EMPHATIC SPEECH GENERATION WITH CONDITIONED INPUT LAYER AND BIDIRECTIONAL LSTMS FOR EXPRESSIVE SPEECH SYNTHESIS

Runnan Li[1,2], Zhiyong Wu[1,2,3], Yuchen Huang[1,2], Jia Jia[2,], Helen Meng[1,3], Lianhong Cai[2]*
[1] MJRC, Graduate School at Shenzhen, Tsinghua University, China
[2] Dept. of Computer Science and Technology, Tsinghua University, China
[3] Dept. of Systems Engineering & Engineering Management, CUHK, Hong Kong SAR, China
{lirn15, huang-yc15}@mails.tsinghua.edu.cn, zywu@sz.tsinghua.edu.cn,
hmmeng@se.cuhk.edu.hk, {hanwentao,clh-dcs}@tsinghua.edu.cn

ABSTRACT

By highlighting the focus of an utterance to draw attention, emphasis in speech interaction plays an important role for speaker intention expressing and understanding. Therefore, emphatic speech synthesis draws increasing interest in the text-to-speech (TTS) area. For emphatic speech synthesis, three problems still exist: 1) sparseness of emphatic speech data; 2) flexibility of trained model; 3) modelling shortage for secondary emphasis. Recently, recurrent neural networks (RNNs) and their bidirectional long short term memory (BLSTM) variants based statistical parametric speech synthesis (SPSS) systems have shown their adaptability and controllability in acoustic modelling thus can solve aforementioned problems. In this paper, we propose a novel conditional input layer for conventional BLSTM-RNN based approach combining using emphasis-specific vectors and linguistic features as input to produce emphatic speech trajectories. Experimental results from objective and subjective evaluations demonstrate the proposed approach can produce emphatic speech trajectories with high quality and naturalness only requiring an additional small-scale emphatic speech corpus.

Index Terms— Emphatic speech generation, conditioned input layer, bidirectional long short term memory, acoustic model, duration model

1. INTRODUCTION

In human-computer speech interaction scenarios, expressive speech generation is required to properly convey the message [1]. State-of-art text-to-speech (TTS) synthesis systems can generate speech trajectories with high quality and naturalness, but being still weak in producing expressive speech [2][3]. Emphasis, as an important form of expressiveness, can highlight the speech focus to enhance the expression of speaker intension [4], is thus attracting increasing interest.

To synthesize emphasis, techniques have been developed for the two mainstream speech synthesis techniques, unit selection speech synthesis and statistical parametric speech synthesis (SPSS). For unit selection speech synthesis system, [5] proposed a rule-based emphasis generation approach and [6] proposed the use of Gaussian mixture model (GMM) and decision tree (DT) to guide unit selection for emphatic speech generation. Hidden Markov models (HMMs) based SPSS models are also widely researched to produce emphatic speech. This is commonly achieved using specified decision tree based parameter tying approaches to generate low-level speech waveforms from high-level symbol sequences via intermediate acoustic feature sequences [7][8][9][10]. While the DTs are trained based on training data and hardly considered about the influence from emphasis to neighboring words [11], these techniques thus suffer from three problems: 1) sparseness of emphatic speech data; 2) flexibility of trained model; 3) modelling shortage for secondary-emphasis.

Inspired by the intrinsically hierarchical process of human speech production and successful application in automatic speech recognition (ASR), deep learning neural networks based speech synthesis techniques have become increasing popular recently [12][13][14][15]. Comparing to conventional approaches, these techniques use a deep model architecture with multiple hidden layers to provide: high-level abstract and discriminative feature learning; shared parameters to learn implicit dependency and avoid data partitioning; and long range temporal context modelling. These inherently strong abilities provide neural network based SPSS a flexibility in changing speaker characteristics, speaking styles and expressions [12]. But even this flexibility has been successfully used in the area of speaker adaptation [16][17], only a few studies have addressed the question of whether the neural networks based speech synthesis can offer the flexibility in expressive speech synthesis, e.g. the emphatic speech generation.

