Journal of Korea Robotics Society (2018) 13(1):016-025
https://doi.org/10.7746/jkros.2018.13.1.016 ISSN: 1975-6291 / eISSN: 2287-3961

RNN을 이용한 Expressive Talking Head from Speech의 합성

Synthesis of Expressive Talking Heads from Speech with Recurrent Neural Network

사쿠라이 류헤이[1] · 심바 타이키[2] · 야마조에 히로타케[3] · 이 주 호[†]
Ryuhei Sakurai[1], Taiki Shimba[2], Hirotake Yamazoe[3], Joo-Ho Lee[†]

Abstract: The talking head (TH) indicates an utterance face animation generated based on text and voice input. In this paper, we propose the generation method of TH with facial expression and intonation by speech input only. The problem of generating TH from speech can be regarded as a regression problem from the acoustic feature sequence to the facial code sequence which is a low dimensional vector representation that can efficiently encode and decode a face image. This regression was modeled by bidirectional RNN and trained by using SAVEE database of the front utterance face animation database as training data. The proposed method is able to generate TH with facial expression and intonation TH by using acoustic features such as MFCC, dynamic elements of MFCC, energy, and F0. According to the experiments, the configuration of the BLSTM layer of the first and second layers of bidirectional RNN was able to predict the face code best. For the evaluation, a questionnaire survey was conducted for 62 persons who watched TH animations, generated by the proposed method and the previous method. As a result, 77% of the respondents answered that the proposed method generated TH, which matches well with the speech.

Keywords: Talking heads, Recurrent neural network, Acoustic features, Facial features

1. Introduction

When people are learning a foreign language by using a computer, to increase the understanding, utterance face animation is utilized often. It makes better results than using only voice in information presentation for users. For example, in [1], in language learning using a computer, support of understanding is provided by presenting not only texts and voices but also images of faces at the same time. In [2], speech face animation for improving the user's understanding was provided when reading out example sentences in the dictionary site. Generating a speech face animation based on text and speech input is called the *generation of a Talking Head* (TH), and we refer to the generated face animation as *TH*.

In order to generate TH, first, speech is synthesized from texts and the synthesized speech is converted into phonemes. In general, a facial code is predicted by a prediction model that predicts a facial code based on phoneme inputs[2,3].

The speech contains, not only symbolic elements such as vowels and consonants, but also non-symbolic elements such as strength and inflection. For that reason, it is necessary to generate not only the shape of the mouth and the timing of transition, both of which are faithful to the uttered voice, but also TH which adequately reproduces strength and intonation. For example, when angry voice is given as input and the system generates TH with neutral expression or smiling, that TH cannot be said to be consistent with speech. If the speech is further strengthened at some point in the uttered voice, TH should also emphasize facial changes to match it. Furthermore, for the two speech signals, if the utterance contents are the same as the text, but the way of

Received : Jan. 23. 2018; Revised : Feb. 8. 2018; Accepted : Feb. 13. 2018
1. Ritsumeikan University, Shiga, Japan (sakurai@aislab.org)
2. Ritsumeikan University, Shiga, Japan
3. Ritsumeikan University, Shiga, Japan
† Corresponding author: Ritsumeikan University, Shiga, Japan (leejooho@is.ritsumei.ac.jp)

speaking is different, the generated TH should also be different. However, with the conventional technology, it is difficult to generate a TH that matches such utterance voice automatically. In the TH generation method based on phoneme, since phoneme does not include information of strength and intonation, only TH with fixed expression can be generated. In the method of generating TH with facial expression by giving the component of emotion as a parameter, it will be labor-intensive because it is necessary to adjust the parameters manually.

Considering the mechanism behind utterance, potential speech intention drives the muscles related to articulation, which include facial movement, then vocal sound is produced. Therefore, generating TH from voice can be considered as an inverse problem. From this point of view, raw audio preserves more information about facial movement than phoneme or text. It suggests that potentially we can recover facial information from voice. Therefore, in this research, our aim is to automatically generate TH from only speech as input. Furthermore, generated TH reflects the expression of emotions and intonation in speech as facial expressions.

1.1 Related Researches

Research on TH has been done for a long time and various approaches have been proposed.

Hidden Markov Model (HMM) is often used for generation of TH. In the method of Wang et al., TH generation is realized by learning the phoneme sequence and the trajectory of mouth shape using HMM[3]. Since the image obtained as an output is a monochrome image with low resolution, matching between the output image and the high-resolution color image is performed to achieve the corresponding high-resolution image as a realistic output image.

