INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Expressive Speech Driven Talking Avatar Synthesis with DBLSTM using Limited Amount of Emotional Bimodal Data

Xu Li[1,2], *Zhiyong Wu*[1,2,3], *Helen Meng*[1,3], *Jia Jia*[1,2], *Xiaoyan Lou*[4], *Lianhong Cai*[1,2]

[1] Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems,
Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China
[2] Tsinghua National Laboratory for Information Science and Technology (TNList),
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[3] Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR, China
[4] Beijing Samsung Telecom R&D Center, Beijing 100081, China

dongfangyixi@gmail.com, {zywu,hmmeng}@se.cuhk.edu.hk, jjia@tsinghua.edu.cn,
xiaoyan.lou@samsung.com, clh-dcs@tsinghua.edu.cn

Abstract

One of the essential problems in synthesizing expressive talking avatar is how to model the interactions between emotional facial expressions and lip movements. Traditional methods either simplify such interactions through separately modeling lip movements and facial expressions, or require substantial high quality emotional audio-visual bimodal training data which are usually difficult to collect. This paper proposes several methods to explore different possibilities in capturing the interactions using a large-scale neutral corpus in addition to a small size emotional corpus with limited amount of data. To incorporate contextual influences, deep bidirectional long short-term memory (DBLSTM) recurrent neural network is adopted as the regression model to predict facial features from acoustic features, emotional states as well as contexts. Experimental results indicate that the method by concatenating neutral facial features with emotional acoustic features as the input of DBLSTM model achieves the best performance in both objective and subjective evaluations.

Index Terms: Expressive talking avatar, emotion, lip movement, facial expression, deep bidirectional long short-term memory (DBLSTM)

1. Introduction

Talking avatar has been widely used in many human-computer interaction areas such as virtual host / tutor, voice agent, remote telecommunication, etc. An animated face model has much to offer in addition to acoustic speech [1]. Besides lip movements, emotional facial expressions can further enhance interaction through nonverbal communication [2].

Several studies indicate that articulators behave differently under the influence of different emotions [3][4]. Furthermore, such influence varies depending on different phonetic contexts. For example, the lip shape and movements of speaking “what?” are different for happy or sad emotions with different facial expressions. More influences of emotions can be observed for the open lip shape of articulating “[aa]” than the close lip shape of articulating “[w]”. Hence, how to model the interactions between emotional facial expressions and lip movements talking into account contextual information is essential for generating expressive talking avatar.

In some previous studies [5][6], expressive talking avatar synthesis is achieved by augmenting existing neutral avatar generation system. The emotional facial frames or parameters with different expressions are learned from emotional audio-visual corpus and then simply attached to the face of neutral avatar. The problem is that such method has ignored the above interactions by modeling lip movements (with neutral avatar) and facial expressions separately. To address the problem, several statistical approaches with regression models have been proposed, such as support vector regression (SVR) [7], neural networks (NNs) [8], hidden Markov models (HMMs) [9], etc. To train such regression models, large speech corpus containing different emotions are generally required. However, recording large quantity of speech with various emotions is obviously cost-inefficient and time consuming. On the other hand, we have already accumulated a great deal of large-scale neutral audio-visual corpora. How to find a better way to make full use of such large-scale neutral corpus as complementary to the relatively small-sized emotional corpus for expressive talking avatar synthesis needs further investigation.

In this paper, several regression methods for speech driven talking avatar are proposed and tested to address the above issue. Inspired by the promising performance of deep bidirectional long short term memory (DBLSTM) in recent studies [10][11], we apply it as the regression model to incorporate contextual information with acoustic features and emotional states. Based on DBLSTM, we propose five methods to explore different possibilities in capturing the interactions using a large-scale neutral corpus in addition to a small size emotional corpus with limited amount of data. The difference between five methods is their way to utilize neutral information. Concretely, in method (a), DBLSTM network is trained with emotional corpus only; method (b) and (c) capture neutral and emotional information simultaneously by training a single DBLSTM network; while method (d) and (e) capture neutral information by a separate DBLSTM network in addition to emotional DBLSTM.

