MULTI-TASK LEARNING OF STRUCTURED OUTPUT LAYER BIDIRECTIONAL LSTMS FOR SPEECH SYNTHESIS

Runnan Li[1], Zhiyong Wu[1,2], Xunying Liu[2], Helen Meng[1,2], Lianhong Cai[1]

[1]Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, Graduate School at Shenzhen, Tsinghua University

[2]Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong

lirn15@mails.tsinghua.edu.cn, {zywu, xyliu, hmmeng}@se.cuhk.edu.hk, clh-dcs@tsinghua.edu.cn

ABSTRACT

Recurrent neural networks (RNNs) and their bidirectional long short term memory (BLSTM) variants are powerful sequence modelling approaches. Their inherently strong ability in capturing long range temporal dependencies allow BLSTM-RNN speech synthesis systems to produce higher quality and smoother speech trajectories than conventional deep neural networks (DNNs). In this paper, we improve the conventional BLSTM-RNN based approach by introducing a multi-task learned structured output layer where spectral parameter targets are conditioned upon pitch parameters prediction. Both objective and subjective experimental results demonstrated the effectiveness of the proposed technique.

Index Terms— text-to-speech, acoustic model, multi-task learning, structured output layer, deep bidirectional long short-term memory

1. INTRODUCTION

A central task of statistical parametric text-to-speech (TTS) synthesis systems is to learn the complex non-linear mapping from abstract linguistic features to their acoustic representation [1][2]. Statistical parametric TTS models based on hidden Markov models (HMMs) [3][4] produce low-level speech waveforms from high-level symbol sequences via intermediate acoustic feature sequences. This is commonly achieved using decision tree based parameter tying approaches that can lead to data partitioning and poor generalization.

Inspired by the intrinsically hierarchical process of human speech production and by the successful application of deep neural networks (DNNs) to automatic speech recognition (ASR) systems [5], deep learning based speech synthesis techniques have become increasingly popular in recent years. These techniques use a deep model architecture with multiple hidden layers to provide: high-level abstract and discriminative feature learning; shared parameters to avoid data partitioning and improve generalization; and long range temporal context modelling. Earlier works along this line include deep belief networks (DBNs) [6][7], and deep neural networks (DNNs) [8][9][10]. In order to learn longer span temporal information, recurrent neural networks (RNNs) and their bidirectional long short term memory (BLSTM) variants [11][12][13] in particular, have also been proposed in previous research [14][15]. Their inherently strong ability in capturing long range temporal dependencies allow BLSTM-RNN speech synthesis systems to produce higher quality and smoother speech trajectories than conventional deep neural networks.

In conventional speech synthesis systems based on BLSTM-RNNs, or DNNs in general, the output layer is commonly constructed to produce acoustic features that concatenate spectral and pitch contour parameters. Two issues arise when using this form of output layer architecture. First, it is difficulty to model the dependency of spectral features on pitch contour parameters, for example, voicing decision. Second, due to the difference in dimensionality between spectral and pitch features, a larger part of network connections is used to model spectral features prediction. This can bias the gradient statistics accumulated at intermediate hidden layers to those associated with spectral features generation, while those obtained from the pitch parameter prediction unduly suppressed. It is therefore preferable to introduce additional controllability over the weighting assigned to the error costs incurred in spectral features generation and those in pitch parameter prediction during model training.

In order to address these issues, this paper proposes the use of a structured output layer (SOL) [16] for conventional BLSTM-RNNs where the spectral features outputs are set to be dependent on the prediction of pitch contour parameters. In order to further appropriately balance the error cost functions associated with spectral feature and pitch parameter targets, the proposed structured output layer BLSTM-RNN models are trained using a multi-task learning [17][18] approach. Both objective and subjective experimental results suggest the proposed technique improved the quality and naturalness of synthetic speech over the baseline BLSTM-RNN synthesis system.

