Scientific Research Publishing

Journal of Computer and Communications, 2017, 5, 50-65
http://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

HMM-Based Photo-Realistic Talking Face Synthesis Using Facial Expression Parameter Mapping with Deep Neural Networks

Kazuki Sato, Takashi Nose, Akinori Ito

Department of Communication Engineering, Graduate School of Engineering, Tohoku University, Sendai, Japan
Email: tnose@m.tohoku.ac.jp

How to cite this paper: Sato, K., Nose, T. and Ito, A. (2017) HMM-Based Photo-Realistic Talking Face Synthesis Using Facial Expression Parameter Mapping with Deep Neural Networks. *Journal of Computer and Communications*, **5**, 50-65.
https://doi.org/10.4236/jcc.2017.510006

Received: July 11, 2017
Accepted: August 20, 2017
Published: August 23, 2017

Abstract

This paper proposes a technique for synthesizing a pixel-based photo-realistic talking face animation using two-step synthesis with HMMs and DNNs. We introduce facial expression parameters as an intermediate representation that has a good correspondence with both of the input contexts and the output pixel data of face images. The sequences of the facial expression parameters are modeled using context-dependent HMMs with static and dynamic features. The mapping from the expression parameters to the target pixel images are trained using DNNs. We examine the required amount of the training data for HMMs and DNNs and compare the performance of the proposed technique with the conventional PCA-based technique through objective and subjective evaluation experiments.

Keywords

Visual-Speech Synthesis, Talking Head, Hidden Markov Models (HMMs), Deep Neural Networks (DNNs), Facial Expression Parameter

1. Introduction

In our daily life, facial information is important to enrich speech communication. The face-to-face communication can give us not only linguistic information but also the facial identity and expressions, which sometimes plays an essential role to make a person be relieved, attracted, or affected. The same thing can be said in human-computer interaction. A spoken dialogue system with facial information is richer than that with only speech, and it often gives friendlier impression to users. For example, a virtual agent with the face of a famous person could easily attract consumers in a shop or a public space. Therefore, a visu-

al-speech synthesis, *i.e.*, creating a talking head with synthetic speech and facial animation, is an interesting topic for more advanced man-machine interfaces.

There have been many studies for visual-speech synthesis [1] [2]. While some studies aimed to generate faces of simple [3] and detailed [3] 3DCG characters, the target of the most studies was in synthesizing photo-realistic human faces, which is a more challenging task. In the early years of visual-speech synthesis, they only focused on synthesizing the image of a speaker's mouth area [4] [5] due to the limitation of computational power. As the performance of computers improved, researchers started to examine entire-face synthesis with a variety of approaches. When we can prepare a large amount of facial video samples, a promising approach is to use synthesis techniques based on visual unit selection [6] [7] [8] that was inspired by the idea in speech synthesis (e.g., [9]). Since video snippets of tri-phone have been used as basic concatenation units, the resulting database can become very large. The use of smaller units, *i.e.*, image samples, showed their effectiveness in improving the coverage of candidate units with smaller footprint [10] [11].

Although the unit-selection-based synthesis has an advantage in the quality of synthetic facial motion, there are restrictions that the recording cost is high and the face position is fixed to keep the continuity between visual units. One approach to overcoming the problem is to use a facial 3DCG model [12] [13]. In this approach, the face model has 3D mesh and photo-realistic texture information, and a high-quality rendered animation can be produced. However, the rendering needs high computational cost, and hence real-time rendering is not always possible in low-resource devices such as mobile phones and tablets. From the viewpoint of the footprint and computational cost, a 2D image-based modeling approach can be an alternative choice [5] [14] [15]. In [14] and [16], multidimensional morphable model (MMM) [17] and active appearance model (AAM) [18] were used to model and parametrize the 2D face images, where the face images are represented by shape and texture (appearance) parameters. By parametrizing 2D face image, the parameter sequences can be statistically modeled using context-independent [14] or context-dependent [16] models. A limitation in this approach is that facial key points must be labeled by hand for the training images of facial models. In contrast, the facial animation generation based on hidden Markov models (HMMs) with non-parametric features has an advantage [5] since no manual labeling of facial parameters is necessary.

In this paper, we revise the HMM-based visual-speech synthesis to synthesize not only lip images [5] but also entire-face images. The conventional technique used principal component analysis (PCA) for visual features instead for speech features in the HMM-based speech synthesis [19]. The HMM-based speech synthesis, one of the statistical parametric speech synthesis techniques, has been widely studied [20] and has high flexibility such as style control of synthetic speech [21] for the expressive speech synthesis [22]. However, the HMM-based visual speech synthesis is difficult to be applied to the face images because the

movement other than the lip region affects the PCA coefficients and degrades the synthesis performance. For this problem, we propose two-step synthesis by introducing intermediate features, *i.e.*, low-dimensional facial expression parameters [23] [24], into the modeling process. The expression parameters are associated with face images with non-linear mapping using deep neural networks (DNNs). Since the expression parameters are not affected by the face movement but well correspond to the lip movement, the proposed technique is expected to improve the modeling accuracy compared to the conventional PCA-based synthesis.

The contributions of this paper are summarized as follows. The proposed technique achieves facial animation synthesis using the HMM-based 2D image synthesis framework with facial expression parameters and DNNs. The advantage of the HMM-based system is its small footprint size [25] compared to the unit-selection and 3DCG models. In addition, our technique uses no manual labeling in the model training, which is essential to realize a visual-speech synthesizer of arbitrary speakers at a low cost. We investigate the amount of training data required for HMMs and DNNs through experiments and finally show the superiority of the proposed technique to the conventional PCA-based technique through the objective and subjective evaluation tests.

