LOW LEVEL DESCRIPTORS BASED DBLSTM BOTTLENECK FEATURE FOR SPEECH DRIVEN TALKING AVATAR

Xinyu Lan[1,2], Xu Li[1,2], Yishuang Ning[1,2], Zhiyong Wu[1,2,3], Helen Meng[1,3], Jia Jia[1,2], Lianhong Cai[1,2]

[1]Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems,
Shenzhen Key Laboratory of Information Science and Technology,
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

{syslxy1991, dongfangyixi}@gmail.com, ningys13@mails.tsinghua.edu.cn
{zywu, hmmeng}@se.cuhk.edu.hk, {jjia, clh-dcs}@tsinghua.edu.cn

ABSTRACT

Speech is bimodal in nature. There are close correlations between the acoustic speech signals and the visual gestures such as lip movements, facial expressions and head motions. For speech driven talking avatar, how to derive more representative acoustic features from which to predict more accurate and realistic visual gestures still remains the research problem. Inspired by the promising performance of low level descriptors (LLD) in speech emotion recognition, in this work, we investigate the usage of LLD feature for the task of speech driven talking avatar. Furthermore, visual gestures also demonstrate correlations with not only context information of past or future acoustic features (e.g. anticipatory co-articulation phenomena) but also textual information (e.g. textual hints for lip movement). To incorporate such information, we also propose to use deep bidirectional long short-term memory (DBLSTM) as the bottleneck feature extractor, which can combine LLD feature with contextual information. Experimental results indicate that the proposed LLD based DBLSTM bottleneck feature outperforms the conventional spectrum related features for the task of speech driven talking avatar, and more sophisticated contextual information can further improve the performance.

Index Terms—bottleneck feature, deep bidirectional long short-term memory (DBLSTM), low level descriptors (LLD), talking avatar

1. INTRODUCTION

Talking avatar has drawn extensive attention for its wide use in human-computer interaction fields, e.g. voice agent, virtual teacher or host, intelligent computer assistant, etc. Speech is bimodal in nature. The visual gestures such as lip movements, facial expressions and head motions are closely correlated with the acoustic speech. Hence lots of research interests have been devoted to find and model such correlations.

To build a speech driven talking avatar, extracting appropriate acoustic features is very important. Traditional spectrum related features such as mel-frequency cepstral coefficients (MFCC) were widely used in facial animation and head motion generation [1-3]. Recent works have mainly focused on developing machine learning models for the audio-visual mapping problem, but put less emphasis on how to find more representative acoustic features for more realistic and expressive visual gesture generation. Speech emotion recognition provides us some hints. [4] and [5] chose pitch (F0), root mean square energy (RMSE) and formants as prosodic features to perform affect recognition. Afterwards, low level descriptors (LLD) attracted lots of attentions and achieved state-of-the-art performance. [6] and [7] developed systems for emotion classification from LLD features. Inspired by the promising performance of LLD in speech emotion recognition, we investigate if LLD could outperform the traditional features in the speech driven talking avatar task.

Most previous studies on talking avatar have focused on facial animation (lip movements and/or facial expressions) only. Some photo-real methods reconstructed face images with lip movement from principle component analysis (PCA) based visual features [8]. [9] adopted active appearance model (AAM) features involving both shape and texture information for more realistic lip animation. [10] used motion unit parameters (MUP) for talking face animation with expression. Recently, lots of work rendered animation using the MPEG-4 facial animation parameters (FAPs) [11][12], where FAPs offer a parameterization approach for the animation of eyes, mouth, tongue, teeth, head motion, etc. However, most works have omitted the head motions that often occur simultaneously with lip movements and facial expressions for an expressive and realistic talking avatar. Although some studies have tried to generate head motions [13][14], the facial animation and head motion are still modeled separately regardless the close correlations between them. In this work, we adopt FAPs as the visual features and predict lip movements, facial expressions and head motions simultaneously from acoustic speech with a single shared regression model.

