End-to-end Learning for 3D Facial Animation from Speech

Hai X. Pham
Rutgers University
hxp1@cs.rutgers.edu

Yuting Wang
Rutgers University
yw632@cs.rutgers.edu

Vladimir Pavlovic
Rutgers University
vladimir@cs.rutgers.edu

ABSTRACT

We present a deep learning framework for real-time speech-driven 3D facial animation from speech audio. Our deep neural network directly maps an input sequence of speech spectrograms to a series of micro facial action unit intensities to drive a 3D blendshape face model. In particular, our deep model is able to learn the latent representations of time-varying contextual information and affective states within the speech. Hence, our model not only activates appropriate facial action units at inference to depict different utterance generating actions, in the form of lip movements, but also, without any assumption, automatically estimates emotional intensity of the speaker and reproduces her ever-changing affective states by adjusting strength of related facial unit activations. For example, in a happy speech, the mouth opens wider than normal, while other facial units are relaxed; or both eyebrows raise higher in a surprised state. Experiments on diverse audiovisual corpora of different actors across a wide range of facial actions and emotional states show promising results of our approach. Being speaker-independent, our generalized model is readily applicable to various tasks in human-machine interaction and animation.

ACM Reference Format:
Hai X. Pham, Yuting Wang, and Vladimir Pavlovic. 2018. End-to-end Learning for 3D Facial Animation from Speech. In *2018 International Conference on Multimodal Interaction (ICMI '18), October 16–20, 2018, Boulder, CO, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3242969.3243017

1 INTRODUCTION

Face synthesis is essential to many applications, such as computer games, animated movies, teleconferencing, talking agents, among others. Traditional facial capture approaches have gained tremendous successes, reconstructing high level of realism. Yet, active face capture rigs utilizing motion sensors/markers are expensive and time-consuming to use. Alternatively, passive techniques capturing facial transformations from cameras, although less accurate, have achieved very impressive performance.

There lies one problem with vision-based facial capture approaches, however, where part of the face is occluded, e.g. when a person is wearing a mixed reality visor, or in the extreme situation where the entire visual appearance is non-existent. In such cases, other input modalities, such as audio, may be exploited to infer facial actions. Indeed, research on speech-driven face synthesis has regained attention of the community in recent time. Latest works [17, 22, 32, 33] employ deep neural networks in order to model the highly non-linear mapping from speech domain, either as audio or phonemes, to visual facial features. Particularly, in approaches by Karras et al. [17] and Pham et al. [22], the reconstruction of facial emotion is also taken into account to generate fully transformed 3D facial shapes. The method in [17] explicitly specifies the emotional state as an additional input beside waveforms, whereas [22] implicitly infers affective states from acoustic features, and represents emotions via blendshape weights.

In this work, we further improve the approach of [22] in several ways, in order to recreate a better 3D talking avatar that can naturally perform micro facial actions to represent the time-varying contextual information and emotional intensity from speech in real-time. Firstly, we forgo using handcrafted, high-level acoustic features which, as the authors of [22] conjectured, may cause the loss of important information to identify some specific emotions, e.g. happy. Instead, we directly use spectrogram as input to our neural network. Secondly, we employ convolutional neural networks (CNN) to learn meaningful acoustic feature representations, taking advantage of the locality and shift invariance in the time-frequency domain of audio signal. Lastly, we combine these convolutional layers with recurrent layer in an end-to-end network, which learns both temporal transition of facial movements, as well as spontaneous actions and varying emotional states from only speech sequences. Experiments on the RAVDESS [20] and VIDTIMIT [31] audiovisual corpora demonstrate promising results of our approach in real-time speech-driven 3D facial animation.

2 RELATED WORK

"Talking head", is a research topic where an avatar is animated to imitate human talking. Various approaches have been developed to synthesize a face model driven by either speech audio [13, 29, 39] or transcripts [9, 35]. Essentially, every talking head animation technique develops a mapping from an input speech to visual features. Early research on talking head used Hidden Markov Models with some successes [36, 37].

