A Real-time Speech Driven Talking Avatar based on Deep Neural Network

Kai Zhao[*†], Zhiyong Wu [*†] and Lianhong Cai [*†]
[*]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, China
[†]Tsinghua National Laboratory for Information Science and Technology,
Department of Computer Science and Technology, Tsinghua University, Beijing, China
E-mail: zk69052@163.com, zywu@sz.tsinghua.edu.cn, clh-dcs@tsinghua.edu.cn Tel: +86-755-26036870

***Abstract*— This paper describes our initial work in developing a real-time speech driven talking avatar system with deep neural network. The input of the system is the acoustic speech and the output is the articulatory movements (that are synchronized with the input speech) on a 3-dimentional avatar. The mapping from the input acoustic features to the output articulatory features is achieved by virtue of deep neural network (DNN). Experiments on the well known acoustic-articulatory English speech corpus MNGU0 demonstrate that the proposed audio-visual mapping method based on DNN can achieve good performance.**

I. INTRODUCTION

This paper describes our attempt towards the development of a real time speech driven talking avatar system based on deep neural network. The input of the system is the acoustic speech and the output is the articulatory movement animation on a virtual talking avatar. The generated movements of the articulators (e.g. lip, tongue, velum, etc.) are synchronized with the input speech. Such system can offer multimedia and multimodal presentation for entertainment applications, for remote telecommunications, and as an aid for the hearing-impaired where the simulated lip movements can help the user decipher the acoustic speech [1].

In speech production, there are direct connections between the configurations of articulators, which are the positions and movements of the lips, tongue, velum, etc., and the speech. In speech driven talking avatar, the most important step is audio-visual mapping, which aims to learn a function whose input is the features representing the acoustic speech and output is the articulatory features.

That the audio-visual mapping between the acoustic and the articulatory features is a non-linear mapping and not a one-to-one mapping makes itself a challenging problem. In the past decades, several techniques have been applied to tackle the audio-visual mapping problem including the artificial neural network [2], hidden Markov model [3][4]. Gaussian mixture model was used to model the joint distribution of acoustic and articulatory features based on a parallel acoustic-articulatory speech corpus [5]. A dynamic Bayesian network based audio-visual articulatory model was proposed in [6] to model the correlation between the audio and video features, and Baum-Welch inversion algorithm was presented to generate optimal facial parameters from audio with the proposed model.

Although the dynamic Bayesian network with Baum-Welch inversion algorithm can achieve realistic mouth-synching, the recursive steps for computing the optimal articulatory features has prevent the method from being used in the real-time applications. The performance of audio-visual mapping may degrade a lot if the artificial neural network (ANN) is over-trained. Deep neural network (DNN) [7] has shown many superior characteristics over traditional ANN [8]. First, the unsupervised pre-training step of the DNN can make effective use of large amount of unlabeled training data. Second, the over-fitting problem of the traditional ANN can be effectively addressed by the pre-training step.

This paper describes the development of a real-time speech driven talking avatar system, in which the acoustic to visual articulatory feature mapping is achieved by virtue of the deep neural network (DNN). The rest of the paper is organized as follows. Section 2 illustrates the detailed architecture of the proposed system. Section 3 describes the deep belief network (DBN) and the DNN for audio-visual mapping. Experiments and results are then presented in Section 4. Finally, Section 5 lays out the conclusions.

II. ARCHITECTURE OF SPEECH DRIVEN TALKING AVATAR

The architecture of the proposed real-time speech driven talking avatar system including audio-visual mapping with deep neural network is illustrated in Fig. 1.

During training stage, speech waveforms from the training audio-visual bimodal speech corpus are fed to the acoustic feature extraction module to extract the acoustic features for training the deep neural network. Articulatory features are also extracted from the training bimodal corpus. Pre-training technology is then utilized to provide a good initialization point for the parameters of the deep neural network (DNN) by training it as stacked restricted Boltzmann machines (RBMs). After training the stacked RBMs, all the units and weights are treated like a traditional neural network to perform fine-tuning

This work is partially supported by the National Natural Science Foundation of China (60928005, 60805008, 61375027, 61370023), the Upgrading Plan Project of Shenzhen Key Laboratory (CXB201005250038A) and the Science and Technology R&D Funding of the Shenzhen Municipal.

to get better regression performance.

