A New Language Independent, Photo-realistic Talking Head Driven by Voice Only

Xinjian Zhang[12], *Lijuan Wang*[1], *Gang Li*[1], *Frank Seide*[1], *Frank K. Soong*[1]

[1] Microsoft Research Asia, Beijing, China

[2] Department of Computer Science, Shanghai Jiao Tong University, Shanghai, China

zha@sjtu.edu.cn, {lijuanw, ganl, fseide, frankkps}@microsoft.com

Abstract

We propose a new photo-realistic, voice driven only (i.e. no linguistic info of the voice input is needed) talking head. The core of the new talking head is a context-dependent, multi-layer, Deep Neural Network (DNN), which is discriminatively trained over hundreds of hours, speaker independent speech data. The trained DNN is then used to map acoustic speech input to 9,000 tied "senone" states probabilistically. For each photo-realistic talking head, an HMM-based lips motion synthesizer is trained over the speaker's audio/visual training data where states are statistically mapped to the corresponding lips images. In test, for given speech input, DNN predicts the likely states in their posterior probabilities and photo-realistic lips animation is then rendered through the DNN predicted state lattice. The DNN trained on English, speaker independent data has also been tested with other language input, *e.g.* Mandarin, Spanish, *etc.* to mimic the lips movements cross-lingually. Subjective experiments show that lip motions thus rendered for 15 non-English languages are highly synchronized with the audio input and photo-realistic to human eyes perceptually.

Index Terms: deep neural net, voice driven, lip-synching, talking head.

1. Introduction

Talking heads have a wide range of applications, including video games and movie characters, assisted language teachers and virtual guides, *etc.* Highly realistic characters, such as those seen in movies, require team of expert artists and animators and involve months of manual effort. The idea of being able to automatically generate a facial animation from speech is therefore a highly attractive proposition. Given such a technique, an actor's voice track could be used to automatically animate a facial model, particularly lip-synching. This has advantages over *e.g.* performance driven animation which additionally involves physically recording an actor's performance using a capture system. Automatically speech driven animation also has great potential in online video games, such as World of Warcraft. In this case, the voice of a person speaking to their friends may be translated onto their virtual avatar, stepping to a more engaging and vivid user experience.

Besides the quality auto lip-synching desired in these applications, another important aspect of any such system is that it should be robust to the sound of different people such that it should be able to generate appropriate actions given voices it has not heard before. Also, multi-lingual features become more and more indispensable as many applications like online video games and movies are distributed to different countries worldwide. Therefore, lip-synching, speaker and language independence are three problems we are trying to address in the automatic voice driven systems.

In previous studies, two general approaches are usually considered: phoneme driven animation or direct mapping from audio to visual space.

In direct audio-visual conversion, the main challenge in attempting to automatically generate visual parameters from speech is to learn the complex many-to-many mappings between the signals. Massaro, *et al.* [1] use an artificial neural network to map the MFCC to visual parameters. Wang, *et al.* [2] use a single hidden Markov model to realize the mapping between Mel-Frequency Cepstral Coefficients (MFCC) and Facial Animation Parameters (FAP). Xie, *et al.* [3] propose a coupled HMM to realize video realistic speech animation. Fu, *et al.* [4] give a comparison of several single HMM based conversion approaches. Zhuang, *et al.* [5] propose a method using minimum converted trajectory error criterion to optimize the single Gaussian Mixture Model (GMM) training to improve the audio-visual conversion. But these methods are inherently speaker dependent, the challenge is then to make such a system speaker independent, such that it can generate new animations from voice identities it has not heard before.

Phoneme-based methods model the audio-visual data with different phone models. Sun, *et al.* [6] use phone-based keyframe interpolation for lips animation. Xie, *et al.* [7] transform speech signals to phone labels with ASR, then mapping them to visemes using a fixed table, where the visemes are modeled by HMM. These models usually synthesize the visual parameters from a phone sequence that is either provided by human labelers or by an automatic speech recognizer (ASR). While the former is expensive and subject to inconsistency resulting from human disagreement in phone labeling, the latter requires a well-trained speech recognizer that is usually complex and in need of handmade labels for training.