In recent years, deep neural networks (DNNs) have been successfully applied to speech synthesis [25, 40] and facial animation [11, 13, 33, 41] with superior performance. This is because DNNs are able to learn the correlation of high-dimensional data, as well as the highly non-linear mapping between input and output features. Suwajanakorn et al. [32] utilize long short-term memory recurrent neural network (LSTM-RNN) to predict 2D lip landmarks from MFCC features for lip-syncing. Karras et al. [17] propose a deep CNN that jointly takes audio autocorrelation coefficients and emotional state to generate an entire 3D face shape.

In terms of the underlying face model, these approaches can be categorized into image-based [5, 9, 12, 13, 36, 39] and model-based [3, 4, 7, 11, 30, 38] approaches. Image-based methods compose

ICMI '18, October 16–20, 2018, Boulder, CO, USA

ACM ISBN 978-1-4503-5692-3/18/10…$15.00
https://doi.org/10.1145/3242969.3243017

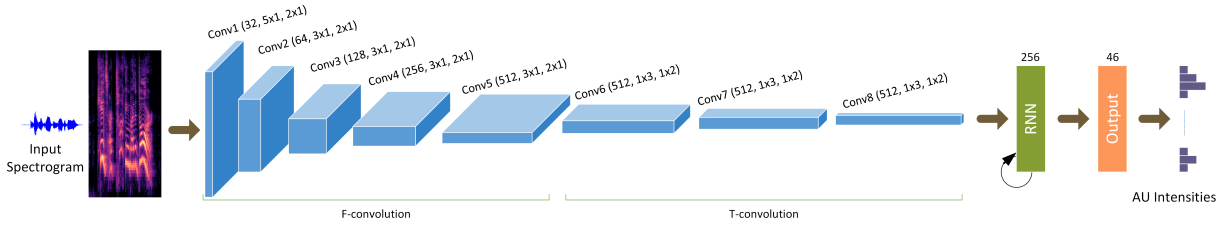

Figure 1: The proposed end-to-end speech-driven 3D facial animation framework. The input spectrogram is first convolved over frequency axis (F-convolution) with five conv. layers, followed by three conv. layers along the time axis (T-convolution).

photo-realistic avatar by concatenating samples from a database together. However, a tremendous amount of image samples is required to cover all possible facial appearances. In contrast, although lacking in realism, model-based approaches enjoy the flexibility of a deformable model, controlled by a set of parameters with more straightforward modeling. Pham et al. [22] proposed a mapping from acoustic features to blendshape weights [6]. This face model allows emotional representation that can be inferred from speech without explicitly defining the emotion from input, and it is also made use of in our work.

CNN-based speech modeling. Convolutional neural networks [19] have achieved great successes in many vision tasks. In recent years, CNNs have been also employed in speech recognition tasks that directly model the raw waveforms [1, 2, 10, 16, 21, 26–28, 34]. In this work, we apply convolutions in the time-frequency domain to learn meaningful features representing both context of speech and facial emotions.

3 PROPOSED METHOD

3.1 Face Representation

Our work makes use of the 3D blendshape face model from the FaceWarehouse database [6], which has been utilized successfully in visual 3D face tracking tasks [23, 24]. An arbitrarily transformed facial shape S can be composed as:

$$S = B_0 + \sum_{i=1}^{N} (B_i - B_0)e_i, \tag{1}$$

where $\{B_i|i = 1..N\}$ are personalized expression blendshape bases of a particular person, B_0 is the neutral posed blendshape and $N = 46$. e is a vector of expression blending parameters and $\{e_i\}$ are constrained within $[0, 1]$.

Slightly different from [22], our deep model only generates e, as it is difficult to infer head pose from speech. Furthermore learning to estimate head pose may disrupt learning the correct acoustic features representing facial actions. We use the 3D face tracker in [24] to extract these parameters from training videos.