The rest of this paper is organized as follows: In Section 2, we briefly overview the conventional HMM-based visual speech synthesis techniques. Section 3 describes the proposed two-step synthesis technique using facial expression parameters and non-linear mapping using DNNs. In Section 4, the performance of the proposed facial animation synthesis technique is evaluated and is compared to the conventional PCA-based approach from objective and subjective perspectives. In Section 5, we summarize this study and give suggestions for future work.

2. Conventional Photo-Realistic Talking Face Synthesis Based on HMMs

In this section, we briefly review the conventional techniques for synthesizing photo-realistic talking face animations. As described in the introduction, the basis of this study is on the HMM-based visual-speech synthesis using PCA-based visual features [5]. This technique was inspired by the HMM-based speech synthesis where sequences of speech parameters, *i.e.*, spectral and excitation parameters, are modeled using context-dependent HMMs. The previous work [5] only focused on mouth area and the tip of the nose. Parallel data of audio and video frames are constructed but each of them are modeled separately using HMM sets with different number of states: five states for audio and three states for video features. The modeling method for the audio data is the same as that for HMM-based speech synthesis.

Since the number of dimensions of pixel image data were very high, it is computationally expensive to apply the HMM-based acoustic modeling to the

image data straightforwardly. Therefore, in the previous work [5], the dimensionality of the image was reduced using PCA, and each lip image was represented by a linear combinations of eigen vectors in a similar manner to the eigenface [26]. Both audio and image features are modeled by the context-dependent phone HMMs, and the durations of each phone is determined by the HMMs trained from the speech. In the model training for visual features, phonetic contextual factors are taken into account, and context-dependent HMMs are trained. State-dependent model-parameter tying using context clustering with contextual decision trees is performed because the number of possible combinations of phonetic contextual factors is enormous. The approach of [5] differs from that in the previous studies [4] [27] because dynamic features are used in addition to the static features, which reflects both static and dynamic properties of the training data to the generated feature sequences.

There have been other studies for visual-speech synthesis based HMMs. [28] proposed the lip animation synthesis where the lip image samples were concatenated using the trajectory-guided sample selection method. The guide trajectory is generated using the similar manner to [5]. Then, the optimal sequence of the image samples is determined by the cost function. The total cost is given by the weighted sum of the target and concatenation costs, which is the same as the unit selection for speech synthesis. Since this is the sample-based approach, the required amount of visual data of the target speaker is larger than that in the parametric visual-speech synthesis. In addition, the synthesis of the entire face was not investigated. There can be the same difficulty with [5] because the guide trajectory is generated from the HMMs with PCA-based features. The HMMs were also used for modeling the AAM parameters [16] instead of the PCA coefficients. In this technique, the face images and emotional expressions are simultaneously modeled and controlled using a framework of a cluster adaptive training (CAT) [29] that is also applied to HMM-based speech synthesis [30][1]. However, our approach has advantages over the AAM-based technique in that our approach does not need clipping the facial region from a image and manual labeling of facial key points.

3. Two-Step Photo-Realistic Talking Face Synthesis Using Facial Expression Parameters

In this section, we present a novel technique for synthesizing a 2D photo-realistic talking face animation. The technique has two steps to model the relation between input context-dependent labels and output pixel images. First, we give an overview of the proposed talking face synthesis system and introduce facial expression parameters as an intermediate features in the modeling. Then, the modeling and parameter generation processes are described in detail. Finally, the conversion from the expression parameters to the pixel images is explained where DNNs are used for the non-linear mapping.

[1]The basic formulation is the same as [21].

3.1. Overview of the Proposed System

Figure 1 illustrates the outline of the proposed talking face synthesis system. As is the same as the conventional PCA-based approach described in Section 2, speech and visual units for synthesis are phone HMMs, and hence lip movements are easily synchronized with auditory speech by using the same phoneme labels for synthesis even when both units are modeled separately. There are two steps for the model training stage. The first step is the modeling of facial expression parameters. In this study, we use Microsoft Kinect v2 to capture the facial video data. The facial expression parameters, called animation units (AUs), are extracted using Microsoft Face Tracking SDK[2]. The details of the expression parameters are explained in Section 3.2. Then, the expression parameter sequences are modeled by HMMs with context-dependent labels. We only use triphone context in this study. The second step is to train the mapping from expression parameters to facial pixel images where DNNs are used to achieve the non-linear