Fan et al. generated TH using a recurrent neural network (RNN) instead of HMM for shape prediction of the face, and gave better results than a method using HMM[2]. RNN has taken the spotlight in recent years as a method that can efficiently learn series signals[4]. For example, it is used for series signals such as machine translation, speech recognition, and video analysis. RNN has a structure in which the hidden layer of the feed-forward neural network (FFNN) has links in the time series direction, so that time series information can be learned. However, neither method trains the relation between phoneme and facial code, so it can only generate TH with fixed facial expression without intonation. In addition, phonemes differ depending on the language, so those TH depends on the language. Moreover, since it cannot be applied to sounds like laughter which cannot be converted to phonemes, there is a problem in using only phonemes as input for generating TH.

Many types of research have been done to add facial expressions to TH in order to realize more natural TH generation. In the sample-based method by Cosatto et. al., a facial part is divided into a forehead, eyebrows, eyes, a jaw, and a mouth, and the shape of each part is adjusted and combined to generate TH with an expression[5]. This is realized by determining the parameters of each part such as the opening width of the mouth and the position of the lips. In [5], the importance of facial expressions during utterance is mentioned, and in addition to the mouth in the example above, by adjusting parameters of other parts such as eyebrows, eyes, chin, and forehead, various expressions can be achieved. However, their method is limited to randomly adding facial expressions to the face generated from phonemes. Furthermore, in order to determine the parameters of each part, facial parts such as eyes and mouth should be detected by image processing. Since for the data set, data such as the size of mouth opening should be collected, the process is weak against noise and detection error. It is a disadvantage that data collection is difficult in this method.

Wan et al. succeeded in adding arbitrary facial expressions by learning the generation of TH based on text input and the expression of TH independently[6]. By adjusting parameters manually, facial expressions of six kinds of emotions such as neutral, kindness, anger, joy, fear, and sadness can be added correspondingly to TH. However, with this method, since it is necessary to adjust the expression parameters manually corresponding to each emotion at the time of generation, it is not possible to generate varying emotion during an utterance.

These TH generation methods cannot generate expressive facial shapes without manual parameters setting. Moreover, intonation cannot be expressed. This is because these methods use phonemes as inputs and information on facial expressions is not used. When speech signals are converted to phonemes, information on expressions and intonation will be lost. For this reason, in these studies, they can only generate TH with a fixed facial expressions unless information on facial expressions is not given manually. In addition, since phonemes are dependent on languages, it is necessary to redesign them all when it is applied to other languages. Even in the same language, facial expressions

based on voice, from which phonemes cannot be extracted, such as screams, cannot be reproduced.

[21] attempted to generate an animation of the adjacent area of lip from the voice signal directly. However, it does not generate TH of the whole face, and the evaluation of the generated image is only quantitative reconstruction error.

Similar to the proposed method in this paper, [22] attempts to generate a direct TH from speech using a neural network. However, this does not focus on emotions and intonation included in speech but reports evaluation results of phoneme recognition by human subjects experiment.

1.2 Research Aim

In this research, instead of converting speech signals into phonemes, we propose a method of converting speech to low-level acoustic features instead and generating TH with facial expression and intonation. By using low level acoustic features including information on facial expressions and intonation, it is possible to generate TH with facial expressions and intonation.

The main contribution of this research is to be able to generate TH with facial expression and intonation. Since the low-level acoustic features of speech do not depend on languages and phonemes, there is no need to adjust for each language, and there is an advantage that TH can be generated from special speech such as laughter which cannot be converted to phonemes. In addition, since it is not necessary to manually label low-level acoustic features and facial codes, it is easy to collect large amounts of data sets. Examples of applications of the proposed method are as follows.

• Generating natural facial expression for Service Robots

If a service robot has a function of displaying facial expressions, the proposed method will be a user-friendly useful interface for generating natural facial expressions. An operator or developer of the service robot does not need to consider facial expression but just input voice with emotions.

• Communication on the VR space

When communicating in the VR space in a virtual reality (VR) game or the like, it is possible to generate TH in real time from the sound input from the microphone of the headset.

• Creating 3D animation

Since facial codes including expressions and intonation can be predicted based on speech input, generation of 3D animation becomes easy.

• Monitoring in a camera-free environment

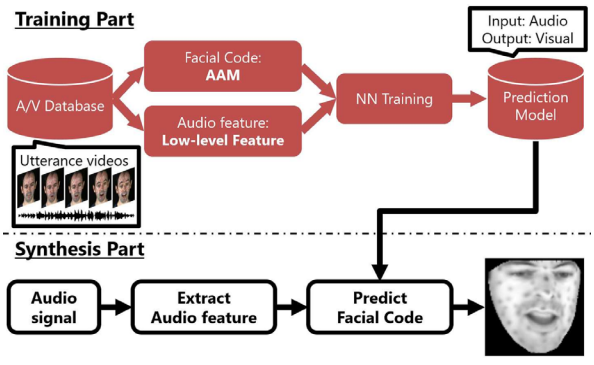
Even in places where cameras cannot be used in consideration of privacy, it is possible to predict facial expressions from voice and thus indirect monitoring becomes possible. Similarly, although the camera cannot acquire facial expression when the object is not in the view angle of cameras, the proposed method is able to predict facial expression if the sound can be monitored.