During prediction stage, the acoustic features of the input speech waveform are also extracted by the acoustic feature extraction module. These acoustic features are served as the input of the DNN for audio-visual mapping. The articulatory features output from the DNN are sent to a 3D talking avatar rendering module to generate the talking avatar animation, which is finally playback together with the input speech.

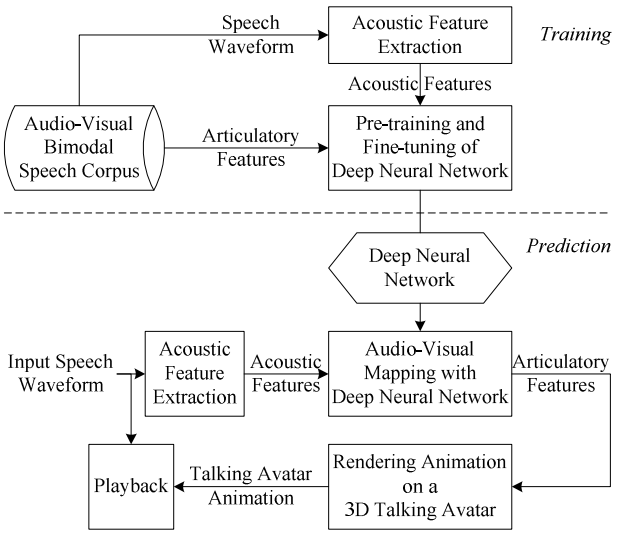

Fig. 1 Architecture of the proposed speech driven talking avatar system based on deep neural network.

III. AUDIO VISUAL MAPPING WITH DEEP NEURAL NETWORK

A. Deep Belief Network (DBN)

Training a deep neural network directly has long been known as a difficult problem due to its highly non-convex property, problem of gradient diffusion and pathological curvature when training with the first order optimizer. The invention of deep belief network (DBN) was proposed to be the first solution to this problem [7][9].

1) Restricted Boltzmann Machine (RBM)

A deep belief network is trained as a stack of restricted Boltzmann machine (RBM) [10]. An RBM is a probabilistic model represented by an undirected graphical model with two layers of probabilistic units: a visible variable layer **v** which represents the data been modeled and a hidden/latent variable layer **h**. All units in one layer are fully connected to the units in the other layer and no connections exist between the units of the same layer.

Each (**v**, **h**) pair is assigned with a energy function $E(\mathbf{v}, \mathbf{h})$, which captures the dependency probability between variables **v** and **h**. The lower the energy value is, the more probable configuration of **v** and **h** could be. The joint probability distribution of **v** and **h** is modeled as:

$$p(\mathbf{v},\mathbf{h}) = \frac{1}{Z} e^{-E(\mathbf{v},\mathbf{h})} \tag{1}$$

where $Z = \sum_{\mathbf{vh}} e^{-E(\mathbf{v},\mathbf{h})}$ is the normalization factor.

If both **v** and **h** are multidimensional binary variables, a Bernoulli-Bernoulli RBM is used. In this case, the energy function is typically defined as:

$$E(\mathbf{v},\mathbf{h}) = -\mathbf{a}^T\mathbf{v} - \mathbf{b}^T\mathbf{h} - \mathbf{v}^T\mathbf{W}\mathbf{h} \tag{2}$$

where **W** is the matrix of connection weights between **v** and **h**, **a** is the bias vector of visible layer and **b** is the bias vector of hidden layer.

For the problems with real-valued input variable **v**, a Gaussian-Bernoulli RBM can be used. The hidden variable **h** is still multidimensional binary variable. The energy function is typically defined as:

$$E(\mathbf{v},\mathbf{h}) = -\frac{1}{2}(\mathbf{a}-\mathbf{v})^T(\mathbf{a}-\mathbf{v}) - \mathbf{b}^T\mathbf{h} - \mathbf{v}^T\mathbf{W}\mathbf{h} \tag{3}$$

The input **v** to Gaussian-Bernoulli RBM is usually normalized over the training data to have mean 0 and standard deviation 1 for each unit.