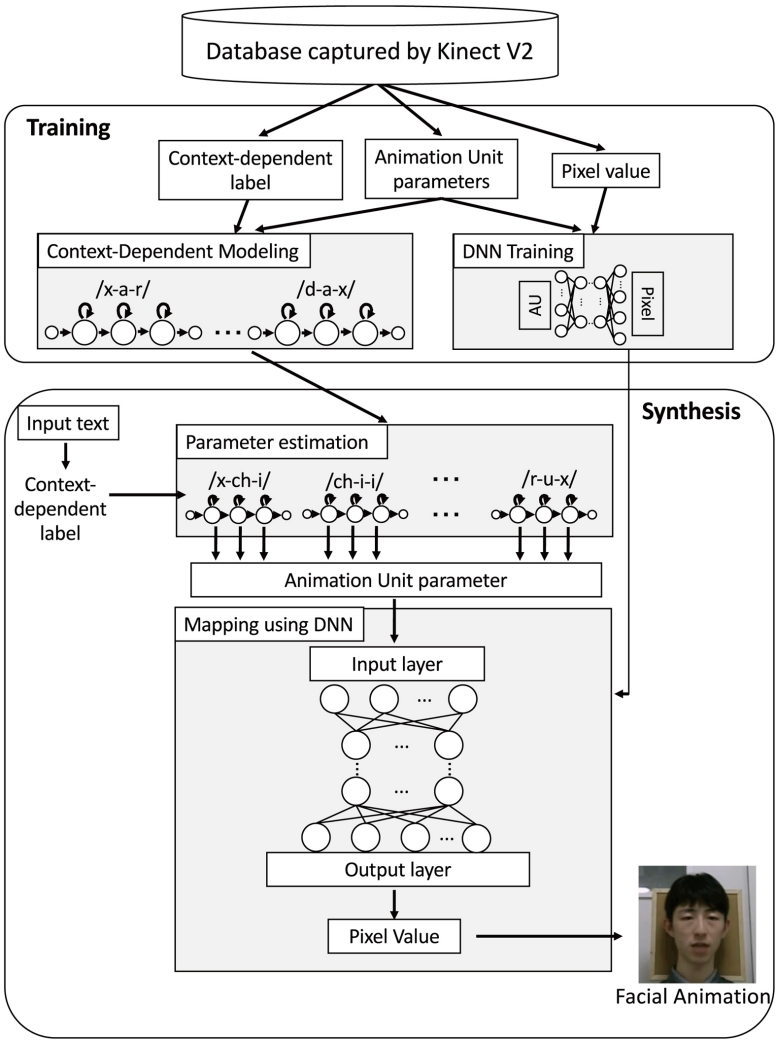

Figure 1. Overview of the proposed system.

[2]The Microsoft face tracking software development kit for Kinect for Windows.

mapping. For the training of HMMs, dynamic features are used to capture the dynamic property of the expression parameters between frames.

In the synthesis stage, the input text is converted to the context-dependent label sequence using text analysis. The context-dependent HMMs are concatenated aligned with the label sequence, and an optimal expression parameter sequence is estimated using the parameter generation algorithm based on maximum likelihood [31], which is described in Section 3.4. In the estimation, both the static and dynamic features are taken into account. Finally, the generated expression parameters are converted to the facial pixel images.

3.2. Animation Units for Facial Expression Parameters

In the conventional PCA-based synthesis, the PCA coefficients obtained from the training pixel images can be viewed as an intermediate representation. Although the PCA efficiently reduces the number of dimensions of the images, the obtained coefficients include the characteristics of the whole images. This means that the representation is sensitive not only to the lip movement but also to the face movement even though the degree is small. As a result, it is difficult to accurately model the facial parts using HMMs with context labels, and hence the applicable region is restricted only to around the lip [5] [28].

Instead of the conventional PCA coefficients, we use animation units (AUs) for the facial expression parameters as intermediate features in the modeling. AUs are seventeen parameters that represent the position and shape of the face and are expressed as a numeric weight as shown in **Table 1**. Three of the parameters, Jaw Slide Right, Right Eyebrow Lowerer, and Left Eyebrow Lowerer,

Table 1. Definition of animation unit parameter.

Number	Parameter
AU 0	Jaw Open
AU 1	Lip Pucker
AU 2	Jaw Slide Right
AU 3	Lip Stretcher Right
AU 4	Lip Stretcher Left
AU 5	Lip Corner Puller Left
AU 6	Lip Corner Puller Right
AU 7	Lip Corner Depressor Left
AU 8	Lip Corner Depressor Right
AU 9	Left cheek Puff
AU 10	Right cheek Puff
AU 11	Left eye Closed
AU 12	Right eye Closed
AU 13	Right eyebrow Lowerer
AU 14	Left eyebrow Lowerer
AU 15	Lower lip Depressor Left
AU 16	Lower lip Depressor Right

vary between −1.0 and 1.0, and the others vary between 0.0 and 1.0. Since these parameters are calculated using color and depth information with the Kinect sensor [32], there is no need to label the face images manually. The advantage of the AUs over PCA coefficients is that the AUs capture the state of the respective facial parts independently and are not affected by each other.

The idea of our approach is similar to the study for the emotional speech synthesis based on a three-layered model using a dimensional approach [33] in contrast to the categorical approach [34]. Similarly to the case of this study, the speech features are sometimes difficult to be predicted directly from the emotion dimensions. They used seventeen semantic primitives as an intermediate representation and improved the accuracy of acoustic feature estimation to synthesize affective speech more similar to that intended in the dimensional emotion space.

3.3. Modeling Facial Expression Parameter Sequences Using HMMs

Since expression parameter sequences generally have continuity in time domain, we use HMMs to model the continuity of the parameter sequences in a similar way to the HMM-based speech synthesis [19]. For the model training, we use a phone as the synthesis unit. The phone labels with phone boundary information are the same as those for the speech modeling. The parameter sequences of the respective phone segments are modeled using context-dependent HMMs. Hidden semi-Markov models (HSMMs) [35] are used for explicit modeling of state duration distribution [36]. State-based decision trees are constructed, and parameter tying using context clustering is performed to reduce the number of model parameters. The stopping criterion based on minimum description length (MDL) [37] is used for the decision tree construction in this study. Dynamic features are used as well as static features to model the dynamic property among multiple frames [38], which is used also in the very low bit-rate coding of spectral [39] and F0 [40] features of speech.