In this work, we proposed the use of a novel conditioned input layer (CIL) with bidirectional long short term memory recurrent neural network (BLSTM-RNN) [18][19] to generate emphatic speech. In particular, we exploit the ability of CIL to generate weighted input features from linguistic features conditioning on emphasis-specific vector at input layer, and then exploit the inherently temporal

* Corresponding author.

modelling ability of BLSTM-RNN to model long term context information as well as the emphasis-related position information at middle level. By performing at different levels, these techniques can be effectively combined. Advantages are taken from the proposed architecture: 1) by sharing learning the representation of phones using both neutral and emphatic speech, the model can relieve the influence sparseness of emphatic data; 2) the model gains higher flexibility in generating emphatic speech using emphasis-specific vector as additional input; 3) the influence from emphasis to neighboring words can be implicitly modelled by BLSTM for its strong inherent ability in capturing long span temporal dependencies. Experimental results from objective and subjective evaluation demonstrate the proposed approach can produce emphatic speech trajectories with high quality and naturalness.

The rest of the paper is organized as follows. Section 2 presents the conditioned input layer. Section 3 presents the proposed emphatic speech system using BLSTM-RNN with conditioned input layer. Section 4 presents objective and subjective experimental results. Conclusions are drawn and future work discussed in section 5.

2. CONDITIONED INPUT LAYER

For emphatic speech synthesis, the embedded features provided by input layer to hidden layers can significantly influence the performance of speech trajectories generation. In this work, we proposed the use of a novel input layer, conditioned input layer (CIL), to generate input features $\hat{\boldsymbol{x}}$ conditioning on an additional information $\boldsymbol{y}$. $\boldsymbol{y}$ could be any kind of auxiliary information, e.g. other correlated data or class labels like emphasis.

As shown in Fig.1, the CIL contains a fully-connected linear dense layer to scale the input $\boldsymbol{x}$ and an embedding layer to augment the binary one hot input $\boldsymbol{y}$ to the same size of $\boldsymbol{x}$. The weighted representation $\hat{\boldsymbol{x}}$ is then calculated by adding scaled $\boldsymbol{x}$ and embedded $\boldsymbol{y}$. The process is described as follow:

$$\boldsymbol{h_x} = \sigma_x(\boldsymbol{W_x x} + \boldsymbol{b_x}) \quad (1)$$

$$\boldsymbol{h_y} = Embedding(\boldsymbol{y}) \quad (2)$$

$$\hat{\boldsymbol{x}} = \boldsymbol{h_x} + \boldsymbol{h_y} \quad (3)$$

where $\sigma_x(\cdot)$, $\boldsymbol{W_x}$, $\boldsymbol{b_x}$ are the linear activation function, weights matrix and bias vector connecting to input $\boldsymbol{x}$ and warped hidden output $\boldsymbol{h_x}$. $\boldsymbol{h_y}$ is the embedded hidden output of $\boldsymbol{y}$. This process is expected to increase the inter-class distance while maintaining the intra-class characteristics of feature distributions to provide class-specific representation.

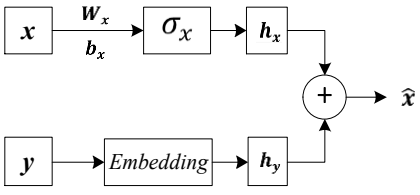

Fig.1. The structure of conditioned input layer.

3. EMPHATIC SPEECH GENERATION WITH CONDITIONED INPUT LAYER BLSTM-RNN

The overall architecture of the proposed emphatic speech synthesis system is illustrated in Fig.2. As the emphasis perception is affected by phone duration and corresponding acoustic parameters [2], two models are implemented respectively in the proposed emphasis synthesis system: 1) duration prediction model; 2) acoustic parameters prediction model. These two model both employ the proposed CIL as input layer and BLSTM-RNN as hidden layers.

By using CIL as input layer, the input features $\widehat{\boldsymbol{X}}$ for duration model and acoustic model are generated by scaling linguistic features $\boldsymbol{X}$ and embedded emphasis-specific vector (ESV). The emphatic input features $\widehat{\boldsymbol{X_E}}$ is thus different from the neutral input $\widehat{\boldsymbol{X_N}}$ while maintaining intra-class features characteristics and distributions. This can help the model to learn the specific representations of emphatic phonemes as well as the difference between their neutral representations. With this specialty, the model can learn the sharing representation of phone using both neutral and emphatic speech, and output specific representation conditioning on emphasis-specific label in prediction. This can help the model to generate emphatic phones even their contexts are out of training set.