In response to the above issues, we propose to use the context dependent triphone tied state as the intermediate representation in converting from speech to lips. This is inspired by the high state accuracy achieved by recent success of context dependent, multi-layer deep neural network in ASR tasks. CD-DNN-HMMs [8], [9] are a recent very promising and possibly disruptive acoustic model. For speaker-independent single-pass recognition, it achieved relative error reductions of 16% on a business-search task, and of up to one-third on the Switchboard phone-call transcription benchmark [10], which are trained with error back-propagation [11] using the frame-based cross-entropy (CE) objective, over discriminatively trained GMM-HMMs. And [12] shows most the gain will be carried over to tasks with much larger acoustic mismatch and variety data sets.

In this paper, we propose a voice driven talking head based on the decoded tied state sequence from a context-dependent, multi-layer, DNN trained over hundreds of hours of speaker independent data. For given speech input, DNN predicts likely states in terms of their posterior probabilities. Photorealistic lip animation is then rendered through the DNN

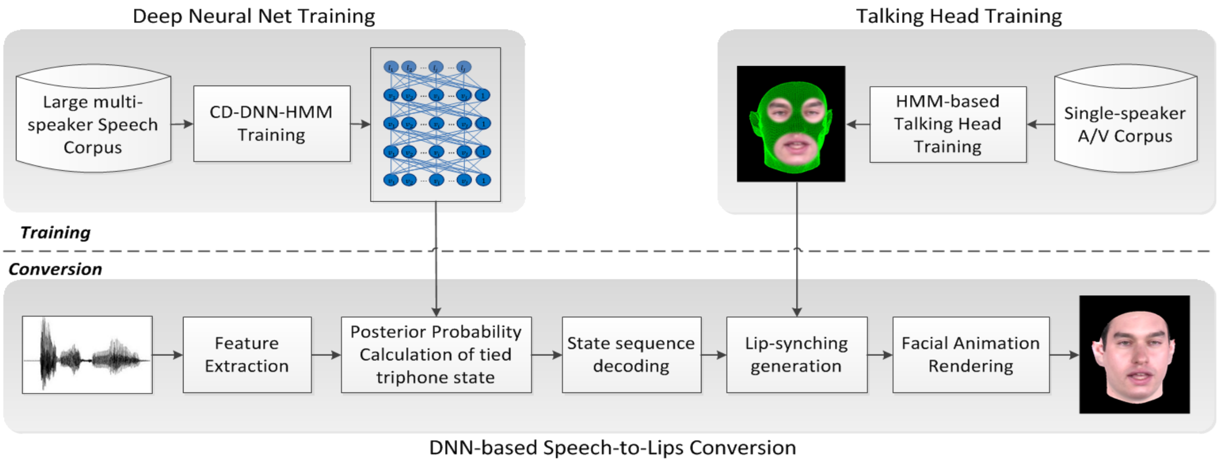
Deep Neural Net Training
Large multi-speaker Speech Corpus
CD-DNN-HMM Training
Talking Head Training
HMM-based Talking Head Training
Single-speaker A/V Corpus
Training
Conversion
Feature Extraction
Posterior Probability Calculation of tied triphone state
State sequence decoding
Lip-synching generation
Facial Animation Rendering
DNN-based Speech-to-Lips Conversion

Figure 1: *Framework of the proposed voice-driven lip-synching with DNN.*

predicted state lattice with the HMM lips motion synthesizer. Objective and subjective experiments show that the voice driven lip-synching is robust to recognition errors, speaker differences, and even language variations.

The rest of the paper is organized as follows: Section 2 gives an overview of the whole system; Section 3 and 4 briefly review the CD-DNN-HMM model training and HMM-based talking head model training; Section 5 introduces our proposed method, followed by experimental results and discussions in Section 6 and conclusions in Section 7.

2. System overview

Fig.1 shows the block diagram of the whole system, which contains two, training and conversion, phases.