3.4. Facial Expression Parameter Generation from HMMs

In the synthesis stage of a face animation, a given text is converted to a context-dependent label sequence using text analysis. The model parameters of the facial expression parameters for unseen context labels are estimated using decision trees constructed during the model training. The context-dependent HSMMs are aligned with the label sequence, and a single sentence HSMM is created. A sequence of facial expression parameters is generated from HMMs using a maximum likelihood parameter generation algorithm [41]. In the parameter generation, both static and dynamic features are taken into account, and consequently, a smooth parameter sequence is obtained. **Figure 2** shows the effect of the dynamic features in the parameter generation. From **Figure 2(b)**, we see that the trajectory of the generated parameter sequence is not smooth when only the static feature is used. There is undesirable fluctuations between frames compared to the trajectory of the original parameter sequence. On the other

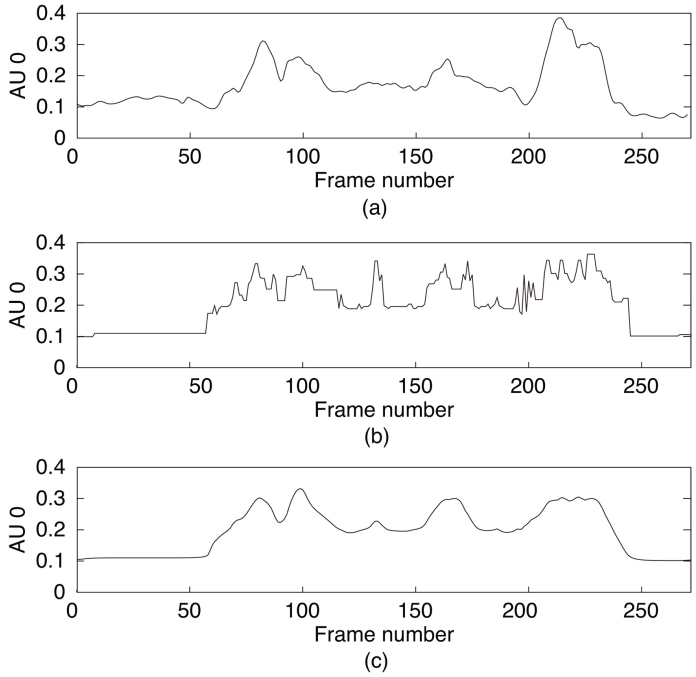

Figure 2. Example of facial expression parameter sequences generated with and without dynamic features. (a) Original; (b) without dynamic features; (c) with dynamic features.

hand, the smooth trajectory is obtained when the dynamic features are taken into account. As a result, the trajectory becomes close to that of the original parameters.

3.5. Mapping Facial Expression Parameters to Face Image Using DNNs

Finally, the facial expression parameters generated from HMMs are converted to the facial pixel image of the target speaker using DNN-based non-linear mapping. The same idea was used in our previous study for the conversion of speaker's face images [24]. Since both the 2D face image and expression parameters, *i.e.*, AUs, are simultaneously obtained using Kinect, there is a good correspondence between them. The variation of the shape of lips and other parts is smaller than that of speech parameters such as spectral and F0 features. Therefore, using whole training data, typically several tens minutes or more, is not necessary for the face image. We randomly choose frames from the training data in the similar manner to the case of PCA of the previous study [5]. The HSV space is used for the color representation on the basis of the report that the HSV space was better than the RGB space in the extraction of the lip region [42].

4. Experiments

In this section, we conducted objective evaluations to examine the appropriate setting for the model training in the proposed talking face synthesis technique. We also compared the proposed technique with the conventional PCA-based

synthesis using objective and subjective evaluation tests to show the effectiveness of introducing the intermediate features, *i.e.*, facial expression parameters, into the model training.

4.1. Experimental Conditions

For the model training, we recorded color video samples of a male speaker who uttered 103 sentences using Kinect v2. The sentences were selected from the subsets A and J of 503 phonetically balanced sentences of the ATR Japanese speech database set B [43]. The size of the images was 400 × 400 pixels. The speech and timestamp data were also recorded as well as the video data. The built-in microphone in Kinect was used for the speech recording. We used all 17 AUs as the facial expression parameters. The head position of the speaker was fixed using a headrest to suppress the face movement during the recording. The frame rate was set to 30 fps in the recording. However, since there were some dropped frames that could not capture the AUs, the frame rate of AUs was converted to 60 fps using cubic spline interpolation with the recorded timestamps. The facial regions were cut out from the recorded images using template matching and were resized to 200 × 200 pixels. In the template matching, a single face image of closed mouth, which was chosen in advance, was applied to the first fame of each utterance. Then the face region was cut out and the image was used as a new template for the next frame to improve the matching accuracy. This template update was performed frame by frame.

From the 103 sentences, 48 sentences were chosen for the candidates of the model training for HMM/DNN, 25 sentences were chosen for the validation data to obtain the optimal number of hidden layers and the number of units for DNNs, and 30 sentences were chosen for the evaluation tests. The AUs and their delta and delta-delta parameters were used as the static and dynamic features. The formulations of the dynamic features were the same as those in the HMM-based speech synthesis [19]. As a result, the total number of dimensions of the feature vector for facial expression parameters was 51. Three state left-to-right triphone HSMMs were used for the modeling of facial expression parameters. We assumed that the probability density functions in the all decision-tree leaf nodes were Gaussian with diagonal covariance matrices, which is a typical implementation in HMM-based speech synthesis. We used standard feed-forward DNNs for the parameter mapping. The conditions for the training of DNNs are listed in **Table 2**, which were the same conditions as [24].

4.2. Required Amount of Training Data

It is important to know the amount of training data that is sufficient for the model training. In this section, we objectively examined the amount of data required for the training of HMMs and DNNs. Root mean square errors (RMSEs) between the original and synthetic features were used as an objective distortion measure. In the evaluation of HMMs, the training data was changed from 4 to 48

Table 2. Structure of DNNs.