By using BLSTM-RNN, the model inherits the stronger ability in capturing long range temporal context dependencies as well as the influence from emphasized words to their neighboring. This can help the model to learn the specific changes of representations in secondary-emphases across temporal axis to produce higher naturalness emphatic speech.

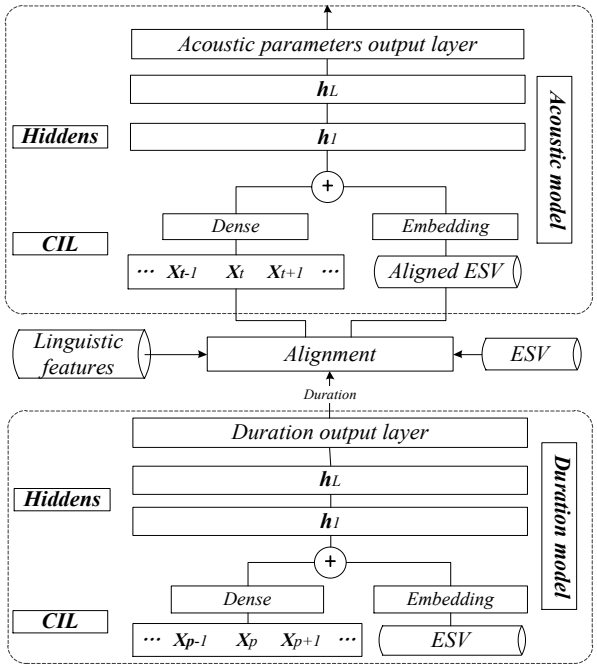

Fig.2. Schematic of proposed system, $\{\boldsymbol{h_1}, \ldots, \boldsymbol{h_L}\}$ presents BLSTM-RNN hidden layers, $\mathbf{X_p}$ and $\mathbf{X_t}$ are the phone-level and frame-aligned linguistic inputs.

Table 1: Results of objective evaluations. MCD and RMSE are Mel-Cepstral Distortion and Root Mean Squared Error calculated between predicted parameters and their reference. V/UV error means frame-level voiced/unvoiced swiping error.

Systems	*Prediction Performance*						*Secondary-emphasis Prediction Performance*					
	DUR RMSE (ms)	**MCEP MCD (dB)**	**ENG MCD (dB)**	**BAP MCD (dB)**	**F0 RMSE (Hz)**	**V/UV Error (%)**	**DUR RMSE (ms)**	**MCEP MCD (dB)**	**ENG MCD (dB)**	**BAP MCD (dB)**	**F0 RMSE (Hz)**	**V/UV Error (%)**
DNN	9.034	5.871	3.659	4.872	7.829	4.232	14.61	5.559	4.004	4.543	6.456	3.403
CDNN	8.704	5.847	3.57	4.858	7.087	4.291	13.94	5.557	3.916	4.555	6.097	3.295
BLSTM	7.801	5.106	3.091	4.653	6.244	**3.294**	12.67	5.099	3.795	4.448	5.681	2.692
CBLSTM	**7.419**	**4.981**	**2.922**	**4.593**	**6.074**	3.369	**12.32**	**4.752**	**2.921**	**4.401**	**4.565**	**2.551**

3. EXPERIMENTS

In this section, we process objective and subjective evaluations to assess the proposed approach.

3.1. Experimental Setup

Corpus. A modified TH-Coss speech corpus [21] from a native Mandarin female speaker is employed to train the models. This corpus contains 2,500 neutral utterances (around 131 mins) and 500 emphatic utterances (around 39 mins), both are phonetically and prosodically rich. The corpora is split with 8:1:1 as training set, validation set and testing set.

Features. The input linguistic features vector is of 1564 dimensions including tri-phone, tone, positional information, word and phrase related information and emphasis-specific label. Phone-level duration is firstly acquired using force-alignment with a HMM model from HTS toolkit and then manually corrected. Three numerical temporal features were appended: 1) the syllable position in sentence; 2) the frame position in syllable; 3) the frame position in sentence. Numerical features are normalized to the range of (0, 1].

For duration modelling, target vector contains one dimensional numerical duration vector, normalized to zero mean and unit variance. For acoustic parameters modelling, target acoustic features vector is of 58 dimensions including 39 dimensional Mel-cepstral coefficients (MCEPs) and 1 dimensional energy representing the spectral envelope, 15 dimensional band aperiodicity (BAPs), 1 dimensional logarithmic fundamental frequency (LF0) and 2 dimensional voiced/unvoiced binary value (V/UV), and then normalized to zero mean and unit variance. These features are extracted using STRAIGHT and SPTK with 25ms window size and 5ms shift.

