 2 and all, means the top 1, 2 and all the layers are served as the style-specific layers in the neural network respectively, while mono denotes the mono-style DNN. From the objective results, multi-style BLSTM with Top1 layer is style-specific layer, and multi-style FNN with top2 layers are style-specific layers get the best performance in corresponding neural network type, and BLSTM with top1 layer achieves the best performance, among all models including HMM-CSMAPLR and mono-style DNN. This indicates the effectiveness of our prosed approach. Also we can see that, when all layers serve as the style-specific layers, all these two neural networks couldn't get the best performance. This can be explained by that the multi-style training corpus is not large enough to adjust all neural networks well. The average objective test results of the generation of interrogative speech are similar to the exclamatory one.

In the subjective test, we compare the multi-style BSLTM (with the optimal topology), mono-style BLSTM and HMM-

CSMAPLR by naturalness MOS. 11 listeners are invited to take part in the evaluation of synthetic speeches.

Figure 3 shows the naturalness MOS results for the generation of the exclamatory and interrogative speech. The proposed multi-style BLSTM behaves best in naturalness test on the generation of both exclamatory and interrogative speech, which indicates the multi-style BSLTM can benefit the synthetized quality with "average style voice" as it can build transformation for different speaking styles.

Table 2.*Objective Measures of multi-style DNN in different network configurations.*

Measures / *Model*	*LSD (dB)*	*V/U Err (%)*	*LogF0 RMSE (Hz)*
BLSTM(top1)	4.67	4.97	0.193
BLSTM(all)	4.90	5.14	0.201
FNN(top1)	5.29	5.42	0.212
FNN(top2)	5.01	5.18	0.206
FNN(all)	5.09	5.22	0.207
BLSTM(mono)	7.87	8.96	0.268
FNN(mono)	9.21	11.5	0.282
HMM-CSMAPLR	5.42	5.58	0.224

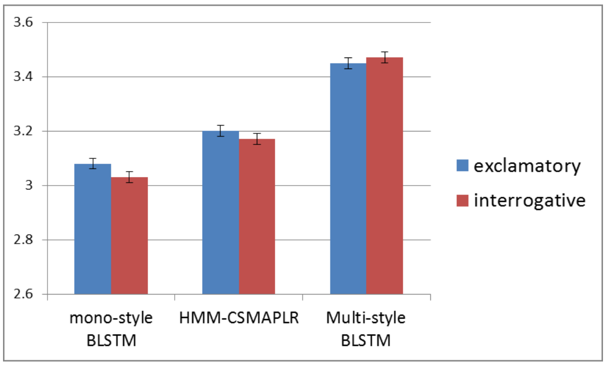

Figure 3: *Naturalness MOS results of Multi-style BLSTM.*

4.2.2. Adaptation parameters

We perform ABX preference test between the multi-style BLSTM-F0 (only excitation parameters are adapted) and multi-style BLSTM to validate whether it is more useful to only adapt the excitation parameters or not. The multi-style BLSTM-F0 has the same topology as multi-style BLSTM. During the test, the subjects are asked to indicate their preference for each test pair where the scale corresponds to prefer A, no preference and prefer B. The subjective ABX test includes 11 listeners, who compared 20 sentence pairs randomly chosen from test database. Test result, given in Figure 4 indicates that, when both the spectral and excitation parameters are adapted, the better results we can get.

4.2.3. Adaptation corpus size

For speech synthesis of low-resource speaking style, it would be valuable to find the least amount data for generating synthetic speech with satisfied speech quality. Based on the above experiments setup, we try to adjust the amount of training utterances of exclamatory corpus. We try different settings of the number of the training utterances, from 550, 300, 250, 200. The objective measures of different settings are shown in Table 3. As we can see, when the number of training utterances is about 250 utterances, these objective measurements are close to the performance of the HMM-CSMAPLR trained with 550 utterances. This indicates that we can add new emotional expressions and speaking styles to multi-BLSTM quickly by only recording a few utterances.