Number of units for input layer	17
Number of units for output layer	120,000
Optimizer	Adam [44]
Activation function	tanh
Batch size	100
Number of epochs	100
Dropout rate	0.5

sentences with an increment of 4 sentences. The sentences were randomly chosen from the all 48 sentences. Since the performance depends on the choice of the sentence set, we made five sets of training data for each target number of sentences. Then, the average value of RMSEs of the five sets was calculated and was used as a final RMSE, which alleviates the dependency to the choice of the sentences. The RMSE was calculated between original and generated AUs where the frames were aligned using the durations of the original speech. **Figure 3** shows the result. From the figure, we found that there was not a large variation of the RMSEs when the number of sentences was over ten. The smallest RMSE was given by the condition that the number of sentences was set to 44 in this experiment. This result indicates that the sufficient amount of training data to model the facial expression parameters using HMMs is around 50 sentences when the phonetically balanced ATR sentences are used for the training.

In the evaluation of DNNs, we randomly chose the frames for training DNNs as is described in Section 3.5. The target number of the frames was doubled from 128 frames up to 4096 frames. As was the case with the evaluation of HMMs, we made five sets of training data and used the average value of RMSEs for the five sets as the final RMSE. For the optimization of DNN structure, the candidate numbers of hidden layers were 1, 2, and 3, and the candidate numbers of hidden units in one layer were 512, 1024, and 2048. Totally, there were nine combinations of the conditions. For each condition, we calculated the RMSEs for the validation set, and finally the best combination, which gave the smallest RMSE, was chosen as the structure in each amount of training data. For the validation and test data, the RMSE was calculated between the original and generated values of pixels in the HSV color space. The frames were aligned using the durations of the original speech. **Figure 4** shows the result. From the figure, it is seen that the variation of RMSE became small when the target number of frames was set to 512 or more.

When comparing the results of DNN and HMM, we found that the required amount of training data for DNNs was much smaller than HMMs. This is because the HMMs model the continuous sequence of facial expression parameters whereas no dynamic features are taken into account in the DNN-based feature mapping and the frame-independent mapping using randomly chosen frames is sufficient.

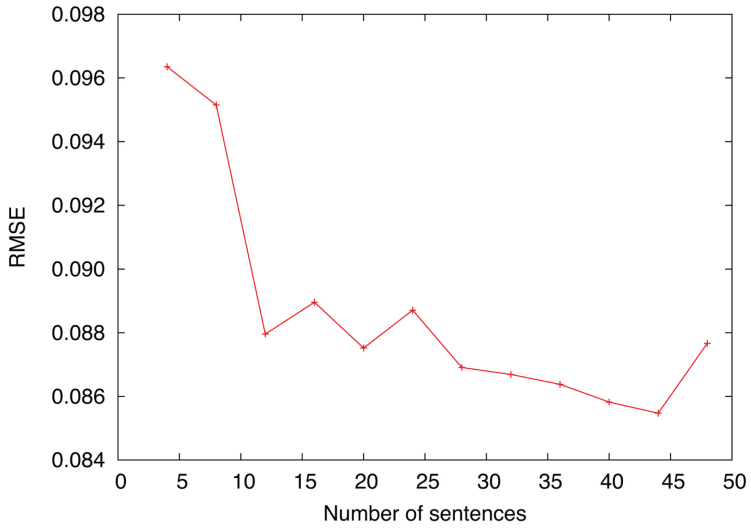

Figure 3. Variation of the objective distortions against the different amounts of training data for HMMs.

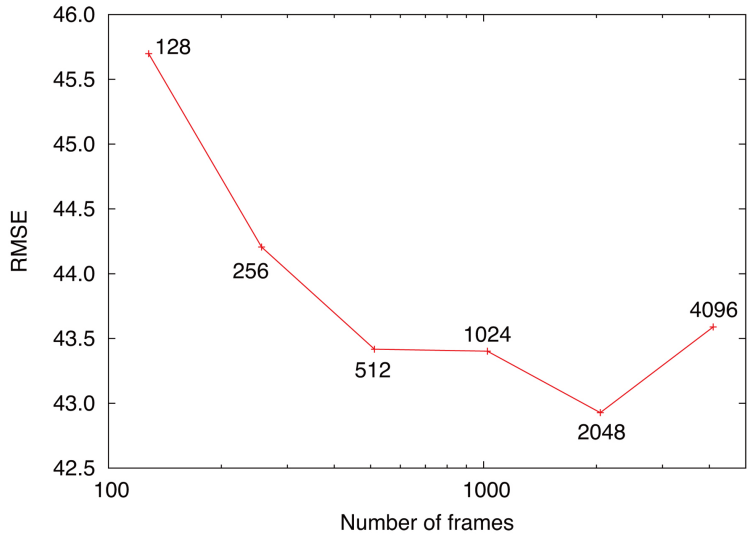

Figure 4. Variation of the objective distortions against the different amounts of training data for DNNs.