3.2 Model Architecture

Our end-to-end deep neural net is illustrated in Fig. 1. The input to our model is raw time-frequency spectrogram of audio signal. Specifically, each spectrogram contains 128 frequency bins across 32 time frames, constructed as a 2D (frequency-time) array suitable for CNN. We apply convolutions on frequency and time separately, similar to [17, 28], as this practice has been empirically shown to reduce overfitting, furthermore, using smaller filters requires less computation, which consequently speeds up training and inference. In particular, the input spectrogram is first convolved on the frequency axis with down-sampling factor of two. Then, two-strided convolution is applied on the time axis. In this work, we report model performance where the recurrent layer is formulated as either LSTM [15] or gated recurrent unit (GRU) [8] cells.

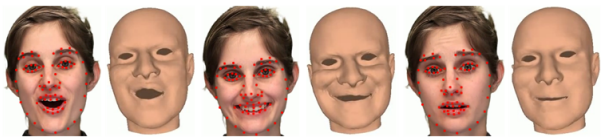

Figure 2: A few samples from the RAVDESS database, where a 3D facial blendshape (right) is aligned to the face of the actor (left) in the corresponding frame. Red dots indicate 3D landmarks of the model projected to the image plane. Rotation parameters are not used in training and inference.

3.3 Model Training

Our framework maps input sequence of spectrograms $x^t, t = 1..T$ to output sequence of shape parameter vectors e^t, where T is the number of video frames. We train the model by minimizing the following objective function:

$$\min \frac{1}{T}\frac{1}{N}\sum_{t=1}^{T}\left(\sum_{i=1}^{N}\left(\frac{e_i^t - \hat{e}_i^t}{\sigma_i}\right)^2 + \lambda\left\|e^t\right\|_1\right), \tag{2}$$

where $\hat{e}^t$ is the expected output, which we extract from training videos, σ_i is the standard deviation of the i^{th} AU component extracted from training data, and λ is the trade-off weight. Essentially, the first term is the $\mathcal{L}2$ loss normalized by variance to reduce the bias towards mouth-related action units, since all videos depict actors speaking but they do not always show a particular expression. The second term encourages parameter sparsity, as not all action units are activated simultaneously. We empirically choose $\lambda = 0.1$, and model parameters are learned by the ADAM optimizer [18]. More details about model architecture and training procedure can be found in the supplementary material [1].

3.4 Audio Processing

For each video frame t in the corpus, we extract a 96ms audio frame sampled at 44.1kHz, including data of the current video frame and previous frames. With the intended application in real-time animation, we do not consider any delay to gather future data, as

[1] code & aux. materials available at: www.cs.rutgers.edu/~hxp1/speechproject.html