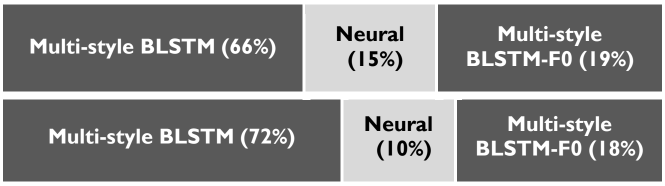

Figure 4: *Preference on exclamatory (above) and interrogative (below) multi-style BLSTM and multi-style BLSTM-F0.*

Table 3. *Objective Measures on different number of training utterances.*

Measures / *Utterances*	*LSD (dB)*	*V/U Err (%)*	*LogF0 RMSE (Hz)*
300	5.29	5.44	0.211
250	5.52	5.69	0.229
200	5.80	6.14	0.240

5. Conclusions

In this paper, we propose a multi-style BLSTM (also FNN) that shares hidden layers across different speaking styles. With this architecture, the shared layers of multi-style BLSTM can exploit the commonalities among different speaking styles so as to transfer learned knowledge to a new speaking styles. This is useful for the training of a text to speech synthesizer for low-resource or resource-limited speaking styles. We can add new emotional expressions and speaking styles to multi-BLSTM quickly by only recording a few utterances. The experiments with the generation of exclamatory and interrogative speech validate the effectiveness of the proposed methods. Both objective and subjective evaluations indicate that multi-style BLSTM can predict more accurate acoustic features than other models, such as mono-style BLSTM and HMM-CSMAPLR. We also validate the multi-style BLSTM (FNN) is more useful than multi-style BLSTM (FNN), and adaption on both spectral and excitation is prior than only excitation parameters is adapted. Our future research will use more speaking styles to evaluate the performance of multi-style BLSTM in a scale-up manner.

6. Acknowledgements

This work is supported by the National High-Tech Research and Development Program of China (863 Program) (No. 2015AA016305), the National Natural Science Foundation of China (NSFC) (No.61305003, No.61425017，No.61403386), the Strategic Priority Research Program of the CAS (Grant XDB02080006) and partly supported by the Major Program for the National Social Science Fund of China (13&ZD189).

7. References

[1] A. Black and N. Cambpbell, "Optimising selection of units from speech database for concatenative synthesis," in *Proc. EUROSPEECH'95*, Sep. 1995, pp. 581–584.

[2] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. ICASSP'96*, May 1996, pp. 373–376.

[3] R. Donovan and P. Woodland, "A hidden Markov-model-based trainable speech synthesizer," *Comput. Speech Lang.*, vol. 13, no. 3, pp. 223–241, 1999.

[4] S King, T Merritt, "Deep neural network-guided unit selection synthesis," in *Proc. ICASSP'16*, March 2016, pp. 333–336.

[5] A. Black, "Unit selection and emotional speech," in *Proc. Eurospeech'03,* Sep. 2003, pp. 1649–1652.

[6] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "HMM-based speech synthesis using dynamic features," (in Japanese) *IEICE Trans.*, vol. J79-D-II, no. 12, pp. 2184–2190, Dec. 1996.

[7] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," (in Japanese) *IEICE Trans.*, vol. J83-D-II, no. 11, pp. 2099–2107, Nov. 2000.

[8] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP'00*, Jun. 2000, pp. 1315–1318.

[9] J.Yamagishi, K. Onishi, T. Masuko, and T.Kobayashi, "Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis," IEICE Trans. Inf. Syst., vol. E88-D, no. 3, pp. 503–509, Mar. 2005.

[10] M. Tachibana, J. Yamagishi, T. Masuko, and T. Kobayashi, "A style adaptation technique for speech synthesis using HSMM and supra segmental features," *IEICE Trans. Inf. Syst.,* vol. E89-D, no. 3, pp. 1092–1099, Mar. 2006.