The rest of the paper is organized as follows. Section 2 proposes modified BLSTM-RNN model architecture with a structured output layer. Section 3 presents the multi-task learning based training for the proposed structured output layer BLSTM-RNNs. Objective and subjective experimental results are presented in section 4. The conclusions are drawn and future work discussed in section 5.

2. STRUCTURED OUTPUT LAYER BLSTM-RNNS

Human speech is produced by the cooperation of vocal folds and articulators. The vibrating vocal folds generate the laryngeal sound via periodically regulating the airflow from the lungs and then the articulators form a filter for the laryngeal sound to generate human sound [19]. In statistical parametric speech systems, pitch parameters are used to represent the state of the vocal folds while the spectral parameters are those associated with the articulators. These two types acoustic features are highly related. Their correlation has been utilized for various purposes in previous research [20][21].

As discussed in Section 1, in conventional BLSTM-RNNs based speech synthesis systems, the output layer is normally constructed to produce acoustic features that concatenate spectral and pitch parameters. In order to model the dependency of spectral features on pitch contour parameters, such as probability of voicing, and appropriately adjust the balance between the error costs incurred in acoustic features generation and pitch parameter prediction during model training, a modified DBLSTM-RNN model architecture with a structured output layer (SOL) is proposed in this paper. This is shown in Figure 1.

Instead of directly predicting concatenated acoustic feature outputs, the network output layer is modified to perform two separate prediction tasks for spectral features and pitch parameters respectively. The main task of spectral feature prediction is further conditioned upon the auxiliary task of pitch parameter generation. This is realized by feeding the pitch parameter generation task's hidden layer output h_p through an activation function $(\cdot)$, such as Softmax or ReLU, to model the correlation between the two tasks before being augmented to the hidden layer output h_s while the weight matrix C used to connect the two tasks is applied.

In the conventional multi-task formulation where no between task dependency is modelled, the two tasks share the same BLSTM-RNN hidden layers $\{h_1, \ldots, h_L\}$, and the prediction of spectral and pitch parameters are computed as the following,

$$h_s = (W_{h_L s} h_L + b_s) \quad (1)$$

$$O_s = \sigma_s(h_s) \quad (2)$$

$$h_p = (W_{h_L p} h_L + b_p) \quad (3)$$

$$O_p = \sigma_p(h_p) \quad (4)$$

where O_s and O_p are predicted spectral and pitch parameter outputs respectively. $\{W_{h_L s},\ b_s\}$ and $\{W_{h_L p},\ b_p\}$ are the weight matrices and bias vectors connecting the shared BLSTM-RNN hidden layer h_L with the outputs associated with the two tasks. $\sigma_s(\cdot)$ and $\sigma_p(\cdot)$ are the linear output activation functions employed to produce the final predicted spectral feature and pitch parameter outputs.

In contrast, the proposed SOL based approach shown in Figure 1 introduces additional dependency of the primary spectrum prediction task on the auxiliary pitch parameter prediction task. The main spectral feature outputs are thus modified as,

$$h_{sp} = (W_{h_L s} h_L + \psi(h_p)\mathrm{C} + b_s) \quad (5)$$

$$O_s = \sigma_s(h_{sp}) \quad (6)$$

Precursors of the same SOL structure have been previously studied for acoustic modelling in speech recognition systems [16] and recurrent neural network language modelling for predicting morphologically decomposed stem and suffix features [22].