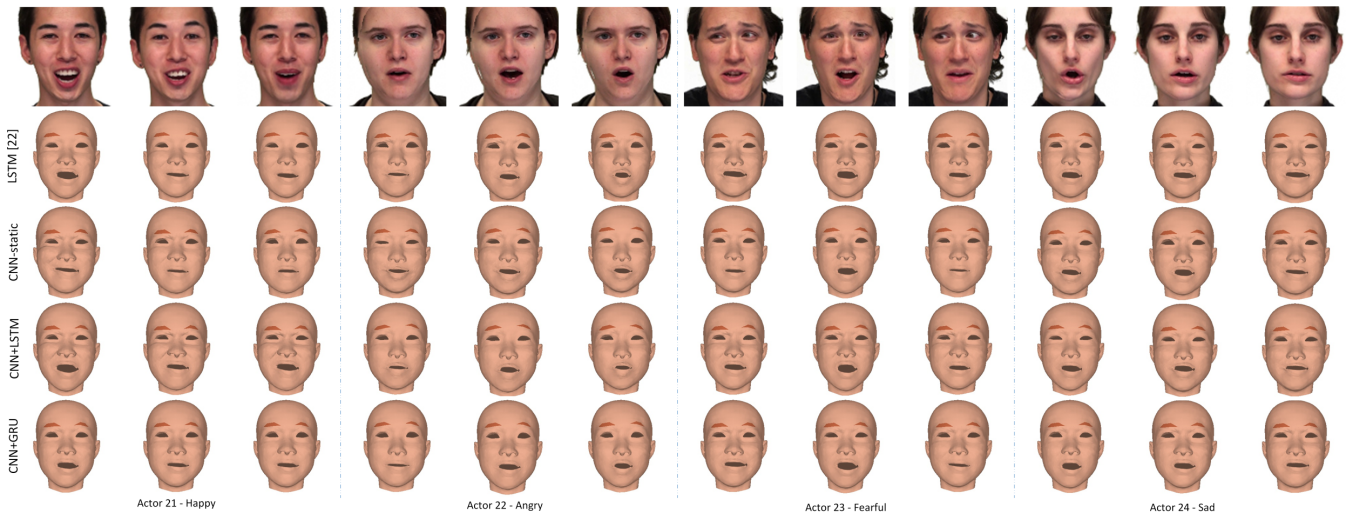

Figure 3: Reconstruction results from the RAVDESS corpus. We sample four sequences of four actors, three frames per sequence and animate a generic 3D face model with parameters generated by each model. From top to bottom: facial texture patch, results of [22], *CNN-static*, *CNN+LSTM* and *CNN+GRU*, respectively. In these samples, *CNN+GRU* captures the temporal dynamics of facial deformation most accurately. (The facial texture (top row) was frontalized for visualization purpose).

they are unknown in a live streaming scenario. Instead, temporal transition will be modeled by the recurrent layer. We apply FFT with window size of 256 and hop length of 128, to recover a power spectrogram of 128 frequency bins across 32 time frames.

4 EXPERIMENTS

4.1 Datasets

We use RAVDESS [20], VIDTIMIT [31] and SAVEE [14] audiovisual corpora for training and evaluation. Specifically, the training set consists of data from the first 20 actors in the RAVDESS dataset, first 40 actors in VIDTIMIT and all four actors in SAVEE, respectively. The test set includes four remaining actors in RAVDESS and three actors in VIDTIMIT, who do not appear in the training set.

4.2 Experimental Settings

Our proposed deep neural network is trained in two configurations: *CNN+LSTM* and *CNN+GRU*, in which the recurrent layer uses LSTM and GRU cells, respectively. As a baseline, we replace the recurrent layer in our proposed model with a fully connected layer of 1,024 units, and denote it as *CNN-static*. This static model cannot handle smooth temporal transition, it estimates facial parameters in a frame-by-frame basis. We compare our proposed models with the method in [22], which uses engineered input features.

We measure performance of these models on *four* metrics: in addition to RMSE of 3D landmarks and MSE of facial action parameters with respect to ground truths recovered by the visual tracker [24], we also report *temporal smoothness*: $\frac{1}{N}\left\|\left(e^{t+1}-e^{t}\right)-\left(\hat{e}^{t+1}-\hat{e}^{t}\right)\right\|_2^2$, and *self-temporal smoothness*: $\frac{1}{N}\left\|e^{t}-\left(e^{t-1}+e^{t+1}\right)/2\right\|_2^2$, which shows how smooth frame transition is within the output sequence itself. Landmark errors are calculated as real-world distances in millimeter from inner landmarks (shown in Fig. 2) on the reconstructed 3D face shape, to those of the ground truth 3D shape (of a generic identity, to avoid inaccuracy in identity recovery). Nevertheless, these error metrics do not truly reflect performance of our deep models on 3D face reconstruction quality, because facial expression may not always relate to speech. For example, the speaker may open her mouth or raise eyebrows but does not utter a sound.