[11] M. Tachibana, J. Yamagishi, T. Masuko, and T. Kobayashi, "Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing," *IEICE Trans. Inf. Syst.*, vol. E88-D, no. 11, pp. 2484–2491, Nov. 2005.

[12] T. Nose, J. Yamagishi, and T. Kobayashi, "A style control technique for HMM-based expressive speech synthesis," *IEICE Trans. Inf. Syst.,* vol. E90-D, no. 9, pp. 1406–1413, Sep. 2007.

[13] Y. Fan, Y. Qian, F.K. Soong, and L. He, "Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis," in Proc. ICASSP, 2015, pp. 4475 – 4479.

[14] Heiga Zen and Hasim Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," *in Proc. ICASSP,* 2015, pp. 4470 – 4474.

[15] Keiichi Tokuda and Heiga Zen, "Directly modeling speech waveforms by neural networks for statistical parametric speech synthesis," *in Proc. ICASSP, 2015,* pp. 4215 – 4219.

[16] Zhizheng Wu, Cassia Valentini-Botinhao, Oliver Watts, and Simon King, "Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis," *in Proc. ICASSP,* 2015, pp. 4460 – 4464.

[17] Y. Fan, Y. Qian, F.K. Soong, and L. He, "Speaker and language factorization in DNN-based TTS synthesis," *in Proc. ICASSP'16,* March 2016, pp. 1005–1008.

[18] Q. Yu, P. Liu and L. Cai, "Learning cross-lingual information with multilingual BLTSM for speech synthesis of low-resource language," *in Proc. ICASSP'16*, March 2016, pp. 1233–1236.

[19] JE Cahn - Master's thesis, Massachusetts Institute of Technology, 1989.

[20] J Tao, Y Kang, A Li, "Prosody Conversion from Neural Speech to Emotional Speech", *IEEE Transactions on Audio, speech and language Processing,* vol.14, Issue 4, pp.1145-1154, 2006.

[21] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Voice characteristic conversion for HMM-based speech synthesis system," *in Proc. ICASSP'97,* Apr. 1997, pp. 1611–1614.

[22] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR," *in Proc. ICASSP'01*, May 2001, pp. 805–808.

[23] J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda, and T.Kobayashi, "A training method of average voice model for HMM-based speech synthesis," *IEICE Trans. Fundamentals,* vol. E86-A, no. 8, pp. 1956–1963, Aug. 2003.

[24] J.Yamagishi and T.Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 2, pp. 533–543, Feb. 2007.

[25] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Speaker adaptation for HMM-based speech synthesis system using MLLR," in *Proc. 3rd ESCA/COCOSDA Workshop Speech Synth.,* Nov. 1998, pp. 273–276.

[26] C. Leggetter and P.Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput. Speech Lang.,* vol. 9, no. 2, pp. 171–185, 1995.

[27] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Speaker adaptation of pitch and spectrum for HMM-based speech synthesis," (in Japanese) *IEICE Trans.,* vol. J85-D-II, no. 4, pp. 545–553, Apr. 2002.

[28] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A hidden semi-Markov model-based speech synthesis system," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 5, pp. 825–834, May 2007.

[29] S. Fang, Z. Wen and J. Tao, "Speech synthesis of questions based on adaptive training," in Proc. NCMMSC, 2015, pp. 092–095, Oct. 2015.

[30] J. Yamagishi, T. Kobayashi, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR Adaptation algorithm," *IEICE Trans. Inf. & Syst*., pp. 66–84, Jan 2009.

[31] Heiga Zen, Andrew Senior, and Mike Schuster, "Statistical parametric speech synthesis using deep neural networks," in Proc. ICASSP'13, 2013.

[32] "Theano: A Python framework for fast computation of mathematical expressions". [OL] [2016-05-09] *http://deeplearning.net/software/theano/*

[33] K. Tokuda, H. Zen, and A.W. Black, "An HMM-based speech synthesis system applied to English," in P*roc. IEEE Speech Synthesis Workshop*, 2002, pp. 227 – 230.