2) Stacked RBM

Suppose we train a series of RBMs, we can stack them together one by one with the visible variable layer of the later RBM being the hidden variable layer of the former RBM. In this way, we can get a multi-layer generative probabilistic model with one visible layer and many hidden layers. The model is called deep belief network (DBN). Such greedy training fashion [9] has been proved to guarantee the improvement of the variational lower bound of the probability of visible variable, if one more RBM is trained and stacked.

B. Deep Neural Network (DNN)

A deep neural network (DNN) is a feed forward neural network (also called multi-layer perceptron, MLP) with many hidden layers (usually larger than 2 and smaller than 10). For the convenience of discussion in the following, we would like to define some notations first.

For a DNN with K layers (input visible layer excluded), the weight matrix and hidden bias from bottom up is denoted as $\mathbf{W}_k$ and $\mathbf{b}_k$ (k=1,2,…,K). $h_{ki}(\mathbf{x})$ is the output of the i-th neuron in hidden layer k. $\mathbf{h}_k(\mathbf{x})=[h_{ki}(\mathbf{x}), i=1,2,\ldots,I]^T$, I is the number of neurons (hidden units) in hidden layer k. Let **x** be the input and **y** be the desired output. Then we have:

$$\mathbf{h}_k(\mathbf{x}) = sigmoid\left(\mathbf{u}_k(\mathbf{x})\right), k = 1, 2, ..., K \tag{4}$$

where

$$\mathbf{u}_k(\mathbf{x}) = \mathbf{W}_k\mathbf{h}_{k-1}(\mathbf{x}) + \mathbf{b}_k \tag{5}$$

$$sigmoid(x) = \frac{1}{1+e^{-x}} \tag{6}$$

Here, $\mathbf{h}_0(\mathbf{x})=\mathbf{x}$ and $\mathbf{u}_K(\mathbf{x})=\mathbf{y}$. With the DNN, suppose we can get an output $\tilde{\mathbf{y}}$ from the input **x**, then the learning of DNN is done by optimizing the following loss function:

$$L(\mathbf{y},\tilde{\mathbf{y}}) = \left\|\mathbf{y}-\tilde{\mathbf{y}}\right\|_2^2 \tag{7}$$

1) Pre-training

The similar structure between DBN and feed forward DNN has made it natural to utilize the weights learnt in a DBN to provide new initializing weights other than the random small Gaussian weights traditionally used to train a neural network. This procedure is called pre-training. Empirically, a neural network with pre-training enjoys faster convergence rate and better convergence results. More importantly, it makes the training of a DNN much easier.

2) Fine-tuning with back-propagation

After pre-training, a DNN can be trained with back-propagation algorithm just as a MLP. This fine-tuning procedure is described as below:

$$\frac{\partial L(\mathbf{y},\tilde{\mathbf{y}})}{\partial \mathbf{u}_K(\mathbf{x})} = -2(\mathbf{y}-\tilde{\mathbf{y}}) \tag{8}$$

and for k=K-1,…,1,

$$\frac{\partial L(\mathbf{y},\tilde{\mathbf{y}})}{\partial u_{ki}(\mathbf{x})} = \frac{\partial L}{\partial h_{ki}(\mathbf{x})} h_{ki}(\mathbf{x})\left(1-h_{ki}(\mathbf{x})\right) \tag{9}$$

$$\frac{\partial L(\mathbf{y},\tilde{\mathbf{y}})}{\partial \mathbf{u}_k(\mathbf{x})} = \mathbf{W}_{k+1}^T \frac{\partial L}{\partial \mathbf{u}_{k+1}(\mathbf{x})} \tag{10}$$

C. DNN for Audio-Visual Mapping

To perform audio-visual mapping with DNN, the input of the DNN is the real-valued acoustic features and the output is the values of the articulator positions. A Gaussian-Bernoulli RBM is used for the bottom two layers of the DNN, and each dimension of the acoustic feature input is normalized over the training set to have mean 0 and variance1. A linear regression layer is added on top of the DNN with one output unit for each articulator position to infer.

IV. EXPERIMENTS

A. Database

The electromagnetic articulography (EMA) dataset of the MNGU0 database [11] was used for the audio-visual mapping experiments.