Comparison approaches. To assess the proposed approach, four different systems are implemented, of each contains one duration model and one acoustic model: 1) deep neural network (DNN) based emphatic speech synthesis system; 2) deep neural network with conditioned input layer (CDNN) based emphatic speech synthesis system; 3) BLSTM-RNN based emphatic speech synthesis system; 4) proposed conditioned input layer BLSTM-RNN (CBLSTM-RNN) based emphatic speech synthesis system. In particular, the output of two DNN based systems comprised acoustic features comprised 39-D MCEPs, 15-D BAPs, 1-D LF0, their corresponding delta and delta-delta features, and 2-D V/UV. For system (1) and system (3), the emphasis-specific vector is directly concatenated with the linguistic features. Each DNN based model contains 4 hidden layers, 1024 nodes per layer. Each BLSTM based model contains 4 hidden layers, 512 nodes per layer (256 forward nodes and 256 backward nodes).

Implementation and training. The proposed approach and its comparisons are implemented using Keras [22] deep learning framework with Theano [23] as backend. Mini-batch-based Adam [24] algorithm is employed as the optimizer and backpropagation through time (BPTT) [25] is employed to train these models by unfolding RNNs into standard feed-forward networks through time steps.

3.2. Objective Evaluation

In objective evaluation, we assess the performance of proposed approach and the comparisons by computing the distortion between the predicted parameters and their targets. Although the objective results might be different with the perceived speech quality and naturalness, they are necessary for optimizing the models, e.g. determining training epoch times and needed data scale. In this evaluation, 50 emphatic utterances from test set are generated from each system and evaluated. Specially, duration extracted from natural speech is used directly in acoustic parameters prediction to better evaluate the performance of acoustic model.

3.2.1. Prediction Performance

The performance comparison result is illustrated in Table 1. By employing conditional emphasis-specific vector as additional input, CDNN slightly outperforms DNN baseline. BLSTM based system gains 13%, 13%, 15%, 4.5%, 19% and 22% relative improvement on duration, MCEP, energy, BAP, F0, V/UV prediction respectively comparing to the DNN baseline. CBLSTM outperforms all the comparisons, gaining 4.9%, 2.5%, 5.5%, 4.5%, 2.8% relative improvement on duration, MCEP, energy, BAP, F0 prediction respectively comparing to the BLSTM based approach.

3.2.2. Secondary-emphasis Prediction Performance

To figure out the prediction performance for secondary-emphasis, we did further analysis on secondary-emphasis

only. As illustrated in Table 1, for secondary-emphases, the overall performance of CDNN is on par with DNN baseline. However, while modifying the hidden layers with BLSTM, the CBLSTM based approach gains 11.7%, 14.5%, 25.4%, 21.3%, 22.6% and relative improvement on duration, MCEP, energy, BAP, F0, V/UV prediction respectively comparing to the CDNN based approach.

3.2.3. Data Scale Comparison

To figure out the influence of emphatic speech data scale, we also assess the proposed approach with different training emphatic speech data sizes. As shown in Fig.3, the system has achieved a better performance using bigger scale of emphatic speech corpus. However, the improvement is leveling out at 200-utterance-scale.

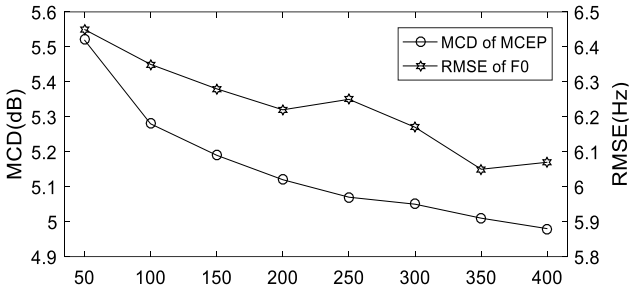

Fig.3. The prediction performance using different sizes of emphatic speech data.

3.3. Subjective Evaluation

We conducted two listening experiments to evaluate the perceived performance of proposed approach: 1) a perceptive experiment to evaluate the emphatic speech generation accuracy; 2) a 5-point perceptive experiment to evaluate the naturalness and quality of generated speech. 10 Mandarin native speakers with no reported listening difficulties are invited to assess 15 emphatic utterances generated by each aforementioned systems. Testing samples are available at: http://mjrc.sz.tsinghua.edu.cn/demo/ess/icassp2018/.