4.3. Comparison with the PCA-Based Synthesis

Next, we compared the proposed two-step synthesis technique with the conventional PCA-based synthesis technique [5] in an objective manner. In the conventional technique, the PCA was applied to the training data and 100 PCA coefficients were obtained for each frame. The cumulative contribution ratio of the obtained eigen vectors was about 90% for the training data. We conducted a preliminary experiment and confirmed that perceived degradation of the re-constructed images by the PCA was small for the original images. The feature vectors consisted of the PCA coefficients with their delta and delta-delta coefficients, and the total number of dimensions was 300. The feature vectors were used for the training of triphone HMMs whose conditions were the same as

those in the proposed technique. For the proposed technique, we used 48 sentences and 4096 frames for the training of HMMs and DNNs, respectively. The structure of the DNNs was determined using the validation data in Section 4.2, and the optimal numbers of hidden layers and units were 3 and 512, respectively, in this condition. The RMSEs of pixel data were calculated for the conventional and proposed techniques. Table 3 shows the result. From the table, we see that the proposed technique using the two-step training can synthesize closer face images than the conventional PCA-based technique.

Finally, we conducted a subjective preference test for the facial animations synthesized by the conventional and proposed techniques. The same samples as those in the objective evaluation were used for the preference test. In this test, the samples with the conventional and proposed techniques were displayed to each participant in random order. 10 sentences were randomly chosen from the 48 sentences for each participant. The participants were asked to choose the sample whose naturalness was better than the other as a facial animation. Since the sample is a photo-realistic facial animation in this evaluation, it is natural that both animation and speech were presented to the participants. Therefore, we added the original speech to the animation of each sample. Note that the lip motion and speech were synchronized because the phone durations of original speech were used in the synthesis of the facial animations. The participants were twelve undergraduate and graduate students.

Figure 5 shows the result of the preference test. The 95% confidence interval is also shown in the figure. From the figure, it is found that the proposed technique synthesized substantially better facial animations than the conventional PCA-based technique, which is consistent with the objective evaluation result. When seeing the synthetic samples of the conventional technique, we found that the motions of mouth open and close were almost not achieved and did not correspond to the phonetic information. A possible reason is that the variation of the whole pixel image affected the PCA coefficients and the contribution of the mouth shape to the coefficients became lower. As a result, the images of the mouth open and close were clustered in the same leaf node in the decision-tree-based context clustering, which crucially made the lip motion unclear. In contrast, although the quality of the synthetic image of the entire face with the proposed technique was at the same level as the conventional one, the lip motion was synthesized because of the two-step modeling and synthesis.

Figure 6 shows an example of the variation of the successive face images extracted from the facial animations of 1) original samples and synthetic samples with 2) conventional and 3) proposed techniques. In the figure of the original

Table 3. Comparison of objective distortions (RMSEs) between the conventional and proposed synthesis techniques.

PCA	Proposed
42.64	41.69

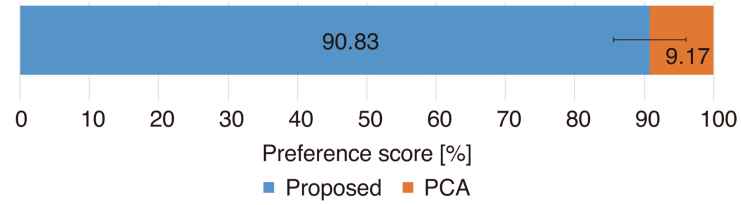

Figure 5. Result of the preference test.

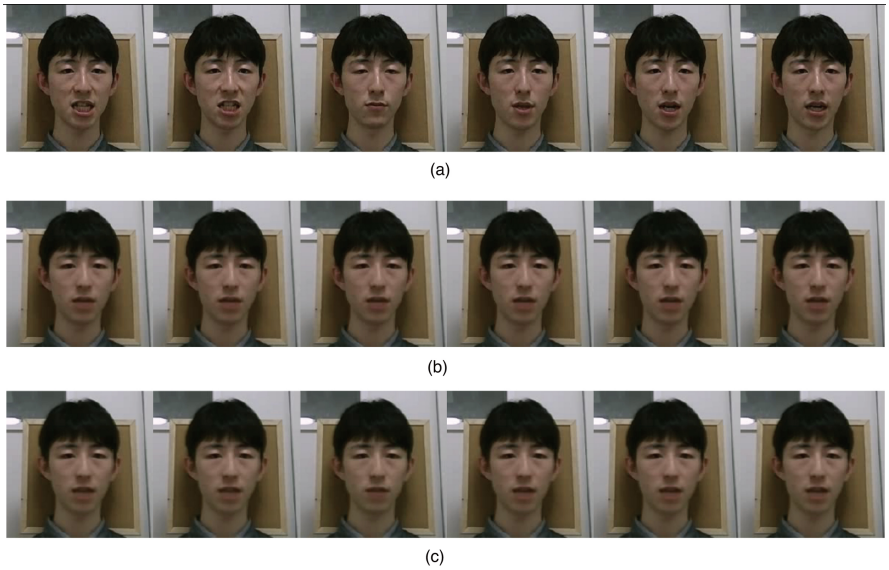

Figure 6. Comparison of the captured and synthesized frames. From left; captured, conventional method, and proposal method. (a) Original; (b) PCA; (c) proposed.

animation, the mouth was opened, closed, and opened again. However, we see no variation in the mouth region of the conventional technique. The proposed technique improves the problem and there is a difference between the frames of mouth open and close.

5. Conclusion

In this paper, we proposed a technique for synthesizing a 2D photo-realistic talking face animation using two-step synthesis with HMMs and DNNs. The key idea of the technique is the introduction of facial expression parameters as an intermediate representation that has a good correspondence both with the input contexts and the output pixel data of the face image. In the proposed technique, the facial expression parameters and pixel images are modeled using HMMs and DNNs, respectively. In the experiments, first we examined the required amount of the training data for HMMs and DNNs. The objective experimental results showed that about 50 phonetically balanced ATR sentences were sufficient for the modeling of facial expression parameters with HMMs. It was also found that the DNN training needed less amount of training data than the HMM training, which saves the computation time for the model preparation. The objective and subjective comparative experiments with the conventional PCA-based synthesis

both results in showing the superiority of the proposed technique. The remaining work includes the synthesis of expressive facial animation and the increase of the contextual factors.