The rest of this paper is organized as follows. Section 2 presents data description and feature extraction. Details of the proposed methods are given in Section 3. Experiments and results are presented in Section 4. Section 5 concludes the paper and gives future directions.

http://dx.doi.org/10.21437/Interspeech.2016-364

2. Data description

2.1. Emotional and neutral bimodal corpus

The neutral corpus adopted in this work contains 321 neutral utterances recorded by a female native English speaker. Each utterance lasts for 3 to 4 seconds.

For emotional corpus, we use eNTERFACE'05 emotion database [12]. 44 subjects reads sentences in 6 basic emotions including anger, disgust, fear, happiness, sadness and surprise. Each emotion category contains 5 different sentences. As none of the subjects is actor or actress, their facial expressions are unprofessional and differ from one another. This leads to the challenges in expressive talking avatar synthesis with the data. The duration is about 4 seconds for each recorded utterance.

Such configuration of neutral-emotional corpora reflects the real situation, where neutral contains high-quality and sufficient amount of data while emotional corpus contains low-quality because of the difficulty in collecting.

2.2. Acoustic and visual features

As for acoustic features, we choose the 384 dimensional feature set of the *INTERSPEECH 2009 Emotion Challenge* [13], which contains 16 low level descriptors (LLDs) and their first order delta coefficients (32 dimensions in total) and 12 functionals. LLD features are extracted using openSMILE [14] with frame length of 40ms and frame shift of 10ms.

As for visual features, we adopt facial animation parameters (FAPs) defined by MPEG-4 specification [16] for talking avatar animation, which include 68 FAPs with 66 low-level ones and 2 high-level ones. The low-level FAPs represent a complete set of basic facial actions; while the 2 high-level FAPs represent visemes and expressions respectively. All low-level FAPs are standard values and expressed in terms of the facial animation parameter units (FAPUs), which allow the interpretation of FAPs across subjects. In this work, we focus on 46 low-level FAPs related to lip shape (21 FAPs), head movement (3) and expressions (22); and use visageSDK [15] to extract them from raw video data. Linear interpolation is used to interpolate FAPs to match the frame rate of LLDs.

3. Methods

In speech driven talking avatar, articulatory parameters are mainly determined by acoustic input. Contextual information should be considered to model the coarticulation phenomena. While in expressive talking avatar, articulatory parameters are also affected by facial expressions related to different emotional states; and such interactions between emotional states, facial expressions and lip movements change over time depending on the acoustic and contextual information. Concretely, regression model should be designed to dig out such relationships from training corpus. In this work, we adopt DBLSTM derived regression models to predict facial parameters.

Five methods are designed for the purpose of exploring the possibility of utilizing neutral corpus information to improve the performance of expressive talking avatar synthesis. The structures of the five methods are shown in Fig.1.

3.1. Method (a): Emotional data trained network

To predict expressive facial parameters from acoustic, the most straightforward way and state-of-the-art method as we know is to train a regression or mapping model using emotional corpus with different emotional states. As illustrated in Fig. 1(a), the regression model (i.e. the emotional DBLSTM) learns temporal varied interactions between facial features and emotion states taking the acoustic features and contextual information into account. Let X_E be the input emotional acoustic features (E-LLD), Y_E be the target expressive facial features (E-FAPs), the learning problem is to find an optimized emotional regression model F_E satisfying:

$$\arg\min_{F_E} \|F_E(X_E) - Y_E\|^2. \quad (1)$$

However, to train such regression model, large corpus with different emotions are required. For this paper's scope with only limited amount of emotional bimodal data, the training result of DBLSTM may suffer from the insufficient training data leading to poor performance.

3.2. Method (b): Mixed data trained network

To make use of neutral corpus, one way is to mix the data from both neutral and emotional corpora together and train the model using the mixed data. As Fig. 1(b) illustrates, the emotional DBLSTM regression model F_E is trained by feeding the training data randomly selected from either neutral corpus or emotional corpus:

$$\arg\min_{F_E} \|F_E(X_{NE}) - Y_{NE}\|^2, \quad (2)$$

where X_{NE} and Y_{NE} denote the LLD and FAPs characterizing the training data from the mixed neutral-emotional corpus.

The optimization of (2) may give us a compromised result of regression model between neutral and emotional corpus. However, the amount of neutral data is much larger than that of emotional data for each emotion category. The unbalanced distribution between neutral and emotional data may lead to unexpected result of neutral facial expression while emotional expression is desired.