3.3.1. Emphasis Perception Evaluation

This experiment is designed to evaluate the perceptive accuracy of synthetic emphatic speech and its corresponding reference. The perceived emphasis in synthetic utterances are labelled by human labelers, and then compared to the used emphasis label in speech generation. As illustrated in Table 2, CBLSTM has been assessed with best perception accuracy.

Table 2: Emphasis perception evaluation using Precision, Recall and F1-measure [26].

Systems	Recall	Precision	F1-measure
DNN	0.47	0.85	0.605
CDNN	0.52	0.82	0.646
BLSTM	0.74	0.98	0.843
CBLSTM	0.84	0.95	0.898

3.3.2. Naturalness and Quality Evaluation

This experiment is designed to evaluate the speech generation performance in naturalness and quality using mean opinion score (MOS). The 5-point scale is designed as: 5 = Excellent (same as natural speech), 4 = Good (close to natural speech), 3 = Fair (synthetic speech but with clear presentation), 2 = Poor (synthetic speech with poor presentation), 1 = Bad (mechanical speech and hard to understand).

The result is shown in Fig.4, all the DBLSTM based systems outperform the baseline DNN based system. Compared with conventional BLSTM based system, the CBLSTM based system gains relative improvements for naturalness and quality at 4.1% and 3.4% respectively.

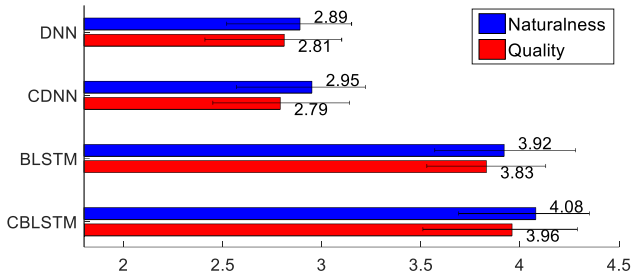

Fig.4. Results of MOS test with 95% confidence intervals for speech naturalness and quality.

4. CONCLUSIONS

In this paper, we proposed the use of conditioned input layer and BLSTM-RNN in emphatic speech generation. By using conditioned input layer, the model can inherit the stronger emphatic speech generation performance facilitated by modeling the specific representation of emphatic words as well as the difference between their corresponding neutral representations. By using BLSTM-RNN, the model can learn the context information as well as the influence from emphasized words to neighboring words in preceding and succeeding directions. In objective evaluation, the proposed CBLSTM based approach outperforms other comparisons in duration and acoustic parameters prediction. Specially, the proposed approach achieves significant improvement in secondary emphasis parameters prediction. The subjective experimental results also confirmed the effectiveness and efficiency of our proposed approach to generate emphatic speech with high quality and naturalness.

The work indicates the flexibility of neutral network based SPSS in expressive speech synthesis, in the future, we will investigate to enhance the synthesis model with other forms of expression, e.g. emotions.

Acknowledgement. This work is supported by National Natural Science Foundation of China (NSFC) (61433018, 61375027), NSFC-RGC (Research Grant Council of Hong Kong) joint fund (61531166002, N_CUHK404/15), National Social Science Foundation of China (13&ZD189) and the Science and Technology Plan of Beijing Municipality under Grant No. Z161100000216147. We would also like to thank Sogou Inc. and Tiangong Institute for Intelligent Computing, Tsinghua University for their support.

5. REFERENCES

[1] F. Meng, Z. Wu, J. Jia, H. Meng, L. Cai, “Synthesizing English emphatic speech for multimodal corrective feedback in computer-aided pronunciation training.” Multimedia tools and applications, vol. 73(1), pp. 463-489, 2014.

[2] Z. Wu, H. Meng, H. Yang, L. Cai. “Modeling the expressivity of input text semantics for Chinese text-to-speech synthesis in a spoken dialog system”. IEEE Transactions on Audio, Speech, and Language Processing, vol. 17(8), pp. 1567-1576, 2009.

[3] K. Yu, F. Mairesse, and S. Young. "Word-level emphasis modelling in HMM-based speech synthesis." Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on. IEEE, 2010.