Acknowledgements

Part of this work was supported by JSPS KAKENHI Grant Number JP15H02720.

References

[1] Ostermann, J. and Weissenfeld, A. (2004) Talking Faces - Technologies and Applications. *Proc. the* 17*th International Conference on Pattern Recognition* (*ICPR*), **3**, 826-833. https://doi.org/10.1109/ICPR.2004.1334656

[2] Mattheyses, W. and Verhelst, W. (2015) Audiovisual Speech Synthesis: An Overview of the State-of-the-Art. *Speech Communication*, **66**, 182-217. https://doi.org/10.1016/j.specom.2014.11.001

[3] Savran, A., Arslan, L.M. and Akarun, L. (2006) Speaker-Independent 3D Face Synthesis Driven by Speech and Text. *Signal Processing*, **86**, 2932-2951. https://doi.org/10.1016/j.sigpro.2005.12.007

[4] Brooke, N.M. and Scott, S.D. (1998) Two- and Three-Dimensional Audio-Visual Speech Synthesis. *Proc. AVSP*98 *International Conference on Auditory-Visual Speech Processing*.

[5] Sako, S., Tokuda, K., Masuko, T., Kobayashi, T. and Kitamura, T. (2000) HMM-Based Text-to-Audio-Visual Speech Synthesis. *Proc. INTERSPEECH*, 25-28.

[6] Huang, F.J., Cosatto, E. and Graf, H.P. (2002) Triphone Based Unit Selection for Concatenative Visual Speech Synthesis. *Proc. International Conference on Acoustics, Speech, and Signal Processing*, **2**, 2037-2040.

[7] Ezzat, T., Geiger, G. and Poggio, T. (2002) Trainable Videorealistic Speech Animation. *Proc. Special Interest Group on Computer GRAPHics and Interactive Techniques*, 388-398. https://doi.org/10.1145/566570.566594

[8] Mattheyses, W., Latacz, L., Verhelst, W. and Sahli, H. (2008) Multimodal Unit Selection for 2D Audiovisual Text-to-Speech Synthesis. *Inter-national Workshop on Machine Learning for Multimodal Interaction*, 125-136. https://doi.org/10.1007/978-3-540-85853-9_12

[9] Hunt, A.J. and Black, A.W. (1996) Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database. *Proc. International Conference on Acoustics, Speech, and Signal Processing*, **1**, 373-376.

[10] Cosatto, E. and Graf, H.P. (2000) Photo-Realistic Talking-Heads from Image Samples. *IEEE Transactions on Multimedia*, **2**, 152-163. https://doi.org/10.1109/6046.865480

[11] Liu, K. and Ostermann, J. (2008) Realistic Facial Animation System for Interactive Services. *Visual Speech Synthesis Challenge*, Brisbane, September 2008, 2330-2333.

[12] Cao, Y., Tien, W.C., Faloutsos, P. and Pighin, F. (2005) Expressive Speech-Driven Facial Animation. *ACM Transactions on Graphics*, **24**, 1283-1302. https://doi.org/10.1145/1095878.1095881

[13] Wang, L., Han, W., Soong, F.K. and Huo, Q. (2011) Text Driven 3D Photo-Realistic Talking Head. *Proc. INTERSPEECH*, 3307-3308.

[14] Chang, Y.-J. and Ezzat, T. (2005) Transferable Video Realistic Speech Animation. *ACM SIGGRAPH Eurographics Symposium on Computer Animation*, 143-151.

https://doi.org/10.1145/1073368.1073388

[15] Wan, V., Anderson, R., Blokland, A., Braunschweiler, N., Chen, L., Kolluru, B., Latorre, J., Maia, R., Stenger, B., Yanagisawa, K., *et al.* (2013) Photo-Realistic Expressive Text to Talking Head Synthesis. *Proc. INTERSPEECH*, 2667-2669.

[16] Anderson, R., Stenger, B., Wan, V. and Cipolla, R. (2013) Expressive Visual Text-to-Speech Using Active Appearance Models. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 3382-3389. https://doi.org/10.1109/CVPR.2013.434

[17] Jones, M.J. and Poggio, T. (1998) Multidimensional Morphable Models. *Proc. 6th International Conference on Computer Vision*, 683-688. https://doi.org/10.1109/ICCV.1998.710791

[18] Cootes, T.F., Edwards, G.J. and Taylor, C.J. (1998) Active Appearance Models. *European Conference on Computer Vision*, 484-498. https://doi.org/10.1007/BFb0054760

[19] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T. and Kitamura, T. (1999) Simultaneous Modeling of Spectrum, Pitch and Duration in HMM-Based Speech Synthesis. *Proc. Eurospeech*, 2347-2350.

[20] Zen, H., Tokuda, K. and Black, A. (2009) Statistical Parametric Speech Synthesis. *Speech Communication*, **51**, 1039-1064.

[21] Nose, T., Yamagishi, J., Masuko, T. and Kobayashi, T. (2007) A Style Control Technique for HMM-Based Expressive Speech Synthesis. *IEICE Transactions on Information and Systems*, **E90-D**, 1406-1413. https://doi.org/10.1093/ietisy/e90-d.9.1406

[22] Nose, T. and Kobayashi, T. (2011) Recent Development of Hmm-Based Expressive Speech Synthesis and Its Applications. *Proc. APSIPA ASC*, 1-4.