3.3. Method (c): Retrained network

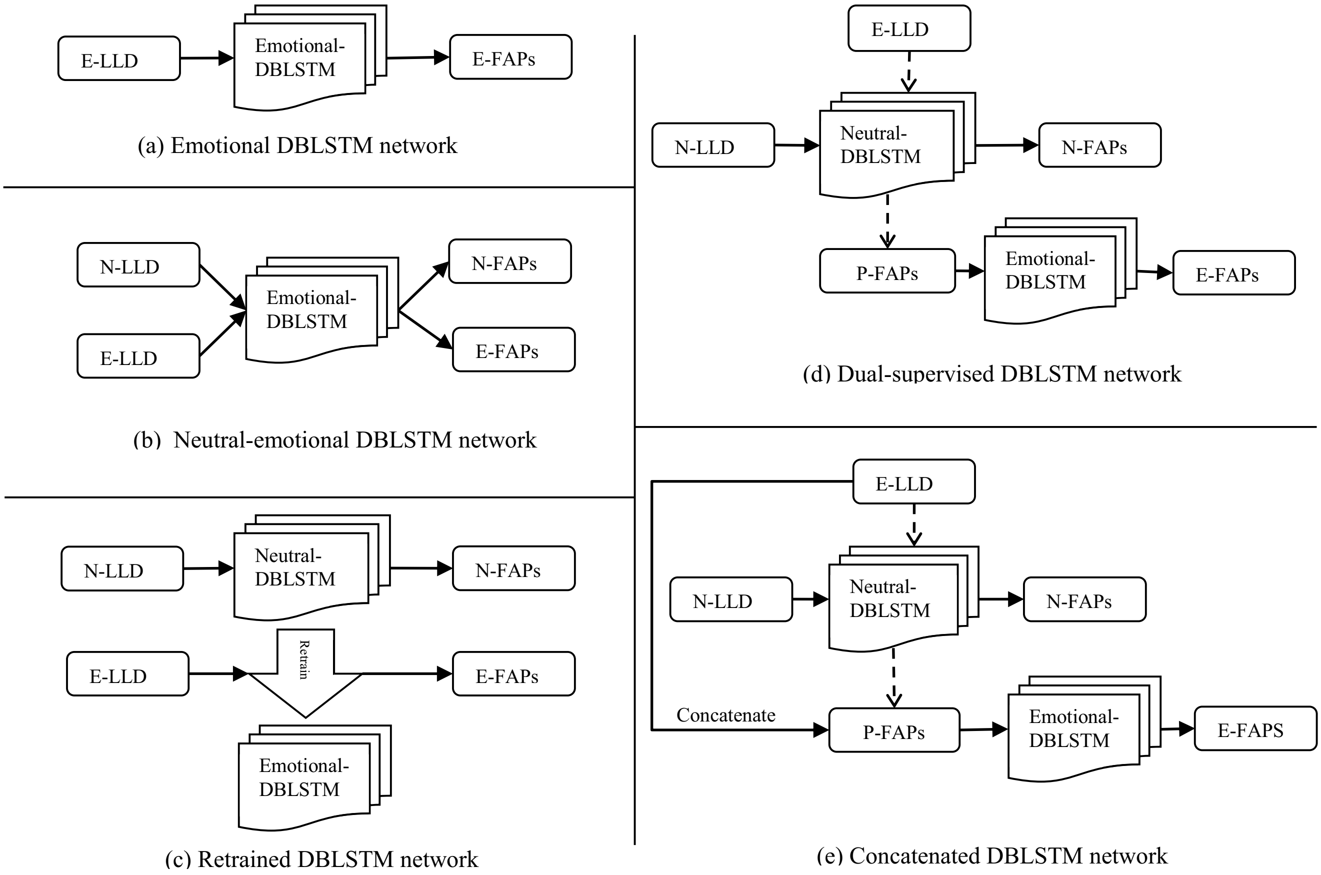
To address the problem in Method (b), we propose another method that can manually adjust the proportion of the influence between neutral and emotional data. As shown in Fig. 1(c), the DBLSTM network is first trained with neutral corpus, and then re-trained (or fine-tuned) in few epochs using the data from emotional corpus to adapt the parameters of initial network into emotional model. The number of the re-train epochs can be manually tuned to adjust the influence of emotional corpus. The method can be formulated as:

$$\arg\min_{F_{N\to E}} \|F_{N\to E}(X_N) - Y_N\|^2, \quad (3)$$

$$\arg\min_{F_{N\to E}} \|F_{N\to E}(X_E) - Y_E\|^2, \quad (4)$$

where $F_{N\to E}$ represents the retrained DBLSTM network, X_N and Y_N be neutral acoustic features (N-LLD) and neutral facial features (N-FAPs) respectively. At the first training stage of (3), nodes in $F_{N\to E}$ sensitive to neutral information are activated. At the second stage of (4), nodes related to emotional information are activated. For nodes sensitive to both corpora, their weights are tuned by both (3) and (4). The degree of such coverage is determined by the number of epochs in the second stage.

By choosing appropriate epoch number, we can get an ideal model balanced between neutral and motional data to achieve better performance than with only emotional data. Whereas, it is quite difficult to determine a proper epoch number as it may



(a) Emotional DBLSTM network

(b) Neutral-emotional DBLSTM network

(c) Retrained DBLSTM network

(d) Dual-supervised DBLSTM network

(e) Concatenated DBLSTM network

Fig. 1. Structures of the five proposed networks.
(The solid-line arrows represent the training process. The dotted-line arrows represent the predicting process.)

vary due to the different quality and quantity of neutral and emotional data. Besides, manual selection of epoch number may introduce unexpected subjective factors into the model.

3.4. Method (d): Dual-supervised network

To avoid manual factors in Method (c) and address unbalanced problem in Method (b), we further propose a dual-supervised DBLSTM network as Method (d). The intuition of this method comes from the idea that emotional facial expressions rely on neutral expressions. The latter provide prime facial animations such as neutral lip movements and expressions; and the former are derived by further deformations from neutral to expressive.

As depicted in Fig. 1(d), we first train a neutral DBLSTM network with N-LLD as input and N-FAPs as output. We then input E-LLD to the neutral network to get the predicted FAPs (P-FAPs). The P-FAPs are frame to frame aligned with the emotional E-FAPs directly extracted from the emotional corpus. The frame aligned P-FAPs and E-FAPs are finally taken as the input and output to train the emotional DBLSTM network:

$$\arg\min_{F_N} \|F_N(X_N) - Y_N\|^2, \tag{5}$$

$$\arg\min_{F_E} \|F_E(F_N(X_E)) - Y_E\|^2. \tag{6}$$

The information related to prime facial animations is derived from neutral corpus and captured by neutral DBSLTM network F_N through (5), which are then transferred to P-FAPs predicted by $F_N(X_E)$. By optimizing (6), F_E is considered to capture the deformation information from neutral to expressive expressions.

3.5. Method (e): Concatenated network

The emotional information implied in the emotional acoustic features is essential for expressive talking avatar. Inspired by this, we slightly modify Method (d) to derive a new Method (e), as in Fig. 1(e). In this method, the emotional E-LLD and the predicted P-FAPs are concatenated to serve as the input to train the emotional DBLSTM network with E-FAPs as the output:

$$\arg\min_{F_N} \|F_N(X_N) - Y_N\|^2, \tag{7}$$

$$\arg\min_{F_E} \left\| F_E \begin{pmatrix} F_N(X_E) \\ X_E \end{pmatrix} - Y_E \right\|^2. \tag{8}$$

Different from Method (d), we concatenated X_E with the predicted P-FAPs $F_N(X_E)$ in (8) to enhance emotional acoustic features' influences to the final expressive facial animations.

4. Experiments and results

4.1. Experimental setup

DBLSTM networks are trained with Theano [17] with Adam [18] as optimizer following the default parameter settings as suggested in [18]. The activation for the hidden BLSTM layers is tanh. The activation for the last regression layer is linear. All DBLSTM networks have three hidden layers with 100 units per layer. Dropout layer (dropout=0.3) is used to prevent overfitting. The dimension of LLD features is 384 and the dimension of FAP features is 46.