2. Proposed method

[Fig. 1] shows the overview of our proposed method. The method consists of a training part and a synthesis part. In the training part, we construct a prediction model that is used to generate THs. In the synthesis part, we generate THs that correspond to given audio signals of speech.

The prediction model is a regressor, which takes as input a sequence of low-level audio features extracted from raw speech signal then outputs a sequence of facial codes corresponding to each time step respectively. In this research, we employ recurrent neural network (RNN) to construct the prediction model.

In the training part, we train the RNN regressor by supervised learning with a dataset of a lot of pairs of parallel sequences of audio features (i.e. inputs) and facial codes (i.e. target labels). Note that in our method the labeled dataset is relatively easily constructed at low costs. It is because of facial codes are automatically obtained from facial images. Thus, we can collect labeled data semi-automatically if we have a speech video which

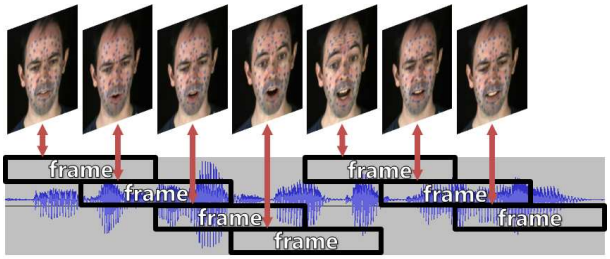


[Fig. 1] Overview of proposed method

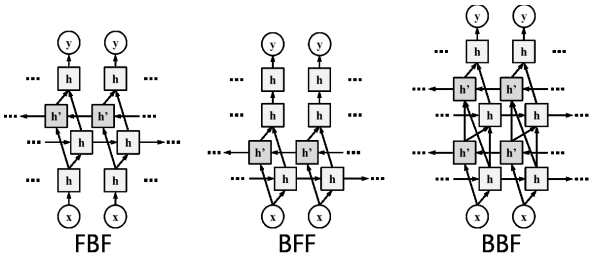
is temporally synchronized with speech audio (i.e. ordinary frontal view speech movies). In this way, we do not need to manually annotate labels except for cases of failed transform of facial codes.

In the synthesis part, we first transform an input raw audio signal to a sequence of audio features. Then the audio feature sequence is input to the appropriately trained prediction model to obtain a sequence of facial codes and finally, a video of facial animation is synthesized by decoding each facial code frame-by-frame.

Since speech movies are sequential signal, we employ bi-directional RNN (BRNN) for prediction model, which is widely used for sequence modeling[7]. However, it is known that naive RNNs or BRNNs cannot learn temporal relationship farther than about 10 steps due to vanishing gradient problem[8]. Thus, in particular, we use bi-directional long short-term memory (BLSTM), which can learn longer dependencies[9].



[Fig. 2] Separate utterance videos into frames



[Fig. 3] Three types of networks used in this research

2.1 Dataset

We create datasets with labels for training and evaluation based on the Surrey Audio-Visual Expressed Emotion database (SAVEE)[10], which is a dataset of movies (i.e. videos with audio) of frontal face speaker. SAVEE contains movies of English speech as 60 fps video and 44.1 kHz audio, by 4 native English male speakers. For each speaker, 120 sentences with 7 types of emotion are recorded. Lengths of the movies are about 3 seconds. The 7 emotions are neutral, anger, disgust, fear, happiness, sadness and surprise. There are 15 sentences for each emotion except for neutral that are 30 sentences. In this research, we used 120 movies of 1 speaker out of 4, who is identified as DC in the database.

Note that SAVEE provides blue markers painted on speaker's face for ease of extraction of facial landmark points. However, we do not use them because there are too few markers around the mouth to sufficiently describe its deformation. In particular, there are no markers on the inner or outer boundary of the lip, which are crucial for our modeling. Instead, we use a set of 68 landmarks defined in Multi-PIE face database[11].

In order to use the dataset as input and output for the network, we need to preprocess the dataset. As the preprocess for an audio waveform, we extract short segments of waveform at the same time intervals of video frames (i.e. 60 fps) by sliding window as shown in [Fig. 2] We refer to the segments as audio frames. The window size is 33.3 ms and the stride is 16.6 ms. Note that the window size of windows is determined as to extract low revel audio feature in the range between 20 ms and 40 ms, which is according to human characteristic of auditory perception. In the SAVEE database, 120 movies of subject DC are converted to 27765 pairs of image and audio frame in total. We describe feature extraction for these frames in section 3.