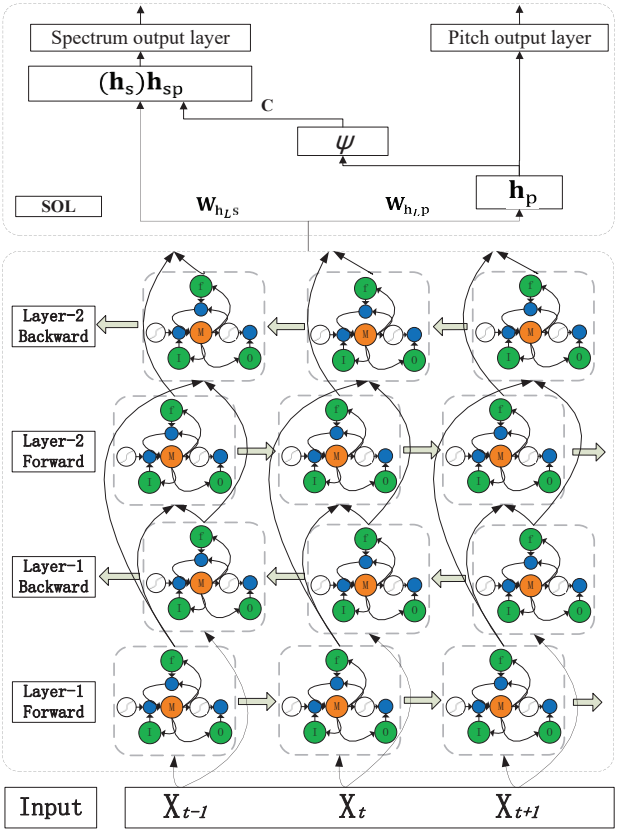

Figure 1: Overall structure of DBLSTM-RNN with SOL. In this network, the tasks share the same BLSTM hidden representations and the spectrum prediction can be benefit by using the hidden layer of pitch prediction. X are the frame-aligned linguistic inputs.

3. MULTI-TASK LEARNING OF SOL BLSTM-RNNS

In common with the conventional multi-task learning framework, structured output layer BSTLM-RNNs can be trained by minimizing a global cost function expressed as a weighted sum of the two task specific separate error costs. This is given by

$$F_g = \alpha F_s + (1 - \alpha) F_p \quad (7)$$

where F_s and F_p are the costs generated by the main task (spectral features) and the auxiliary task (pitch parameters) computed as mean squared errors (MSE),

$$F_s = \frac{1}{NT} \sum_1^N \sum_1^T (O_s - s)^2 \quad (8)$$

$$F_p = \frac{1}{NT}\sum_1^N \sum_1^T (O_p - p)^2 \quad (9)$$

where T is the total length of the input linguistic sequence, N is the mini-batch size, and α is a tunable weighting parameter adjusting the contribution from the main and auxiliary tasks.

Hence, the global error cost in (7) can be re-expressed as

$$F_g = \frac{1}{NT}\sum_1^N \sum_1^T [\alpha(O_s - s)^2 + (1-\alpha)(O_p - p)^2] \quad (10)$$

The gradients used to update all the parameters in the SOL BLSTM-RNN network are then computed as the weighted average gradient statistics computed over both tasks:

$$\frac{\partial F}{\partial \theta^k} = \frac{1}{NT}\sum_1^N \sum_1^T [\alpha \frac{\partial}{\partial \theta_s^k}(O_s - s)^2 + (1-\alpha)\frac{\partial}{\partial \theta_p^k}(O_p - p)^2] \quad (11)$$

where θ presents all the parameters of the model including those in in the hidden layers.

The proposed SOL BLSTM-RNN model inherits the stronger generalization performance and robustness of conventional multi-task learning [16][17][18] facilitated by shared hidden layers and joint training over multiple tasks. The use of a structured output layer further allows both the regularization properties of the comparatively simpler auxiliary task of predicting pitch contour variation and its direct effect on the primary spectral feature generation task to be fully exploited.

4. EXPERIMENTS

4.1 Experimental setup

The TH-Coss speech corpus [23] containing 5429 phonetically and prosodically rich utterances from a native Mandarin female speaker is used as speech dataset in experiments: 5000 utterances (around 8.5 hours) as training set, 200 utterances as validation set and the rest 229 utterances are reserved as test set. Speech signals are sampled at 16K Hz. Statistical parameters including 40 dimensional Mel-frequency cepstral coefficients (MFCCs), 25 dimensional band aperiodicity (BAPs), logarithmic fundamental frequency (log F0) and Voiced/Unvoiced flag (V/UV) are extracted with STRAIGHT [24].