For emotional corpus, we use the first 38 subjects' first 4 utterances (with all 6 emotion categories) as training set, and the remaining 6 subjects' last utterance as test set. From taining set, 10 utterances are ramdomly selected as validation set. This makes the test set totally new to the regression model. The 321 utterances from neutral corpus are all added to the training set in Method (b). In Method (c), (d) and (e), the neutral corpus is used to train the neutral model with 10 utterances randomly selected as validation set and other utterances as training set.

4.2. Objective evaluation for different emotions

We adopt root mean squared error (RMSE) between predicted FAPs and the ground truth to evaluate the proposed approaches. Experimental results are shown in Fig. 2, where RMSE values of 5 different methods for 6 emotions are plotted. This reveals how different regression models work for different emotions.

The average RMSE results indicate concatenated DBLSTM network (Method (e)) achieves the best performance; and the performance of retrained DBLSTM network (Method (c)) is also acceptable. The results validate neutral corpus data provide beneficial information to expressive talking avatar generation.

For surprise and disgust emotions, the model trained with emotional data only (Method (a)) achieves the best result. On the contrary, for sadness emotion, Method (a) gets the worst performance while other methods work pretty well. The reason might be that exaggerated facial expressions and lip movements are necessary to express surprise and disgust emotions, while sadness emotion results in peaceful expressions. In this way, information from neutral corpus is more valuable for peaceful expressions than those exaggerated ones.

4.3. Frame-wise objective evaluation

We further conduct frame-wise comparison of RMSE values between the proposed Method (e) and the superposed method from [6]. The frame-wise RMSE values of a randomly selected utterance is shown in Fig. 3. As can be seen, RMSE values of the predicted results from Method (e) are smaller for most of the frames. The results indicate that the proposed methods are effective in modeling the interactions between emotional states, facial expressions and lip movements comprehensively.

4.4. Subjective evaluation

We further conduct subjective evaluations to test the predicted FAPs on a talking avatar synthesis system [16]. The predicted FAP values for the utterances in the test set are used to generate 3D talking avatar videos. 6 subjects are then asked to score each video based on a five-point scale '1' (bad), '2' (poor), '3' (fair), '4' (good) and '5' (excellent) according to the naturalness and expressiveness of the talking avatar. The video files generated from the 5 methods are presented in the same webpage. The subjects assign the scores using the homegrown rank software. Mean opinion scores (MOS) for each method under different emotion are shown in Fig. 4. As can be observed from the result, in average, concatenated DBLSTM network (Method (e)) performs the best in subjective evaluation which is in consistent with the objective evaluation, indicating the effectiveness of the proposed method.

5. Conclusion and future work

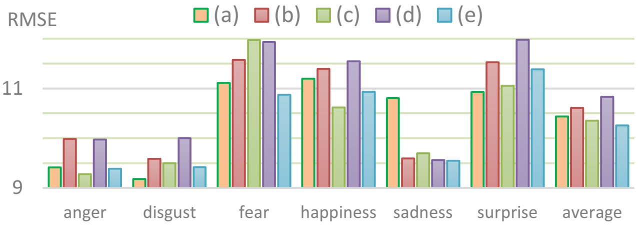


Fig. 2. RMSE for different methods under different emotions.

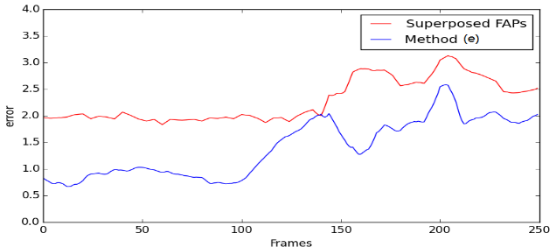


Fig. 3. Comparison of frame-wise RMSEs of different methods.

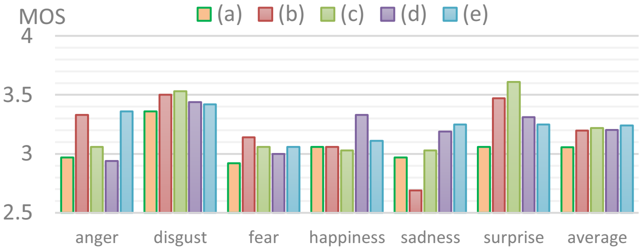


Fig. 4. MOS for different methods under different emotions.