In training, a context-dependent, multi-layer, Deep Neural Network (DNN) is first trained with error back-propagation procedure over hundreds of hours of speaker independent data. A highly discriminative mapping between acoustic speech input and 9000 tied states is thus established. Additionally, an HMM-based lips motion synthesizer is trained over a speaker's audio/visual data and where each state is statistically mapped to its corresponding lips images. In conversion, for given speech input, DNN predicts likely states in terms of their posterior probabilities. Photorealistic lip animation is then rendered through the DNN predicted state lattice with the HMM lips motion synthesizer. Next, we will introduce the training and conversion modules one by one.

3. The context-dependent deep-neural-network HMM

A deep neural network (DNN) is a conventional multi-layer perceptron (MLP[13]) with many hidden layers, where training is typically initialized by a pretraining algorithm. Below, we describe the DNN; briefly touch upon its training in practice. Extra details can be found in [9].

3.1. Deep neural network

A DNN models the posterior probability $P_{s|o}(s|o)$ of a class s given an observation vector o, as a stack of (L+1) layers of log-linear models. The first L layers, $\ell = 0, \dots, L-1$, model posterior probabilities of conditionally independent hidden binary units h^ℓ given input vectors v^ℓ, while the top layer L models the desired class posterior as,

$$P^{\ell}_{\boldsymbol{h}|\boldsymbol{v}}(h^\ell|v^\ell) = \prod_{j=1}^{N^\ell} \frac{e^{z_j^\ell(v^\ell)\cdot h_j^\ell}}{e^{z_j^\ell(v^\ell)\cdot 1}+e^{z_j^\ell(v^\ell)\cdot 0}} \quad , \quad 0 \le \ell < L \qquad (1)$$

$$P^{L}_{\boldsymbol{s}|\boldsymbol{v}}(s|v^L) = \frac{e^{z_s^\ell(v^L)}}{\sum_{s'} e^{z_{s'}^\ell(v^L)}} = softmax_s(z^L(v^L)) \qquad (2)$$

$$z^\ell(v^\ell) = (W^\ell)^T v^\ell + \alpha^\ell \qquad (3)$$

with weight matrices W^ℓ and bias vector α^ℓ, where h_j^ℓ and $z_j^\ell(v^\ell)$ are the j-th component of h^ℓ and $z^\ell(v^\ell)$, respectively. The precise modeling of $P_{s|o}(s|o)$ requires integration over all possible values of h^ℓ across all layers which is infeasible. An effective practical trick is to replace the marginalization with the "mean-field approximation" [14]. Given observation o, we set $v^0 = o$ and choose the conditional expectation $E^\ell_{h|v}\{h^\ell|v^\ell\} = \sigma\left(z^\ell(v^\ell)\right)$ as input $v^{\ell+1}$ to the next layer, with component-wise sigmoid $\sigma_j(z) = 1/(1+e^{-z_j})$.

3.2. Training

DNNs, being 'deep' MLPs, can be trained with the well-known error back-propagation procedure (BP) [11]. Because BP can easily get trapped in poor local optima for deep networks, it is helpful to 'pretrain' the model in a layer-growing fashion. [10] shows that two pretraining methods, deep belief network (DBN) pretraining [15, 16, 17] and discriminative pretraining, are approximately equally effective.

The CD-DNN-HMM's model structure (phone set, HMM topology, tying of context-dependent states) is inherited from a matching GMM-HMM model that has been ML-trained on the same data. That model is also used to initialize the class labels $s(t)$through forced alignment.

DNN training is an expensive operation. The model used in this paper has 7 layers of 2k hidden nodes and 9304 senones. The total number of parameters is 45.4 million, with the majority being concentrated in the output layer. Using a single server equipped with a highend NVidia Tesla S2070 GPGPU, it took 10 days to train this model.

4. HMM-based photo-realistic talking head

The voice driven animation is retargeted to a photo-realistic avatar [18]. Below, we briefly review the process of how to build such a talking head model.

In training, audio/visual footage of a speaker is used to train the statistical audio-visual Hidden Markov Model (AV-HMM). The input of the HMM contains both the acoustic features and the visual features. The acoustic features consist of Mel-Frequency Cepstral Coefficients (MFCCs), their delta and delta-delta coefficients. The visual features include the PCA coefficients and their dynamic features. The contextual dependent HMM is used to capture the variations caused by different contextual features. Also, the tree-based clustering technique is applied to the acoustic and visual features respectively to improve the robustness of the HMM.