Table 1: Error metrics of four models on the test set. Best results are marked in bold.

	LSTM [22]	CNN-static	CNN+LSTM	CNN+GRU
MSE on AU coefficients (× 1e-2)				
RAVDESS	7.179	**6.53**	7.301	6.587
VIDTIMIT	8.271	7.646	7.457	**7.038**
RMSE (unit: mm) on 3D landmarks				
RAVDESS	1.067	**1.038**	1.048	1.044
VIDTIMIT	0.974	0.993	0.968	**0.966**
Temporal smoothness (× 1e-2)				
RAVDESS	**0.284**	1.94	0.291	0.292
VIDTIMIT	0.575	2.888	0.581	**0.573**
Self-smoothness (× 1e-2)				
RAVDESS	0.964	2.654	0.986	**0.908**
VIDTIMIT	1.22	3.678	1.134	**1.088**

4.3 Evaluation

Table 1 shows the aforementioned error metrics of four models on the test set. Based on parameter MSE and landmark RMSE, *CNN-static* performs slightly better than *CNN+GRU* on the RAVDESS dataset (< 1e-3 error difference), whereas *CNN+GRU* is 8% better than *CNN-static* on VIDTIMIT. This can be explained by the inherent difference between two datasets.

In VIDTIMIT, each actor maintains almost uniform expression throughout the sequence, mostly only the mouth region is deformed, i.e. speaking. The facial deformation dynamics is smooth and stable,

Table 2: MSE (× 1e-2) of expression blending weights, organized by categories. Best results are marked in bold.

	Neutral	Calm	Happy	Sad	Angry	Fearful	Disgusted	Surprised	Actor21	Actor22	Actor23	Actor24
LSTM [22]	7.001	7.059	7.378	7.339	6.738	6.928	7.965	7.583	5.049	6.93	7.105	9.576
CNN-static	**6.505**	**6.356**	**7.026**	6.607	6.169	6.281	6.734	**6.87**	4.457	6.191	**6.3**	**9.109**
CNN+LSTM	7.111	7.236	7.861	7.294	6.617	7.222	7.627	7.869	4.684	7.37	6.996	10.081
CNN+GRU	6.911	6.681	7.031	**6.591**	**6.061**	**6.246**	**6.627**	6.943	**4.248**	**6.058**	6.766	9.225

Table 3: RMSE of 3D landmarks in millimeter on RAVDESS, organized by categories. Best results are marked in bold.

	Neutral	Calm	Happy	Sad	Angry	Fearful	Disgusted	Surprised	Actor21	Actor22	Actor23	Actor24
LSTM [22]	1.072	1.052	1.111	1.047	1.073	1.042	1.053	1.100	1.055	0.997	1.055	1.153
CNN-static	1.048	**1.022**	1.092	**1.014**	**1.046**	**1.015**	**1.028**	1.046	1.029	0.991	**0.990**	**1.134**
CNN+LSTM	**1.032**	1.035	1.090	1.018	1.051	1.041	1.066	1.064	**1.003**	1.034	1.020	1.128
CNN+GRU	1.051	1.038	**1.088**	1.019	1.056	1.018	1.037	**1.041**	1.027	**0.964**	1.039	1.135

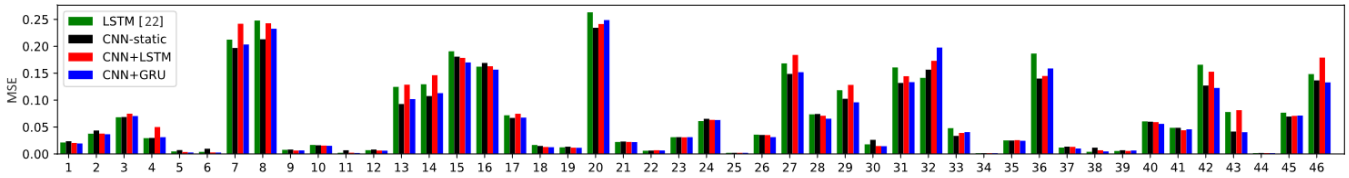

Figure 4: Plotting MSE of individual action unit parameters. In general *CNN+GRU* and *CNN-static* have similar performance across different AUs. Our proposed models outperform [22] in most cases, especially on major AUs: 9, 10, 13-18, 22, 36 and 37.