The input linguistic features vector is of 329 dimensions including tri-syllable, syllable tone, positional information, word and phrase related information and so on, where 291 are binary features for categorical linguistic contexts and the rest are numerical features. Specially, three numerical position features were appended: the syllable position within the sentence, the frame position within syllable and frame position within the sentence. The input numerical features are normalized to the range of (0, 1] and the frame level forced alignment upon the training data is processed with a HMM system implemented by HTS toolkit [3]. The target acoustic features are normalized to zero mean and unit variance before training. Four different models have been implemented for comparison:

- **DNN**: baseline DNN-based approach containing 4 hidden layers with 1024 nodes each.
- **DBLSTM**: conventional DBLSTM-based approach containing 2 BLSTM hidden layers with 512 nodes per layer (256 forward nodes and 256 backward nodes, same setting for the following BLSTM derived approaches) and conventional output layer.
- **MTL-DBLSTM**: multi-task learning DBLSTM-based approach containing 2 BLSTM hidden layers with 512 nodes per layer and two independent output layers. The global cost weighting constant α is 0.9.
- **SOL-DBLSTM**: the proposed MTL DBLSTM with structured output layer approach containing 2 BLSTM hidden layers with 512 nodes per layer and one structured output layer. The weight α used in (7) is 0.9, and tanh function is used as activation function $\psi(\cdot)$. Selection of $\psi(\cdot)$ and α is elaborated in the next section.

The outputs of DNN-based system are acoustic features for speech synthesis that consist of MFCCs, BAPs, log F0, their dynamic counterparts (deltas and delta-deltas) and V/UV, totally 199 dimensions. For the other DBLSTM based approaches, the output contains all the features except the dynamic counterparts, totally 67 dimensions. For MTL-DBLSTM based approaches, the two output layers are for spectrum (MFCCs and BAPs) and pitch (log F0 and V/UV).

Specially, the output features from DNN based system are fed into maximum likelihood parameters generation (MLPG) [4] module with pre-computed variances for smoothing before synthesis. In DBLSTM based systems, MLPG post-processing is skipped. STRAIGHT vocoder is employed to synthesize speech with predicted acoustic features from different aforementioned approaches.

Backpropagation through time (BPTT) [25][26] is employed to train DBLSTM-based approaches by unfolding RNNs into standard feed-forward networks through time steps. Mini-batch-based Adam algorithm [27] is used as the optimizer with Keras [28] deep learning framework, which uses Theano [29] as backend, to implement and evaluate the different approaches.

4.2 Hyper-parameters in structured output layer

For proposed SOL DBLSTM-RNN model, the selection of activation function ψ as well as the value of α in (7) would dramatically influence the performance of predictions. Table 1 presents the objective evaluation of spectrum prediction using different activation functions. The tanh functions was found to be of the best performance outperforming other activation functions and had been chosen as the default ψ. Figure 2 illustrates the Mel-CD of spectrum features and RMSE of F0 when use different weighting constant α in training. As a balance, we use α=0.9 as the final value for the proposed system.

Table 1: Mel-CD (dB) of spectrum prediction with different activation function with α = 0.90.

ψ -activation				
Linear	Softmax	Sigmoid	ReLU	Tanh
5.3226	5.3423	5.3790	5.3041	5.2886

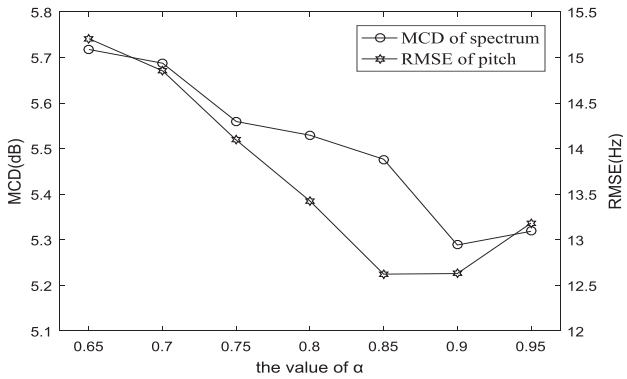

Figure 2: Mel-CD of spectrum prediction and RMSE of pitch prediction using different values of α

4.3 Objective evaluation

In objective evaluation, the generated features are assessed by comparing the distortions between the features extracted from natural speech in the test set and the generated ones predicted from different systems. Specifically, the duration extracted from natural speech is used directly in prediction.