In synthesis, the input phoneme labels and alignments are firstly converted to a context-dependent label sequence. Meanwhile, the decision trees generated in the training stage are used to choose the appropriate clustered state HMMs for each label. Then a parameter generation algorithm is used to generate the visual parameter trajectory in the maximum probability sense. The HMM predicted trajectory is used to guide the selection of succinct mouth sample sequence from the image library. The remaining task is to stitch the lips image sequence into a full face background sequence.

5. DNN-based lip-synching generation

Once the DNN and talking head model get ready, for given speech input, we use DNN predicts likely states in terms of their posterior probabilities. Then realistic lip motion can be rendered from the predicted state sequence with the talking head model synthesizer.

5.1. Feature extraction

13-dimensional PLP features with rolling-window mean-variance normalization and up to third-order derivatives, which for the GMM-HMM systems is reduced to 39 dimensions by HLDA, while in DNN training we directly use 52 dimensions feature before HLDA, because [10] shows DNN can learn HLDA implicitly.

5.2. State sequence decoding

The CD-DNN-HMM model gets the features as input and generates the posterior probability of every state for every frame according to Eq. 1-3. For decoding and lattice generation, the "senone" posteriors are converted into the HMM's emission likelihoods by dividing the "senone" priors $P(s)$:

$$\log p(o|s) = \log P(s|o) - \log P(s) + \log p(o) \quad (4)$$

where o is a regular acoustic feature vector augmented with neighbor frames (5 on each side in our case), $p(o)$ is unknown but can be ignored as it cancels out in best-path decisions.

After converting DNN generated state posteriors to likelihoods, standard decoding can be carried out within the HMM framework. With phone list and phone trigram, phone decoding results can be generated; with word dictionary and word trigram language model, we can get word decoding results. Both word and phone decoding can generate "senone" sequences as byproduct. However, we find it beneficial to simplify it to do state sequence decoding directly, which is time saving, no language dependent constraints.

State sequence decoding is to find an optimal state sequence given the tied state lattice estimated by the DNN. One way is to simply choose the most likely tied state at each frame, but this will cause different states switching frequently along the path so that the faces finally rendered are shaky. To avoid this, we further constrain the state transition between neighboring frames. The optimization function is formulated as the product of likelihood and the state transition probability:

$$P(O_1^T|S_1^T) = p(o_1|s_1) * \prod_{t=2}^{T} p(o_t|s_t)p(s_{t-1}, s_t) \quad (5)$$

$S_1^T = s_1 s_2 \ldots s_T$ is the tied state sequence, $p(s_{t-1}, s_t)$ is the non-normalized state transition probability between neighboring frames. If $s_{t-1}$and s_t are the same state, or they belong to the same central phone class, $p(s_{t-1}, s_t)$ is set to 1; otherwise $p(s_{t-1}, s_t)$ is set to a constant value less than 1 and serves as a penalty to this transition. Adding transition cost forces the state path to be relatively smooth while maximizing the total probability. The value of transition penalty is later determined through a greedy search experiment on a development data set, where under different penalty setting the difference of the final converted lips movement trajectory between the ground truth is calculated and the one that minimizes the difference is chosen. Our goal is to find the best state sequence $\hat{S}_1^T$ that maximizes $P(O_1^T|S_1^T)$. Applying Viterbi search to Eq. 5, the best path $\hat{S}_1^T$ can be found.

5.3. Lip motion rendering

Once the optimal state sequence $\hat{S}_1^T$ is ready, the audio-visual HMM λ trained for the talking head in section 4 can predict the lip motion visual trajectory in a maximum probability sense [19]. The best visual trajectory $V = [V_1^\top, V_2^\top, \cdots, V_T^\top]^\top$ is determined by maximizing the following likelihood function.