In collecting the database, 6 coils were attached to the speaker's articulators in the midsagittal plane: 3 on the tongue, one on the lower incisor and one each on the upper and lower lips. The x- and y-coordinates of the 6 coils in the midsagittal plane were recorded and used. So, the articulatory data used for this experiment included 12 channels of EMA data at sampling frequency of 200Hz. The related audio data was converted to frequency-warped line spectral frequencies (LSFs) of order 40 plus a gain value. The LSFs were derived from the spectral envelope estimated with STRAIGHT [12], with 5msec frame shift to match the EMA sampling rate. The initial and final silences were removed. Both EMA and LSF feature vectors were z-score normalized by subtracting their respective global mean and dividing by 4 times the standard deviation for each dimension.

The database contains 1,354 utterances and is divided into three subsets: a validation and test set each with 63 utterances, and a training set containing the rest 1,228 utterances [11].

B. Experimental Setup

In the experiments, the audio-visual mapping performances of traditional artificial neural network (ANN) and deep neural network (DNN) were compared. The network structures of ANN and DNN were the same. The difference lay in that the initial weights of the DNN were initialized by pre-training, while the initial weights of the ANN were initialized with the random small Gaussian weights.

To measure the performance of audio-visual mapping, root mean-squared error (RMSE) was used as the measurement, which is defined as

$$RMSE = \sqrt{\frac{1}{N}\sum_i \left(e_i - t_i\right)^2} \tag{11}$$

where e_i is the estimated articulatory trajectory value and t_i is the actual measured articulatory value.

1) Network structure

Four hidden layers were used in the experiments for both DNN and ANN, with 1000 hidden units for each hidden layer. For the input, a context window of 11 acoustic frames (5 left frames, 1 current frame, and 5 right frames) was used, with 40 order LSFs for each frame. Hence, the input layer contained 40×11=440 units. As for the output, the 12 order EMA frame corresponding to the time of the current frame was used as the articulatory data related to the input 11 acoustic frames. The delta and acceleration (delta delta) of the EMA data were also considered. So the output layer contained 12×3=36 units.

2) Pre-training

All the RBMs were trained with contrastive divergence algorithm [9]. Stochastic gradient descent was used with a mini-batch size of 128. A momentum of 0.9 and no weight decay was applied. For Gaussian-Bernoulli RBM, 100 epochs and learning rate of 0.001 was used. For all Bernoulli-Bernoulli RBMs, 50 epochs and learning rate of 0.01 was used.

3) Back-propagation

The DNN was trained with back-propagation and stochastic gradient decent algorithm with a mini-batch of 128. The learning rate starts from 0.001 and exponentially decays by factor of 0.998. The training was stopped if no further evident decrease of RMSE was observed.

C. Experimental Results

Two experiments were conducted. First, the performances of the traditional ANN and DNN for audio-visual mapping were compared in terms of RMSE on the validation subset of the MNGU0 database. Results are shown in Fig. 2, where the x-axis indicates the number of epochs and the y-axis shows the RMSE averaged over all 12 articulatory features. As can be seen, due to the procedure of pre-training, the DNN has

higher performance (lower average RMSE) than ANN for audio-visual mapping.

Fig. 3 shows the estimated articulatory trajectory values (green) for 1000 frames of the tongue body feature (measured by T2 coil in MNGU0 database) and compares the estimated values with the actual measured values (red). As we can see, the estimated values follow the same trend and are very close to the measured actual values, which indicate the feasibility of DNN for audio-visual mapping.

However, we can also find some big differences between estimated and measured values still exist, which might be caused by the fact the mapping between acoustic and visual features is not a one-to-one mapping. Furthermore, as can be seen from the figure, the estimated values have shown some dynamic variances, while the actual measured curve is much smoother. This may be due to that the DNN only performs regression from a context window of acoustic features to one frame of articulatory positions, and the continuity properties of the articulatory trajectories are not considered.