Visual gestures at a particular time step are also correlated with the context information of past or future acoustic features (e.g. the anticipatory co-articulation phenomena). How to model such correlation provides another challenge. In previous work, there are mainly two streams of approaches. The first one is hidden Markov model (HMM) based method [2][8] motivated by the ideas from automatic speech recognition (ASR). In this method, the textual context information can be easily incorporated by state-transition probabilities. The second one is the direct acoustic to visual feature mapping approach using long short-term memory (LSTM) [15][9], bidirectional LSTM (BLSTM), etc. These LSTM derived models have shown superior performance over HMM with capabilities in capturing long-range context information [9]. Whereas, the textual

information such as phoneme labels that are most valuable for lip movements and head motions are not well considered in these models. On the other hand, deep bidirectional LSTM (DBLSTM) as the probabilistic feature extractor, can involve context of both acoustic features and phoneme labels during feature extraction [16][17]. In this paper, we propose a LLD based DBLSTM bottleneck feature that takes into account not only the contextual acoustic feature correlations but also the textual information.

The rest of the paper is organized as follows. Section 2 gives an overview of our system architecture. Detailed acoustic visual features and their extraction method are listed in Section 3. Section 4 describes the LLD based DBLSTM bottleneck feature and the training method of the DBLSTM for bottleneck feature extraction. Section 5 discusses the objective and subjective experiments comparing the performance of different features and network architectures. Finally, our conclusions are drawn in Section 6.

2. SYSTEM FRAMEWORK

Fig. 1 shows the block diagram of the proposed system, which involves a training stage for two DBLSTMs and a prediction stage. The first DBLSTM is for bottleneck feature extraction and the second DBLSTM is for bottleneck to visual feature mapping.

In the training stage, given the audio visual bimodal corpus, we extract acoustic features and visual features (FAPs). Meanwhile, forced alignment between acoustic features and contextual labels (e.g. phoneme labels) is performed with a homegrown HMM based speech recognizer. The DBLSTM feature extractor is then trained with cross entropy error; thus a discriminative mapping between the acoustic feature and latent contextual sequence is established. Thereafter, we can derive the DBLSTM bottleneck feature. Then the second DBLSTM is trained to learn the regression model between the bottleneck feature and FAPs.

In the prediction stage, given an input speech utterance, the DBLSTM feature extractor first extracts bottleneck features from the raw acoustic features, and then the DBLSTM mapping model predicts the FAP sequences from the bottleneck features. Finally, the visual gestures including lip movements, facial expressions and head motions are reconstructed on a 3 dimensional avatar using the technologies of our previous work [12][13][21].

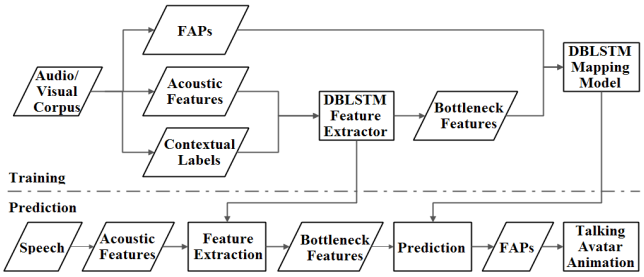

Fig 1. System framework

Table 1. 16 low level descriptors used

Feature Group	Features in Group	No.
RMSE	Root mean square signal frame energy	0
MFCC	Mel-frequency cepstral coefficients	1~12
PCM zcr	Zero-crossing rate of time signal	13
Voice Probability	Voicing probability computed from the autocorrelation functions (ACF)	14
F0	The fundamental frequency computed from the Cepstrum	15

Table 2. 12 functionals applied to LLD contours

Functionals	No.
The max/min value of the contour	1~2
Range (max – min)	3
The absolute position of the max/min value (in frames)	4~5
The arithmetic mean of the contour	6
The slope of a linear approximation of the contour	7
The offset of a linear approximation of the contour	8
The quadratic error computed as the difference of the linear approximation and the actual contour	9
The standard deviation of the values in the contour	10
The skewness and the kurtosis	11~12

3. ACOUSTIC AND VISUAL FEATURES

3.1. Low level descriptors (LLD)

Taking advantage of the attracting performance of LLD in emotion recognition task, we chose a 384 dimensional acoustic feature vectors that served as the feature set of the *INTERSPEECH 2009 Emotion Challenge* [22], which contains 16 low level descriptors and their first order delta regression coefficients (32 dimensions in total) and 12 functionals. Table 2 lists the statistical functionals that are applied to the low level descriptors shown in Table 1.