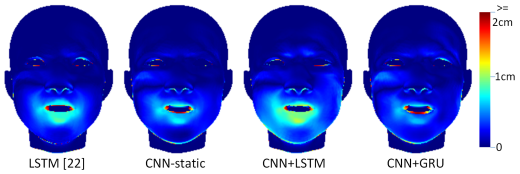

Figure 5: Surface error on a sample of *Actor 21 - Happy*.

hence, recurrent models are able to estimate temporal changes of facial action intensities effectively. Notice that in this case, LSTM [22] also performs better than *CNN-static*, in terms of landmark RMSE. This is understandable, since engineered features (MFCCs) can represent the context of speech sufficiently.

On the other hand, RAVDESS actors manifest spontaneous and varying facial expressions while speaking. Thus, the static model has a slight edge in estimating those sudden expression changes. This also explains why the performance of *CNN+LSTM* is worse than two other end-to-end models. LSTM model is more complex than GRU, and has the tendency to smooth the output sequence more. Thus, on VIDTIMIT test data, *CNN+LSTM* outperforms the static baseline, but it is bested by the GRU model. It performs worse than both *CNN+GRU* and *CNN-static* on RAVDESS. Some resulting 3D face syntheses on RAVDESS are demonstrated in Fig. 3.

To further understand how each model performs on the RAVDESS test set, we break down parameter MSE and landmark RMSE corresponding to different emotions and actors, as shown in Table 2 and 3. Generally, *CNN+GRU* and *CNN-static* have comparable performance in terms of reported error metrics. Overall, our proposed models consistently perform better than [22] which uses engineered features. However, *CNN-static* does not handle temporal smoothness, thus sequential transitions in its generated animation are often jarring, unlike recurrent models. Indeed, this is demonstrated through smoothness metrics in Table 1. Both recurrent models generate satiny animated sequences, which are reflected in their self-smoothness metrics. Temporal smoothness measures of *CNN+GRU*, *CNN+LSTM* and [22] are roughly similar, and are an order of magnitude better than that of *CNN-static*. These results show that our proposed *CNN+GRU* model achieves both accuracy and temporal smoothness of predicted facial action sequences.

Fig. 4 provides further insight on how each model estimates the most prominent facial action parameters from speech. *CNN+GRU* and *CNN-static* demonstrate superior performance across different action units, whereas *CNN+LSTM* has significantly higher errors on AU13 and AU14, which are raising eyebrow actions. Nonetheless, these error metrics are calculated from compressed information (landmarks and parameters), they do not fully indicate the quality of shape reconstruction. Fig. 5 shows errors when comparing the reconstructed surfaces by four models to the shape estimated by the visual tracker. It is observed that *CNN+GRU* often has the smallest surface errors overall, especially in the mouth region.

Speed. On a Quadro K1000M GPU, recurrent models process one frame in 5.2ms, while *CNN-static* is slightly faster at 5ms.

5 CONCLUSION

This paper introduces a deep learning framework for real-time speech-driven 3D facial animation from sequence of input spectrograms. Our proposed deep neural network learns a mapping from audio signal to the temporally varying context of the speech, as well as emotional states of the speaker represented implicitly via blending weights of a 3D face model. Experiments demonstrate that our approach could estimate lip movements with emotional intensity of the speaker reasonably from just her speech. In future work, we will improve the generalization of our deep neural network, and explore other generative models to increase the quality and realism of facial reconstruction.