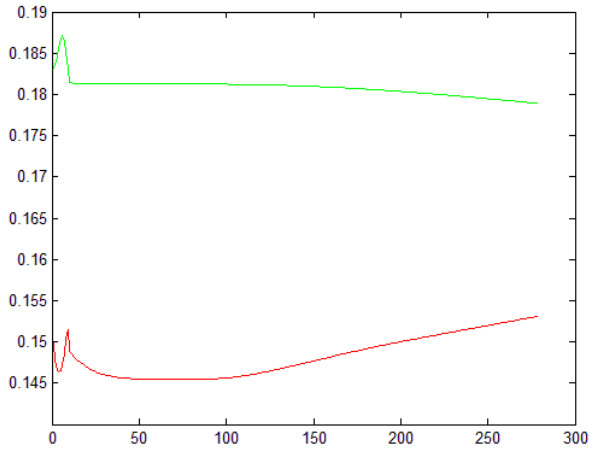

Fig. 2 Average generation error (RMSE) of articulatory features for DNN (red) and ANN (green) on the validation set of MNGU0 database, where DNN was pre-trained as DBN while ANN was initialized randomly.

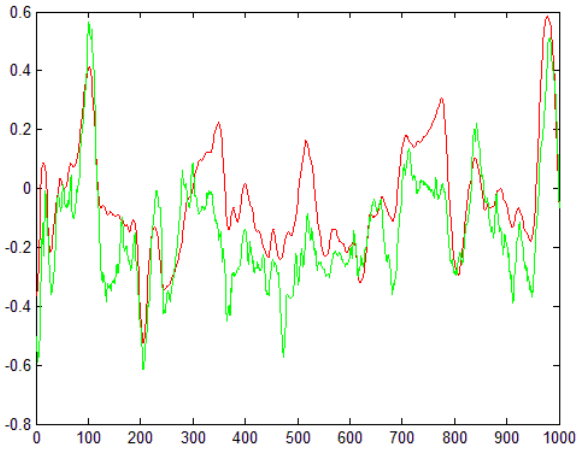

Fig. 3 Comparison between the estimated (green) and measured (red) articulatory trajectories for 1000 frames of the tongue body feature.

V. CONCLUSIONS

This paper describes our initial work in developing a real-time speech driven talking avatar system with deep neural network. The input of the system is the acoustic speech and the output is the articulatory movements synchronized with the input speech on a 3-dimentional avatar. The mapping from the input acoustic features to the output articulatory features is achieved by virtue of deep neural network (DNN). Experiments on a well known acoustic-articulatory English speech corpus MNGU0 demonstrate that the proposed audio-visual mapping method based on DNN can achieve good performance. Future work can be devoted to get smoother estimated articulator trajectories by considering the continuity properties of the articulator trajectories.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their helpful comments.

REFERENCES

[1] I. Karlsson, A. Faulkner and G. Salvi, "SYNFACE – a talking face telephone," in *Proc. of Eurospeech*, pp. 1297-1300, Geneva, Sweden, September 2003.

[2] K. Richmond, *Estimating Articulatory Parameters from the Acoustic Speech Signal*. PhD thesis, The Centre for Speech Technology Research, Edinburgh University, 2002.

[3] S. Hiroya and M. Honda, "Acoustic-to-articulatory inverse mapping using an HMM-based speech production model," in *7th International Conference on Spoken Language Processing*, 2002.

[4] L. Zhang and S. Renals, "Acoustic-articulatory modeling with the trajectory HMM," *IEEE Signal Processing Letters,* vol. 15, pp. 245-248, 2008.

[5] T. Toda, A. W. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," *Speech Communication*, vol. 50, pp. 215-227, 2008.

[6] L. Xie and Z. Q. Liu, "Realistic mouth-synching for speech-driven talking face using articulatory modelling," *IEEE Transactions on Multimedia*, vol. 9, no. 3, pp. 500-510, 2007.

[7] G. E. Hinton, S. Osindero, and Y. W. The, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527-1554, 2006.

[8] L. Deng, "An overview of deep-structured learning for information processing," in *Proc. Asian-Pacific Signal & Information Processing Annual Summit & Conference*, pp. 1-14, 2011.

[9] A. Mohamed, G. Dahl, and G. Hinton, "Deep belief networks for phone recognition," in *Proc. NIPS 2009 Workshop on Deep Learning for Speech Recognition and Related Applications,* 2009.

[10] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[11] K. Richmond, P. Hoole, and S. King, "Announcing the electromagnetic articulography (day 1) subset of the MNGU0 articulatory corpus," in *Proc. Interspeech*, pp. 1505-1508, 2011.

[12] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight," in *Proc. Int. Workshop Models and Analysis of Vocal Emissions for Biomedical Application*, 2001.