3.2. Facial animation parameters (FAPs)

The MPEG-4 specification defines totally 68 FAPs, including 66 low-level FAPs and 2 high-level FAPs [21]. The low-level FAPs, based on the movements of facial definition points, 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 the FAPs to fit with any face models. MPEG-4 standard defines a neutral face and all the FAPs are expressed as displacements from the positions defined in the neutral face.

Our work focuses on controlling the facial definition points directly. The 2 high-level FAPs are not considered. We adopt the open source toolkit visageSDK [18] to extract FAPs from raw video data. Some of the parameters related to tongue, teeth, nose and ears currently cannot be reliably estimated and do not affect the animation of talking avatar on our following prediction steps. The values of these FAPs are simply set to be zeros.

4. DBLSTM BOTTLENECK FEATURE

4.1. DBLSTM bottleneck feature extractor

As a kind of recurrent neural network (RNN), long short term memory (LSTM) has been demonstrated to be one of the most effective architectures to map the long-term history of inputs to the current output by solving the vanishing gradient problem that traditional RNN faces. To retrieve both past and future contextual information, bidirectional LSTM (BLSTM) with two separate hidden layers scanning the input sequences in both directions has

been popular recently [20]. Inspired by its promising performance in sequence classification and regression, deep BLSTM (DBLSTM) has been developed to extract potential features from input features. This paper proposes a LLD based DBLSTM bottleneck feature extractor, which accepts LLD acoustic feature vectors and outputs potential contextual sequences, to extract the bottleneck feature.

Fig. 2 illustrates the network structure in detail. The feature extractor can be divided into two parts, namely training stage and extracting stage. During the training stage, we train the DBLSTM as a speech-based recognizer on the training dataset. LLD acoustic feature serves as the input of DBLSTM network, while the frame wise contextual label serves as the output. In this work, two kinds of contextual labels (phoneme label and HMM state label) are involved to generate two contextual levels of bottleneck features. The frame wise correspondence information are generated by the forced alignment process as described in Section 2. Three hidden LSTM layers are used for both forward and backward directions. Taking advantage of the results from previous work, we adopted two hidden layers and a bottleneck layer with stationary size of 40 [17]. In the extracting stage, the activations of the output layer of DBLSTM are ignored, as we focus on the output of the bottleneck layers. As shown in Fig. 2, the final 80 dimensional bottleneck feature is generated by combining the outputs of the two directional bottleneck layers of DBLSTM.

4.2. Training method

Training a DBLSTM feature extractor can be regarded as learning a special kind of neural network for speech recognition from the LLD acoustic features to the contextual label sequences. It can be trained to minimize the cross-entropy error between the predicted contextual sequence C_k^p and the ground truth C_k^o. The loss function of the kth sequence can be:

$$H_k\left(C_k^o, C_k^p\right) = -\sum_n c_{nk}^o \log c_{nk}^p , \quad (1)$$

where c_{nk}^o and c_{nk}^p are the nth dimension of ground truth contextual label vector and predicted one separately. In addition, we adopt one-hot representation (a vector with a specific dimension set to 1 while all other dimensions 0) for the contextual label (phoneme label and/or HMM state label).

We feed forward the DBLSTM bottleneck feature extractor like traditional DBLSTM and adopt the back propagation through time (BPTT) algorithm to train the network [19]. By applying the chain rule, we could obtain that:

$$\frac{\partial H\left(w_{ij}\right)}{\partial w_{ij}} = \sum_{t=1}^{T} \frac{\partial H\left(w_{ij}\right)}{\partial a_j^t} \frac{\partial a_j^t}{\partial w_{ij}} , \quad (2)$$

where ∂w_{ij} denotes the weight unit i and unit j, and a_j^t indicates the input activations of unit j at time t. Then we could feed back the network error and train the extractor as traditional DBLSTM.