As illustrated in Table 2, the conventional DBLSTM based system outperforms DNN baseline on MFCC prediction with 10% relative improvement. By using MTL style, the F0 trajectories generation gains a 6% relative improvement while the MFCC generation is on par. By employing SOL, the MFCC trajectories generation and F0 trajectories generation gain a further optimization with 1.3% and 3.1% relative improvement respectively over MTL-DBLSTM based approach. V/UV error drops from 4.3% with conventional one-task style to 4.2% with multi-task style. The BAP trajectories generation by SOL-DBLSTM based system is on par with that by other systems.

Table 2: Objective evaluation results of different features generated by aforementioned systems.

Systems	MFCC MCD(dB)	BAP MCD(dB)	F0 RMSE(Hz)	V/UV error(%)
DNN	5.9237	3.3689	13.6753	4.352
DBLSTM	5.3586	3.2665	13.8390	4.378
MTL-DBLSTM	5.3544	3.2590	13.0472	4.234
SOL-DBLSTM	5.2886	3.2517	12.6236	4.211

4.4 Subjective evaluation

Mean opinion score (MOS) is used to evaluate the perceived naturalness and quality of synthesized speech. 25 utterances from the test set are selected as the testing material. For each utterance, 4 synthetic speeches are generated from the aforementioned approaches and randomly shuffled to avoid preferential bias; the original natural speech is also available[1]. 10 native Chinese listeners with no reported listening difficulties are invited to score the synthetic speeches at 5-point scale in naturalness and speech quality by comparing the synthetic speeches with the natural one, in which the grades are standardized as 5 = Excellent (same as the natural speech), 4 = Good, 3 = Fair, 2 = Poor, 1 = Bad.

As illustrated in Figure 3, all the DBLSTM based systems outperform the baseline DNN based system. However, the MTL-DBLSTM based system is on par with the conventional DBLSTM based system for naturalness and quality. Compared with MTL-DBLSTM based system, the proposed SOL-DBLSTM based system gains relative improvements for naturalness and quality at 2.3% and 7.5% respectively.

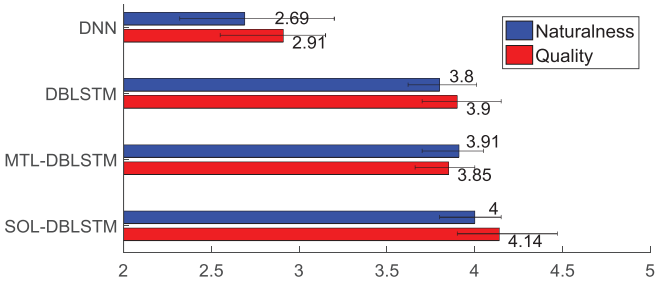

Figure 3: Results of MOS test for speech naturalness and quality, the original natural speech has set as 'Excellent'.

5. CONCLUSIONS

In the paper, we have proposed to use multi-task learned structured output layer in conventional DBLSTM where spectral parameter targets are conditioned upon pitch parameters prediction to improve the performance of TTS synthesis system. Experiments results show the proposed approach outperforms DNN based, DBLSTM based and MTL-DBLSTM based approaches on pitch prediction and spectrum prediction. Objective results illustrate using SOL is helpful in improving the trajectories generation and subjective results show the improvements in naturalness and speech quality.

This work indicates the possibility to use different but related tasks in training a better acoustic model for TTS with the SOL framework. In the future, more related tasks could be investigated in the framework for generating more animated speeches carrying different characteristics such as emphasis, interactive styles, etc.