REFERENCES

[1] Osama Abdel-Hamid, Abdel-Rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu. 2014. Convolutional Neural Networks for Speech Recognition. *IEEE Transaction on Audio, Speech, and Language Processing* 22, 10 (October 2014).

[2] Osama Abdel-Hamid, Abdel-Rahman Mohamed, Hui Jiang, and Gerald Penn. 2012. Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*.

[3] V. Blanz, C. Basso, T. Poggio, and T. Vetter. 1999. Reanimating faces in images and video. In *SIGGAPH*. 187–194.

[4] V. Blanz and T. Vetter. 2003. A morphable model for the synthesis of 3d faces. In *Eurographics*. 641–650.

[5] C. Bregler, M. Covell, and M. Slaney. 2007. Video rewrite: driving visual speech with audio. In *SIGGRAPH*. 353–360.

[6] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. 2014. FaceWarehouse: A 3D Facial Expression Database for Visual Computing. *IEEE Transactions on Visualization and Computer Graphics* 20, 3 (March 2014), 413–425.

[7] Yong Cao, Wen C. Tien, Petros Faloutsos, and Fred Pighin. 2005. Expressive Speech-Driven Facial Animation. *ACM Transactions on Graphics* 24, 4 (2005), 1283–1302.

[8] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. In *NIPS 2014 Deep Learning and Representation Learning Workshop*.

[9] E. Cosatto, J. Ostermann, H. P. Graf, and J. Schroeter. 2003. Lifelike talking faces for interactive services. *Proc IEEE* 91, 9 (2003), 1406–1429.

[10] Li Deng, Ossama Abdel-Hamid, and Dong Yu. 2013. A deep convolutional neural network using heterogeneous pooling for trading acoustic invariance with phonetic confusion. In *IEEE International Conference on Acoustics, Speech and Signal Processing*.

[11] Chuang Ding, Lei Xie, and Pengcheng Zhu. 2015. Head Motion Synthesis from Speech using Deep Neural Network. *Multimed Tools Appl* 74 (2015), 9871–9888.

[12] T. Ezzat, G. Geiger, and T. Poggio. 2002. Trainable videorealistic speech animatio. In *SIGGRAPH*. 388–397.

[13] Bo Fan, Lei Xie, Shan Yang, Lijuan Wang, and Frank K. Soong. 2016. A Deep Bidirectional LSTM Approach for Video-realistic Talking Head. *Multimed Tools Appl* 75 (2016), 5287–5309.

[14] S. Haq, P.J.B. Jackson, and J. Edge. 2008. Audio-visual feature selection and reduction for emotion classification. In *Proc. Int. Conf. on Auditory-Visual Speech Processing (AVSP'08), Tangalooma, Australia*.

[15] S. Hochreiter and J. Scmidhuber. 1997. Long short-term memory. *Neural Comput* 9, 8 (1997), 1735–1780.

[16] Yedid Hoshen, Ron J. Weiss, and Kevin W. Wilson. 2015. Speech acoustic modeling from raw multichannel waveforms. In *IEEE International Conference on Acoustics, Speech and Signal Processing*.

[17] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaako Lehtinen. 2017. Audio-driven facial animation by joint end-to-end learning of pose and emotion. In *SIGGRAPH*.

[18] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference for Learning Representations*.

[19] Yann LeCun and Yoshua Bengio. 1998. Convolutional Networks for Images, Speech, and Time-Series. (1998), 255–258.

[20] S. R. Livingstone, K. Peck, and F. A. Russo. 2012. RAVDESS: The Ryerson Audio-Visual Database of Emotional Speech and Song. In *22nd Annual Meeting of the Canadian Society for Brain, Behaviour and Cognitive Science (CSBBCS)*.

[21] Dimitri Palaz, Ronan Collobert, and Mathew Magimai-Doss. 2013. Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks. In *Interspeech*.