$$\begin{aligned} \log p\left(V\middle|\hat{S}_1^T, \lambda\right) &= \log p\left(V|\hat{\mu}^{(V)}, \widehat{U}^{(VV)}\right) \\ &= -\frac{1}{2} V^\top \widehat{U}^{(VV)^{-1}} V + V^\top \widehat{U}^{(VV)^{-1}} \hat{\mu}^{(V)} + const, \quad (6) \end{aligned}$$

where

$$\hat{\mu}^{(V)} = \left[\hat{\mu}_{\hat{s}_1}^{(V)}, \hat{\mu}_{\hat{s}_2}^{(V)}, \cdots, \hat{\mu}_{\hat{s}_T}^{(V)}\right]^\top, \quad (7)$$

$$\widehat{U}^{(VV)^{-1}} = diag\left[\widehat{\Sigma}_{\hat{s}_1}^{(VV)^{-1}}, \widehat{\Sigma}_{\hat{s}_2}^{(VV)^{-1}}, \cdots, \widehat{\Sigma}_{\hat{s}_T}^{(VV)^{-1}}\right]^\top. \quad (8)$$

By setting $\frac{\partial}{\partial C} \log p\left(V\middle|\hat{S}_1^T, \lambda\right) = 0$, where $V = W_c C$[19], we obtain $\widehat{V}_{opt}$ by solving a weighted least square solution.

The HMM predicted visual trajectory $\widehat{V}_{opt}$ is then used to render the photo-realistic lip movement for our talking head.

6. Experimental results

6.1. Experiment setup

The CD-DNN-HMMs model in the paper is trained using the 309-hour Switchboard-I training set [20]. The system uses 13-dimensional PLP features with rolling-window mean-variance normalization and up to third-order derivatives, 52 dimensions in CD-DNN-HMM, reduced to 39 dimensions by HLDA in GMM-HMM. The speaker-independent cross-word triphones use the common 3-state topology and share 9304 CART-tied states. The model is trained on alignment by 60 mixtures GMM-HMM with 7 data sweep, consistent of 52x11 dimensions in input layer, 7 layers of 2k hidden nodes and 9304 senones in output layer. The WER on Hub5'00 SWB test set is reduced from 26.2 to 17.2.

The HMM-based talking head model is trained with an AV database recorded by ourselves, called MT dataset for convenience. This dataset has 497 video files with corresponding audio track, each being one English sentence spoken by a single native speaker with neutral emotion. The video frame rate is 30 frames/sec. For each image, Principle Component Analysis (PCA) projection is performed on automatically detected and aligned mouth image, resulting in a 60-dimensional visual parameter vector. Mel-Frequency Cepstral Coefficient

(MFCC) vectors are extracted with a 20ms time window shifted every 5ms. The visual parameter vectors are interpolated up to the same frame rate as the MFCCs. The A-V feature vectors are used to train the HMM models using HTS 2.1 [21] for lip motion rendering.

To evaluate the performance of our proposed method, we first test it on the MT dataset which has the AV recordings so that the voice driven lip motion can be compared with the original recordings by objective measurement. We also compare the method using tied state decoding with the traditional word and phone decoding. Then we test it on a more challenging dataset which contains 15 different languages spoken by different speakers. As this multi-lingual dataset is audio only, the results are evaluated subjectively by AB test.

6.2. Objective results

We try the three different decoding methods on the MT dataset, state, phone, and word decoding, to compare their impact on the final lip rendering results. The DNN decoded state accuracy on the MT test set is about 50%, similar to the number reported on Switchboard test set. Table 1 shows the word error rate (WER) and phone error rate (PER) of word and phone decoding.

The voice driven lip rendering results are first compared with the results of the ground truth label (Table 2). Then they are compared with the original lip recordings (Table 3). Both objectively measured by root-mean-square error (RMSE), average correlation coefficient (ACC) of the PCA parameter trajectories. In each cell of Table2&3, the first number represents the average results of all the 20 PCA dimensions; the second number represents the results of the first PCA dimension. Both the RMSE and ACC results show that the result of using state decoding is statically close to that of using word or phone decoding. In some cases, word decoding generates slightly better results than the state decoding method by considering syntactic information (dictionary and language model). However, word decoding may also suffer serious errors when encountering out of vocabulary (OOV) words which are unavoidable. Fig. 2 shows a test case in our dataset in which "*herb was as ready for new adventures as he was for new ideas*." is misrecognized as "*i heard was ready ...*" We can see that when the word decoding errors happen at the beginning, the derived PCA trajectory of the first 300 frames drift away from the ground truth trajectory. In contrast, state decoding is robust to OOVs and pronunciation variations because there are no phone set, dictionary, and language model constraints.