[23] Gui, J., Zhang, Y., Li, S., Xu, P. and Lan, S. (2015) Real-Time 3D Facial Subtle Expression Control Based on Blended Normal Maps. 8*th International Symposium on Computational Intelligence and Design*, Vol. 1, 466-469. https://doi.org/10.1109/ISCID.2015.200

[24] Saito, Y., Nose, T., Shinozaki, T. and Ito, A. (2015) Conversion of Speaker's Face Image Using PCA and Animation Unit for Video Chatting. 2015 *International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, 433-436.

[25] Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J. and Oura, K. (2013) Speech Synthesis Based on Hidden Markov Models. *Proc. the IEEE*, **101**, 1234-1252. https://doi.org/10.1109/JPROC.2013.2251852

[26] Turk, M.A. and Pentland, A.P. (1991) Face Recognition Using Eigenfaces. *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 586-591. https://doi.org/10.1109/CVPR.1991.139758

[27] Williams, J.J., Katsaggelos, A.K. and Randolph, M.A. (2000) A Hidden Markov Model Based Visual Speech Synthesizer. *Proc. International Conference on Acoustics, Speech, and Signal Processing*, Vol. 6, 2393-2396. https://doi.org/10.1109/ICASSP.2000.859323

[28] Wang, L., Qian, X., Han, W. and Soong, F.K. (2010) Synthesizing Photo-Real Talking Head via Trajectory-Guided Sample Selection. *Proc. INTERSPEECH*, 446-449.

[29] Gales, M. (2000) Cluster Adaptive Training of Hidden Markov Models. *IEEE Transactions on Speech and Audio Processing*, **8**, 417-428. https://doi.org/10.1109/89.848223

[30] Latorre, J., Wan, V., Gales, M.J., Chen, L., Chin, K., Knill, K., Akamine, M., *et al.*

(2012) Speech Factorization for HMM-TTS Based on Cluster Adaptive Training. *Proc. INTERSPEECH*, 971-974.

[31] Tokuda, K., Masuko, T., Yamada, T., Kobayashi, T. and Imai, S. (1995) An Algorithm for Speech Parameter Generation from Continuous Mixture HMMs with Dynamic Features. *Proc. Eurospeech*, 757-760.

[32] Zhang, Z. (2012) Microsoft Kinect Sensor and Its Effect. *IEEE Multimedia*, **19**, 4-10. https://doi.org/10.1109/MMUL.2012.24

[33] Xue, Y., Hamada, Y. and Akagi, M. (2015) Emotional Speech Synthesis System Based on a Three-Layered Model Using a Dimensional Approach. *Proc. APSIPA ASC*, 505-514. https://doi.org/10.1109/APSIPA.2015.7415323

[34] Yamagishi, J., Onishi, K., Masuko, T. and Kobayashi, T. (2005) Acoustic Modeling of Speaking Styles and Emotional Expressions in HMM-Based Speech Synthesis. *IEICE Transactions on Information and Systems*, **E88-D**, 503-509. https://doi.org/10.1093/ietisy/e88-d.3.502

[35] Levinson, S. (1986) Continuously Variable Duration Hidden Markov Models for Automatic Speech Recognition. *Computer Speech and Language*, **1**, 29-45.

[36] Zen, H., Tokuda, K., Masuko, T., Kobayashi, T. and Kitamura, T. (2007) A Hidden Semi-Markov Model-Based Speech Synthesis System. *IEICE Transactions on Information and Systems*, **E90-D**, 825-834. https://doi.org/10.1093/ietisy/e90-d.5.825

[37] Shinoda, K. and Watanabe, T. (2000) MDL-Based Context-Dependent Subword Modeling for Speech Recognition. *Journal of the Acoustical Society of Japan* (*E*), **21**, 79-86. https://doi.org/10.1250/ast.21.79

[38] Masuko, T., Tokuda, K., Kobayashi, T. and Imai, S. (1996) Speech Synthesis Using HMMs with Dynamic Features. *Proc. International Conference on Acoustics, Speech, and Signal Processing*, 389-392. https://doi.org/10.1109/ICASSP.1996.541114

[39] Tokuda, K., Masuko, T., Hiroi, J., Kobayashi, T. and Kitamura, T. (1998) A Very Low Bit Rate Speech Coder Using HMM-Based Speech Recognition/Synthesis Techniques. *Proc. International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, 609-612. https://doi.org/10.1109/ICASSP.1998.675338

[40] Nose, T. and Kobayashi, T. (2012) Very Low Bit-Rate F0 Coding for Phonetic Vocoders Using MSD-HMM with Quantized F0 Symbols. *Speech Communication*, **54**, 384-392.

[41] Tokuda, K., Kobayashi, T. and Imai, S. (1995) Speech Parameter Generation from HMM Using Dynamic Features. *Proc. International Conference on Acoustics, Speech, and Signal Processing*, 660-663. https://doi.org/10.1109/ICASSP.1995.479684

[42] Kuroda, T. and Watanabe, T. (1995) Method for Lip Extraction from Face Image Using HSV Color Space. *Transactions the Japan Society of Mechanical Engineers Series C*, **61**, 4724-4729.

[43] Kurematsu, A., Takeda, K., Sagisaka, Y., Katagiri, S., Kuwabara, H. and Shikano, K. (1990) ATR Japanese Speech Database as a Tool of Speech Recognition and Synthesis. *Speech Communication*, **9**, 357-363.

[44] Kingma, D. and Ba, J. (2014) Adam: A Method for Stochastic Optimization.