[22] Hai X. Pham, Samuel Cheung, and Vladimir Pavlovic. 2017. Speech-driven 3D facial animation with implicit emotional awareness: a deep learning approach. In *The 1st DALCOM workshop, CVPR*.

[23] Hai X. Pham and Vladimir Pavlovic. 2016. Robust Real-time 3D Face Tracking from RGBD Videos under Extreme Pose, Depth, and Expression Variations. In *3DV*.

[24] Hai X. Pham, Vladimir Pavlovic, Jianfei Cai, and Tat jen Cham. 2016. Robust Real-time Performance-driven 3D Face Tracking. In *ICPR*.

[25] Y. Qian, Y. Fan, and F. K. Soong. 2014. On the training aspects of deep neural network (DNN) for parametric TTS synthesis. In *ICASSP*. 3829–3833.

[26] Tara N. Sainath, Brian Kingsbury, George Saon, Hagen Soltau, Abdel rahman Mohamed, George Dahl, and Bhuvana Ramabhadran. 2015. Deep convolutional neural networks for large-scale speech tasks. *Neural Network* 64 (2015), 39–48.

[27] Tara N. Sainath, Oriol Vinyals, Andrew Senior, and Hasim Sak. 2015. Convolutional, long short-term memory, fully connected deep neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing*.

[28] Tara N. Sainath, Ron J. Weiss, Andrew Senior, Kevin W. Wilson, and Oriol Vinyals. 2015. Learning the speech front-end with raw waveforms CLDNNs. In *Interspeech*.

[29] S. Sako, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. 2000. HMM-based text-to-audio-visual speech synthesis. In *ICSLP*. 25–28.

[30] G. Salvi, J. Beskow, S.A. Moubayed, and B. Granstrom. 2009. Synface: speech-driven facial animation for virtual speech-reading support. *URASIP journal on Audio, speech, and music processing* (2009).

[31] C. Sanderson and B.C. Lovell. 2009. Multi-Region Probabilistic Histograms for Robust and Scalable Identity Inference. *Lecture Notes in Computer Science (LNCS)* 5558 (2009), 199–208.

[32] Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Schlizerman. 2017. Synthesizing Obama: learning lip sync from audio. In *SIGGRAPH*.

[33] Sarah Taylor, Taehwan Kim, Yisong Yue, Moshe Mahler, James Krahe, Anastasio Garcia Rodriguez, Jessica Hodgins, and Iain Matthews. 2017. A deep learning approach for generalized speech animation. In *SIGGRAPH*.

[34] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A. Nicolaou, Björn Schuller, and Stefanos Zafeiriou. 2016. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In *Interspeech*.

[35] Alice Wang, Michael Emmi, and Petros Faloutsos. 2007. Assembling an Expressive Facial Animation System. *ACM SIGGRAPH Video Game Symposium (Sandbox)* (2007), 21–26.

[36] L. Wang, X. Qian, W. Han, and F. K. Soong. 2010. Synthesizing photo-real talking head via trajectoryguided sample selection. In *Interspeech*. 446–449.

[37] L. Wang, X. Qian, F. K. Soong, and Q. Huo. 2011. Text driven 3D Photo-realistic talking head. In *Interspeech*. 3307–3310.

[38] Z. Wu, S. Zhang, L. Cai, and H. Meng. 2006. Real-time synthesis of chinese visual speech and facial expressions using mpeg-4 fap features in a three-dimensional avatar. In *Interspeech*. 1802–1805.

[39] L. Xie and Z. Liu. 2007. Realistic mouth-synching for speech-driven talking face using articulatory modeling. *IEEE Trans Multimed* 9, 23 (2007), 500–510.

[40] H. Zen, A. Senior, and M. Schuster. 2013. Statistical parametric speech synthesis using deep neural networks. In *ICASSP*. 7962–7966.

[41] X. Zhang, L. Wang, G. Li, F. Seide, and F. K. Soong. 2013. A new language independent, photo realistic talking head driven by voice only. In *Interspeech*.