5. EXPERIENMENTS

5.1. Experimental setup

This work adopted an emphatic audio visual database with 700 English utterances spoken by a female, including 350 emphatic and 350 neutral utterances, for experimentation. The video frame rate is 25fps and well formed in AVI format. We divided the whole corpus into 3 parts randomly, 600 utterances as training set, 60 as test set and others as validation set.

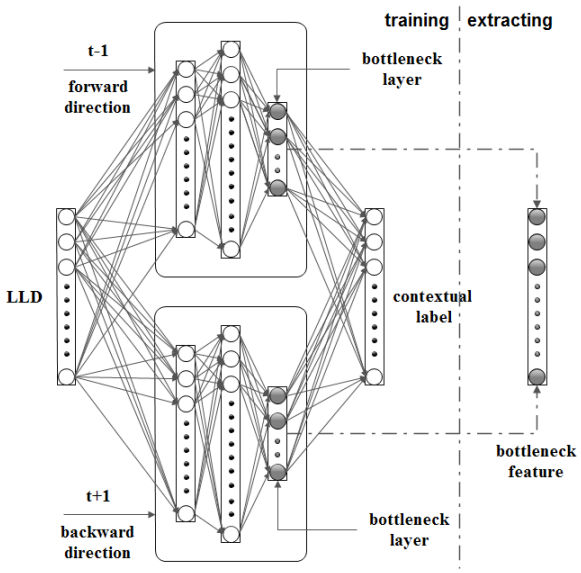

Fig 2. DBLSTM bottleneck feature extractor

To evaluate the proposed approach, we take the advantage of the well-known objective evaluations criterion root mean squared error (RMSE) and correlation coefficient (CORR) between the predicted FAPs and the ground truth. These metrics are defined as:

$$RMSE = \frac{\sum_{k=1}^{M_T}\sum_{t=1}^{T_k}\sqrt{\left\|f_{tk}^o - f_{tk}^p\right\|^2 / N_{fap}}}{\sum_{k=1}^{M_T} T_k} , \quad (3)$$

$$CORR = \frac{\sum_{k=1}^{M_T}\sum_{t=1}^{T_k} corr\left(f_{tk}^o, f_{tk}^p\right)}{\sum_{k=1}^{M_T} T_k} , \quad (4)$$

where M_T denotes the scale of the test set, T_k is the frame number of the kth test set, f_{tk}^o and f_{tk}^p mean the frame wise ground truth and the predicted result separately, *corr*(.) indicates the correlation coefficient between the two vectors of each frame.

Different architecture of the DBLSTM mapping model may affect the prediction accuracy from acoustic features to FAPs. To extensively explore the performances of different acoustic features, we conducted experiments on several network topologies for DBLSTM mapping model with different numbers of hidden layers and numbers of hidden units per hidden layer.

For both the DBLSTM networks, the learning rate is set to be 1e-4, and the momentum is 0.9. In addition, we choose the steepest optimizer and consider the network to be the best when it meets the stopping criterion.

5.2. Objective experiment on LLD and MFCC-RMSE

To compare the performance of the proposed LLD feature and the traditional feature, a 39 dimensional MFCC-RMSE feature containing 12 order MFCCs and RMSE together with their first and second order derivatives was used as the baseline. The MFCC-RMSE and LLD were extracted with 25ms frame size and 10ms frame shift. The dimension of the LLD acoustic feature is 384.

Experimental results are shown in Table 3, where the second column means different network structures of DBLSTM mapping model with one (100), two (100-100) or three hidden layers (100-100-100) respectively. The hidden unit numbers per hidden layer

are all set to 100. As can be seen, MFCC-RMSE performs better than LLD on the condition that BLSTM structure is simple, such as one or two bidirectional hidden layers. While when the network goes deeper with three hidden layers, LLD outperforms MFCC-RMSE significantly indicating that high dimensional LLD carries more information that can be learned with complex deep network.