Table 1. *WER & PER for word and phone DNN decoding*

	WER(%)	PER(%)
word	16.20	11.85
phone	N/A	18.00

Table 2. *Voice driven results vs. Ground truth label*

	Word	Phone	Tied state
RMSE	185/490	241/638	234/616
ACC	0.85/0.94	0.76/0.90	0.76/0.91

Table 3. *Voice driven results vs. Original recordings*

	Word	Phone	Tied State	Ground Truth
RMSE	385/923	411/996	353/833	408/993
ACC	0.54/0.83	0.49/0.81	0.49/0.81	0.60/0.87

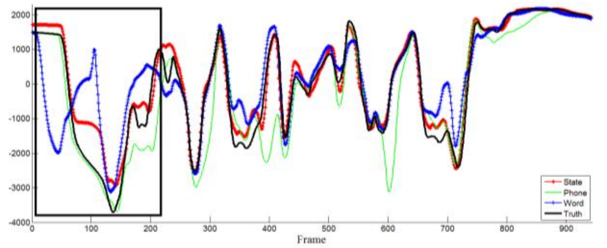

Figure 2: *PCA trajectory in presence of a recognition error.*

6.3. Subjective results

We do A/B subjective test between our state decoding voice driven results and the results with the ground truth labels. Ten pairs of video sentences are generated from the audios in MT dataset. Each pair of video clips is shuffled randomly. Eight volunteers participant this AB test, they are asked to choose the one they think better lip-synched, or choose equal if they can't decide. Fig.3 shows no dominate preference to either the ground truth or the state decoding results. It means the voice driven lip motion is close to as if we know the ground truth.

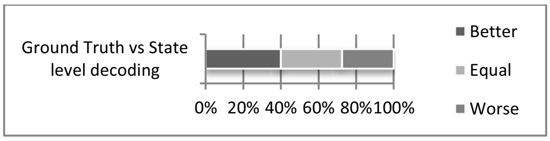

Figure 3: *Results of A/B test: ground truth vs. state decoding.*

In another subjective experiment, we test the proposed method on 15 different non-English languages. We choose 2 audio sentences from each language, so there are total 30 sentences for each decoding method and in total 90 pairs between the three decoding methods. We divide the 90 pairs into 3 sessions. Each participant takes one session. There are 9 people taking part in this test. Fig.4 shows that in most cases, state decoding results are better than phone and word decoding results. It is interesting to see that the English trained DNN can decode other foreign languages as a sequence of "seones" and use them to render convincing lip motion highly synchronized with audio. The results demonstrate that the proposed voice driven lip synching is language independent.

Video stimuli used in the experiments are available at: research.microsoft.com/en-us/projects/voice_driven_talking_head/

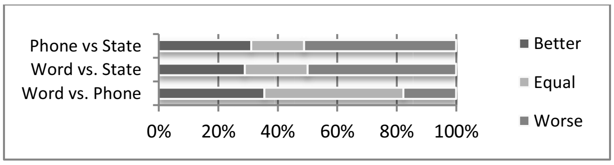

Figure 4: *Results of A/B test in 15 non-English languages.*

7. Conclusions

We propose a voice driven talking head based on the decoded tied state sequence from a context-dependent, multi-layer, DNN trained over speaker independent English data. By using the context dependent triphone tied state as the intermediate representation in converting from speech to lips, the proposed method is independent of speaker and language variations. Objective and subjective experiments show that lip motions thus rendered are highly synchronized with the audio input and photo-realistic to human eyes perceptually.

8. References

[1] Massaro, D.W., Beskow, J., Cohen, M.M., Fry, C.L. and Rodriguez, T., "Picture My Voice: Audio to Visual Speech Synthesis using Artificial Neural Networks", in Audio-Visual Speech Processing, 1999.

[2] Wang, G.-Y., Yang, M.-T., Chiang, C.-C., Tai, W.-K., "A Talking Face Driven by Voice using Hidden Markov Model", in Journal of Information Science and Engineering. 22(5):1059-1075, 2006.