In this paper, we propose several different regression models to predict emotional facial parameters from acoustic speech for expressive speech driven talking avatar. In expressive speech driven talking avatar, articulatory parameters are determined by acoustic input as well as affected by facial expressions with different emotional states. Such interactions between emotional states, facial expressions and lip movements change over time depending on acoustic and contextual information. To capture such complex interactions, we adopt DBLSTM as regression models. For limited amount of emotional data, we propose several methods to investigate the different possibilities in capturing the interactions using a large-scale neutral corpus in addition to a small size emotional corpus. Experiments indicate that utilizing neutral corpus can improve the performance of expressive talking avatar generation. In addition, we figure out the best method to employ the neutral corpus is concatenated DBLSTM network.

In the future, we plan to combine the different structures we proposed in this paper to synthesis facial animation and try to generate expressive facial features of all emotions in a single regression framework.

6. Acknowledgement

This work is supported by National Basic Research Program of China (2012CB316401), National High Technology Research and Development Program of China (2015AA016305), Natural Science Foundation of China (NSFC) (61375027, 61433018, 61370023 and 61171116), joint fund of NSFC-RGC (Research Grant Council of Hong Kong) (61531166002, N_CUHK404/15) and Major Program for National Social Science Foundation of China (13&ZD189).

7. References

[1] J. Cassell, “Embodied conversational agents: representation and intelligence in user interface,” AI magazine, 22(3): 67-83, 2001.

[2] B. Branstrom, D. House, “Multimodality and speech technology: verbal and non-verbal communication in talking agents,” in Proc. EUROSPEECH, pp. 2901-2904, 2003.

[3] E. Provost, I. Zhu, S. Narayanan, “Using emotional noise to uncloud audio-visual emotion perceptual evaluation,” in Proc. ICME, pp. 1-6, 2013.

[4] M. Caldognetto, P. Cosi, F. Cavicchio, “Modifications of speech articulatory characteristics in the emotive speech,” in Proc. ADS, pp. 233-239, 2004.

[5] M. Malcangi, “Text-driven avatar based on artificial neural networks and fuzzy logic,” Int. Journal of Computers, 4(2), 2010.

[6] J. Jia, Z. Wu, S. Zhang,H. Meng, L. Cai, “Head and facial gestures synthesis using PAD model for an expressive talking avatar,” Multimedia Tools and Applications, 73(1): 439-461, 2014.

[7] Y. Cao, W. Tien, P. Faloutsos, F. Pighin, “Expressive speech-driven facial animation,” ACM Trans Graph, 24: 1283-1302, 2005.

[8] P. Hong, Z. Wen, T. Huang, “Real-time speech-driven face animation with expressions using neural networks,” IEEE Trans Neural Netw, 13: 916-927, 2002.

[9] Y. Zhang, Q. Ji, Z. Zhu, B. Yi, “Dynamic facial expression analysis and synthesis with MPEG-4 facial animation parameters,” IEEE Trans Circ Syst Video Technol, 18: 1383-1396, 2008.

[10] S. Hochreiter, J. Schmidhuber, “Long short-term memory,” Neural Computation, 9: 1735-1780, 1997.

[11] B. Fan, L. Wang, F. Soong, L. Xie, “Photo-real talking head with deep bidirectional LSTM,” in Proc. ICASSP, pp. 4884-4888, 2015.

[12] O. Martin, I. Kotsia, B. Macq, I. Pitas, “The eNTERFACE'05 audio-visual emotion database,” in Proc. ICDEW, 2006.

[13] B. Schuller, S. Steidl, A. Batliner, “The INTERSPEECH 2009 emotion challenge,” in Proc. INTERSPEECH, pp. 312-315, 2009.

[14] F. Eyben, F. Weninger, F. Gross, B. Schuller, “Recent developments in openSMILE, the Munich open-source multimedia feature extractor,” in Proc. ACM MM, pp. 835-838, 2013.

[15] Visage | SDK - Face tracking tools [OL]. [2015-07-10]. http://www.visagetechnologies.com/products/visagesdk.

[16] Z. Wu, S. Zhang, L. Cai, H. Meng, “Real-time synthesis of Chinese visual speech and facial expressions using MPEG-4 FAP features in a three-dimensional avatar,” in Proc. ICSLP, pp. 1802-1805, 2006.

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

[18] D. Kingma, J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.