5.3. Objective experiment on bottleneck features

Furthermore, to compare the performance of different contextual levels of bottleneck features, we trained two kinds of DBLSTM bottleneck feature extractors for extracting phoneme level and HMM state level bottleneck features respectively. The phoneme level bottleneck feature was extracted by a well-trained DBLSTM with the output layer of 41 dimensional phoneme labels, while the HMM state level bottleneck feature was extracted with the output layer of 123 (41×3) dimensional HMM state label.

As for the structure of DBLSTM bottleneck feature extractor, we adopted a commonly used DBLSTM structure. The network contains three bidirectional hidden layers, with 100, 100 and 40 hidden units for each hidden layer respectively [17]. The structure of DBLSTM mapping model is the same as the above experiment.

Experimental results are shown in Table 4. The first column indicates different kinds of bottleneck features. MFCC-phoneme (LLD-phoneme) is the MFCC-RMSE (LLD) based phoneme level bottleneck feature where the DBLSTM feature extractor is trained with MFCC-RMSE (LLD) and phoneme labels; while MFCC-state (LLD-state) means the DBLSTM is trained with MFCC-RMSE (LLD) and HMM state labels. From Table 3 and Table 4, we can find that, for the same architecture of DBLSTM mapping model, bottleneck feature shows superior performance for both RMSE and CORR. Furthermore, we can see that bottleneck features at state level perform best for the complex and deep network structure of DBLSTM mapping model.

5.4. Subjective evaluation

We further conducted a set of subjective experiments to evaluate the naturalness of synthetic animation of the talking avatar driven by the above mentioned four different acoustic features.

10 speech utterances were randomly selected from the test set and used to generate synthetic lip movements, facial expressions as well as head motions on a 3D talking avatar. The synthetic visual animation together with the acoustic speech input were saved as 10 video files. 20 subjects were asked to watch the video file and then assign a score at 5-point scale for each file based on naturalness and synchronization between the visual animations and acoustic speech. Higher score means more natural and closer synchrony. The mean opinion score (MOS) over 10 speech utterances and 20 subjects are computed and presented in Table 5. As can be seen, the MOS score of raw LLD feature is a bit higher than MFCC-RMSE feature, while LLD based bottleneck features lead to further performance improvement. Furthermore, LLD based bottleneck feature of state labels achieves the best MOS score. The results suggest that our proposed LLD based bottleneck feature is able to capture inherent information of lip movements, facial expressions and head motions simultaneously from acoustic speech signals, and can achieve more nature speech driven talking avatar.

6. CONCLUSION

In this paper, we investigated new acoustic features and proposed LLD based DBLSTM bottleneck feature for speech driven talking avatar. Our work shows that LLD can enhance the performance of regression model for audio visual speech mapping. In addition, the bottleneck feature shows strong representation ability and gets further performance improvement. Specifically, bottleneck feature involving HMM state contextual label information performs better than the bottleneck feature with phoneme label information. In the future, we plan to investigate in detail the performance of different LLD feature sets for the task.

Table 3. Results for LLD and MFCC-RMSE, where "Map Model" means network structure of DBLSTM mapping model

Acoustic Feature	Map Model	RMSE	CORR
MFCC-RMSE	100	199.0083	0.5667
MFCC-RMSE	100-100	108.3619	0.6259
MFCC-RMSE	100-100-100	108.2610	0.6607
LLD	100	224.5075	0.5637
LLD	100-100	115.3336	0.6144
LLD	**100-100-100**	**91.1433**	**0.6824**

Table 4. Results for different bottleneck features, where "Map Model" mains network structure of DBLSTM mapping model

Bottleneck Feature	Map Model	RMSE	CORR
MFCC-phoneme	100	161.9562	0.6209
MFCC-phoneme	100-100	103.5360	0.6209
MFCC-phoneme	100-100-100	93.9756	0.6439
LLD-phoneme	100	186.6465	0.6336
LLD-phoneme	100-100	103.5785	0.6750
LLD-phoneme	**100-100-100**	**89.2479**	**0.6969**
MFCC-state	100	159.7956	0.5776
MFCC-state	100-100	118.3734	0.6797
MFCC-state	100-100-100	96.2563	0.6507
LLD-state	100	176.0278	0.5769
LLD-state	100-100	87.8979	0.6602
LLD-state	**100-100-100**	**77.2061**	**0.7242**