[3] Xie, L., Liu, Z.-Q., "A Coupled HMM Approach to Video-Realistic Speech Animation", in Pattern Recognition, 40(8):2325-2340, 2007.

[4] Fu, S., Gutierrez-Osuna, R., Esposito, A., Kakumanu, P. K. and Garcia, O. N., "Audio/Visual Mapping with Cross-Modal Hidden Markov Models," in IEEE Transactions on Multimedia, 7(2):243–252, April 2005.

[5] Zhuang, X.-D., Wang, L.-J., Soong, F.K., Hasegawa-Johnson, M., "A Minimum Converted Trajectory Error (MCTE) Approach to High Quality Speech-to-Lips Conversion", in Interspeech, 1736-1739, 2005.

[6] Sun, N., Suigetsu, K., Ayabe, T., "An Approach to Speech Driven Animation", in IIH-MSP, 113-116, 2006.

[7] Xie L., Jiang, D., Ilse R., Wemer, V., Hichem, S., Velina, S., Zhao, R., "Context Dependent Viseme Models for Voice Driven Animation", in EC-VIP-MC 2003.4th EURASIP Conference Focused on Video / Image Processing and Multimedia Communications, 2: 649-654, 2003.

[8] Yu, D., Deng, L., and Dahl, G., "Roles of Pretraining and Fine-Tuning in Context-Dependent DNN-HMMs for Real-World Speech Recognition," in NIPS Workshop on Deep Learning and Unsupervised Feature Learning, Dec. 2010.

[9] Dahl, G., Yu, D., Deng, L., Acero, A. "Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition", in IEEE Transactions on Audio, Speech and Language Processing 20(1):30-42, 2012.

[10] Seide, F., Li, G., Chen, X., Yu, D. "Feature Engineering in Context-Dependent Deep Neural Networks for Conversational Speech Transcription", in ASRU, 24-29, 2011.

[11] Rumelhart, D., Hinton, G., Williams, R., "Learning Representations by Back-Propagating Errors", in Nature, vol. 323, Oct.,1986.

[12] Li, G., Zhu, H.-F., Cheng, G., Thambiratnam, K., Chitsaz, B., Yu, D., Seide, F., "Context-dependent Deep Neural Networks for Audio Indexing of Real-life Data", SLT, 143-148, 2012.

[13] Rosenblatt, F., "Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms", Spartan Books, Wash. DC, 1961.

[14] Saul, L. K., Jaakkola, T., and Jordan, M. I., "Mean Field Theory for Sigmoid Belief Networks", in Journal: Computing Research Repository–CORR, 61-76, 1996.

[15] Hinton, G., Osindero, S., and The, Y., "A Fast Learning Algorithm for Deep Belief Nets", in Neural Computation, 18:1527–1554, 2006.

[16] Hinton, G., "A Practical Guide to Training Restricted Boltzmann Machines", in Technical Report UTML TR 2010–003, University of Toronto, 2010.

[17] Mohamed, A., Dahl, G., and Hinton, G., "Deep Belief Networks for Phone Recognition", in NIPS Workshop Deep Learning for Speech Recognition, 2009.

[18] Wang, L.-J., Qian, Y., Scott, M.R., Chen, G., Soong, F.K., "Computer-Assisted Audiovisual Language Learning", in IEEE Computer 45(6):38-47, 2012.

[19] Wang, L.-J., Han, W., Qian, X.-J., Soong, F., "Synthesizing Photo-Real Talking Head via Trajectory-Guided Sample Selection", Interspeech, 446-449, 2010.

[20] Godfrey, J. and Holliman, E., "Switchboard-1 Release 2", in Linguistic Data Consortium, Philadelphia, 1997.

[21] Tokuda, K., Zen, H., etc., "The HMM-based speech synthesis system (HTS)", Online: http://hts.ics.nitech.ac.jp/, accessed on 13 March 2013.

[22] Salvi, G., Beskow, J., Moubayed, S.A., Granström, B., "Syn-Face-Speech-Driven Facial Animation for Virtual Speech-Reading Support", EURASIP J. Audio, Speech and Music Processing, 2009.