[4] Y. Ning, J. Jia, Z. Wu, R. Li, Y. An, Y. Wang and H. Meng, “Multi-task Deep Learning for User Intention Understanding in Speech Interaction Systems”, [in] Proc. AAAI, 2017.

[5] Li, A. J. "Duration charateristics of stress and its synthesis rules on standard Chinese." Report of phonetic research, 1994.

[6] W. Zhu. "A Chinese speech synthesis system with capability of accent realizing." Journal of Chinese Information Processing, 21.3, pp. 122-128, 2007.

[7] Y. Maeno, T. Nose, T. Kobayashi, Y. Ijima, H. Nakajima, H. Mizuno, O. Yoshioka, "HMM-based emphatic speech synthesis using unsupervised context labeling." [in] Proc. Interspeech, 2011.

[8] K. Morizane, K. Nakamura, T. Toda, H. Saruwatari, K. Shikano, "Emphasized speech synthesis based on hidden Markov models." Speech Database and Assessments, 2009 Oriental COCOSDA International Conference on. IEEE, 2009.

[9] K. Tokuda , H. Zen, J. Yamagishi, T. Masuko, S. Sako, A. Black, T. Nose The HMM-based speech synthesis system (HTS) version 2.1, http://hts.sp.nitech.ac.jp/

[10] Z. Wu, Y. Ning, X. Zang, J. Jia, F. Meng, H. Meng, L. Cai, "Generating emphatic speech with hidden Markov model for expressive speech synthesis." Multimedia Tools and Applications, vol.74.22, pp. 9909-9925, 2015.

[11] Y. Xu, C. Xu, “Phonetic realization of focus in English declarative intonation”. J Phon 33:159–197.

[12] H. Zen, K. Tokuda, and A. Black, “Statistical parametric speech synthesis,” Speech Commn., vol. 51, no. 11, pp. 1039-1064, 2009.

[13] S. Kang, X. Qian, and H. Meng, “Multi-distribution deep belief network for speech synthesis”, [in] Proc. ICASSP, pp. 7962-7966, 2013.

[14] H. Zen, A. Senior and M. Schuster, “Statistical Parametric Speech Synthesis Using Deep Neural Networks”, [in] Proc. ICASSP, pp. 8012-8016, 2013.

[15] Y. Qian, Y.-C. Fan, W.-P. Hu and F. K. Soong, “On the training aspects of deep neural network (DNN) for parametric TTS synthesis”, [in] Proc. ICASSP, pp. 3829-3833, 2014.

[16] B. Potard, P. Motlicek, and D. Imseng, “Preliminary work on speaker adaptation for dnn-based speech synthesis,” Idiap, Tech. Rep., 2015.

[17] Z. Wu, P. Swietojanski, C. Veaux, S. Renals, S. King. “A study of speaker adaptation for DNN-based speech synthesis”. [in] Proc. Interspeech, pp. 879-883, 2015.

[18] S. Hochreiter, and J. Schmidhuber, “Long short-term memory,” Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[19] M. Schuster, and K. K. Paliwal, “Bidirectional recurrent neural networks,” IEEE Transactions on Signal Processing, vol. 45, pp. 2673–2681, Nov 1997.

[20] AM. Rush, C. Sumit, and W. Jason, "A neural attention model for abstractive sentence summarization." arXiv, preprint arXiv:1509.00685, 2015.

[21] L. Cai, D. Cui, and R. Cai. "TH-CoSS, a mandarin speech corpus for TTS." Journal of Chinese Information Processing, vol. 21, no. 2, pp. 94-99, 2007.

[22] F. Chollet, Keras [OL]. [2017-10-11]. GitHub repository. https://github.com/fchollet/keras.

[23] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio, “Theano: A CPU and GPU math compiler in Python,” [in] Proc. SciPy, pp. 1-7, 2010.

[24] D. P. Kingma, and J. L. Ba, “Adam: A method for stochastic optimization,” [in] Proc. ICLR, 2015.

[25] P. J. Werbos, “Backpropagation through time: what it does and how to do it,” [in] Proc. IEEE, vol. 78, no. 10, pp. 1550–1560, 1990.

[26] D. Powers, “Evaluation: from precision, recall and Fmeasure to ROC, informedness, markedness and correlation”. Journal of Machine Learning Technologies Vol. 2(1), pp. 37-63, 2011.