Table 5. Results for subjective evaluation

Acoustic feature	MOS score
MFCC-RMSE	3.3
LLD	3.4
LLD-phoneme	3.6
LLD-state	**3.8**

7. ACKNOWLEDGEMENT

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

8. REFERENCES

[1] E. Yamamoto, S. Nakamura, K. Shikano, "Lip movement synthesis from speech based on hidden Markov models," *Speech Communication*, vol. 26, pp. 105-115, 1998.

[2] G. Wang, M. Yang, C. Chiang, W. Tai, "A talking face driven by voice using hidden Markov model," *Journal of information science and engineering*, vol. 22, pp. 1059-1075, 2006.

[3] N. Nalini, A. Chakraborty. "Speech emotion recognition using MFCC and AANN," in *Proc. International Conference on Engineering and Technology*, pp. 223-225, 2013.

[4] Z. Zeng, J. Tu, M. Liu, T. Huang, B Pianfetti, D. Roth, S. Levinson, "Audio-visual affect recognition," *IEEE Transactions on Multimedia*, vol. 9, pp. 424-428, 2007.

[5] M. Song, J. Bu, C. Chen, N. Li, "Audio-visual based emotion recognition-a new approach," *Computer Vision and Pattern Recognition* CVPR, vol. 2, pp. 1020-1025, 2004.

[6] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, V. Aharonson, "The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals," in *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, pp. 2253-2256, 2007.

[7] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Cowie, R. Cowie, "Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies," in *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, pp. 597-600, 2008.

[8] L. Wang, X. Qian, W. Han, F. Soong, "Synthesizing photo-real talking head via trajectory-guided sample selection," in *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, vol. 10, pp. 446-449, 2010.

[9] B. Fan, L. Wang, F. Soong, L. Xie, "Photo-real talking head with deep bidirectional LSTM," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4884-4888, 2015.

[10] P. Hong, Z. Wen, T. Huang, "Real-time speech-driven face animation with expressions using neural networks," *IEEE Transactions on Neural Networks*, vol. 13, pp. 916-927, 2002.

[11] J. Jia, S. Zhang, F. Meng, Y. Wang, L. Cai, "Emotional audio-visual speech synthesis based on PAD," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 570-582, 2011.

[12] S. Zhang, Z. Wu, H. Meng, L. Cai, "Facial expression synthesis using PAD emotional parameters for a Chinese expressive avatar," *Affective Computing and Intelligent Interaction*, Springer, Berlin Heidelberg, pp. 24-35, 2007.

[13] 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*, vol. 73, pp. 439-461, 2014.

[14] C. Ding, L. Xie, P. Zhu, "Head motion synthesis from speech using deep neural networks," *Multimedia Tools and Applications*, pp. 1-18, 2014.

[15] S. Hochreiter, J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, pp. 1735–1780, 1997.

[16] F. Grézl, M. Karafiát, S. Kontár, J. Cernocky, "Probabilistic and bottleneck features for LVCSR of meetings," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, pp. 757-760, 2007.

[17] M. Wöllme, B. Schuller, G. Rigoll, "A novel bottleneck-BLSTM front-end for feature-level context modeling in conversational speech recognition," *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 36-41, 2011.

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

[19] R. Williams, D. Zipser, "Gradient-based learning algorithms for recurrent networks and their computational complexity," *Back-propagation: Theory, architectures and applications*, pp. 433–486, 1995.

[20] A. Graves, *Supervised Sequence Labelling with Recurrent Neural Networks*, Springer, Berlin, vol. 385, pp. 5-13, 2012.

[21] 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. International Conference on Spoken Language Processing (ICSLP)*, pp. 1802-1805, 2006.

[22] Schuller B, Steidl S, Batliner A, "The INTERSPEECH 2009 emotion challenge," in *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH),* pp. 312-315, 2009.