6. ACKNOWLEDGEMENTS

This work is supported by National Natural Science Foundation of China - Research Grant Council of Hong Kong (NSFC-RGC) joint fund (61531166002, N_CUHK404/15), National High Technology Research and Development Program of China (2015AA016305), National Social Science Foundation of China (13&ZD189) and NSFC (61375027, 61433018).

[1]Samples are accessible at
http://mjrc.sz.tsinghua.edu.cn/demo/tts/icassp2017/

7. REFERENCES

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

[2] A. Hunt and A. Black, “Unit selection in a concatenative speech synthesis system using a large speech database,” [in] Proc. ICASSP, pp. 373-376, 1996.

[3] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, “Simultaneous Modeling of Spectrum, Pitch and Duration in HMM Based Speech Synthesis,” Proc. of EUROSPEECH, vol.5, pp.2347–2350, 1999.

[4] K. Tokuda, T. Kobayashi, T. Masuko, T. Kobayashi, and T. Kitamura, “Speech parameter generation algorithms for HMM based speech synthesis”, [in] Proc. ICASSP, pp. 1315-1318, 2000.

[5] G. Hinton, L. Deng, D. Yu, G.E. Dahl, A.R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath and B. Kingsbury, “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups.” IEEE Signal Processing Magazine, vol.29, no.6, pp.82-97, 2012.

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

[7] S. Kang and H. Meng. "Statistical parametric speech synthesis using weighted multi-distribution deep belief network." [in] Proc. InterSpeech, pp. 1959-1963. 2014.

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

[9] 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.

[10] Z. Wu, C. Valentini-Botinhao, O. Watts, and S. King, “Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis,” [in] Proc. ICASSP, pp. 4460-4463, 2015.

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

[12] A. Graves, and J. Schmidhuber, “Framewise phoneme classification with bidirectional LSTM and other neural network architectures,” Neural Networks, vol. 18, no. 5, pp. 602–610, 2005.

[13] A. Graves, A.-R. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” [in] Proc. ICASSP, 2013.

[14] Y. Fan, Y. Qian, F.L Xie, and F.K. Soong, “TTS synthesis with bidirectional LSTM based recurrent neural networks.” [in] Proc. InterSpeech, pp. 1964-1968. 2014.

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

[16] P. Swietojanski, P. Bell, and S. Renals. "Structured output layer with auxiliary targets for context-dependent acoustic modelling." [in] Proc. InterSpeech, pp. 1964-1967, 2015.

[17] R. Caruana, Multitask learning, Springer, 1998

[18] M. L. Seltzer, and J. Droppo, “Multi-task learning in deep neural networks for improved phoneme recognition,” [in] Proc. ICASSP, 2013.

[19] K.N. Stevens, Acoustic Phonetics, MIT Press, 2000, ISBN 0-262-69250-3, 978-0-262-69250-2.

[20] S. Arthi, and T. V. Sreenivas. "Influence of time-varying pitch on timbre: "Coherence and incoherence" based on spectral centroid." [in] Proc. ICASSP, pp. 4240-4244, 2015.

[21] Karimian-Azari, Sam, Nasser Mohammadiha, Jesper R. Jensen, and Mads G. Christensen. "Pitch estimation and tracking with harmonic emphasis on the acoustic spectrum." [in] Proc. ICASSP, pp. 4330-4334, 2015.

[22] E. Arisoy, M. Saraclar. “Compositional Neural Network Language Models for Agglutinative Languages.” [in] Proc. InterSpeech, 2016.

[23] 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.

[24] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency based F0 extraction: Possible role of a repetitive structure in sounds,” Speech communication, vol. 27, no. 3, pp. 187–207, 1999.

[25] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” Cognitive modeling, 1988.

[26] 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.

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

[28] F. Chollet, Keras [OL]. [2016-09-04]. GitHub repository. https://github.com/fchollet/keras.

[29] 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.