2.2 Sequential regression by RNN

Speech movie is a pair of sequential frames of audio and video that are tied and temporally synchronized. Thus, the relationship between audio features and facial codes can be modeled as regression, which is the main idea of our proposed method.

Speech movie has a nature that the mouth moves prior to the occurrence of the vocal sound. In addition, at the end of the utterance, the mouth is moving to close after the stop of the voice. Therefore, audio features and facial codes in speech movie have dependencies not only on the forward direction of time but also on the reverse direction of time. In order to predict facial codes from audio features, the information in both forward and reverse time directions should be captured. Thus, we employ BLSTM as the regression model whose input is audio feature and output is facial code. By training the regression model, we predict facial codes only from audio signals and then synthesize THs.

In this research, we train and compare 3 variants of 3 layers network. Each network consists of a composition of BLSTM

layer and feed-forward layer, which we refer to them as B and F. In particular, we use FBF, BFF, and BBF as shown in [Fig. 3]. We apply ReLU activation function after F layer except for the last F layer, since the output (i.e. facial code) is valued on real vector. The loss function is defined as the mean squared error between predicted outputs and labels. We minimize the loss by the stochastic gradient descent.

3. Feature representation and extraction

We extract sequences of audio features and facial codes from sequences of audio frames and images (i.e. movies) of the dataset as described in 2.1 in order to use them as input and output for networks. In this section, we consider the design of the low-level audio feature and the facial code, which represent expression and movement of mouth well.

3.1 Audio feature

Previous work uses phoneme as audio feature, which is converted from raw audio signal. However, using phoneme leads to discard some information about expression or accent. On the other hand, we employ other low-level audio feature which retains expressive information more than phoneme.

Emotion recognition from voice has a long history and there are comparison studies about audio features that are suited to capture expression or intonation[12,13]. These researches developed low-level audio features that well represent expression in voices. Note that, the information that represents expression or intonation, is information such as pitch or loudness of voice.

Mouth shape is mainly determined by vowels and several consonants. Therefore, audio feature for speech recognition, which must convey information about phonemes, can be considered as correlating to mouth shape. Mel frequency cepstral coefficients (MFCCs) are audio feature that approximates the characteristics of human auditory perception. In particular, MFCC itself and its dynamics (ΔMFCC, Δ^2MFCC) are classical features that are universally used for speech recognition[14]. In addition, F0 and energy are the most important features for emotion recognition[12]. F0 represents the perceptual pitch of the voice and energy represents the strength of the voice. Therefore, in this research, we employ MFCC, energy, their dynamics, and F0 as the feature that captures mouth movement and expression.

We extract the feature from all audio frames in the dataset. We use Hidden Markov Model Toolkit (HTK)[15] to extract MFCC, energy and their dynamics, and we use SPTK[16] to extract F0. More precisely, we extract MFCC with 0.97 of pre-emphasis coefficient and 20 channels of Mel filter bank, and we use 12 lower coefficients as MFCC feature. As the result, we obtain 40 dimensional vector of audio feature by concatenating them as $(\mathbf{c}; e; \Delta(\mathbf{c}); \Delta(e); \Delta^2(\mathbf{c}); \Delta^2(e); z)$, where $\mathbf{c} \in \mathbb{R}^{12}$ is MFCC, $e \in \mathbb{R}$ is energy, $z \in \mathbb{R}$ is F0 and $\Delta(\cdot)$ is an operator that compute regression coefficients.

3.2 Facial code

Directly predicting facial images from audio features by a prediction model is difficult since raw facial images are very high dimensional signal. Thus, we have to describe images by low dimensional codes that can be approximately reconstructed to original images. Active appearance model (AAM) is a well-established technique that can efficiently encode and decode a face image[17]. AAM is a parametric model of deformable objects in a 2D image, which is described by shape parameters and appearance parameters. AAM can represent facial images of a variety of individuals or expressions by shape and appearance parameters, which are weights of the principal component analysis (PCA). In this research, we employ shape and appearance parameters as facial code. In particular, let $\mathbf{p} \in \mathbb{R}^8$ be the shape parameters and lambda in $\boldsymbol{\lambda} \in \mathbb{R}^8$ be the appearance parameters of AMM, then we define our facial code as their concatenation $(\mathbf{p}; \boldsymbol{\lambda})$.

In order to encode images in the dataset into facial codes, we first construct AAM then encode the videos into the sequences of facial codes by the AAM. Since AAM uses iterative registration and requires a good initial crop of image that is close to the location of the face. Therefore, for the first image in a video, we apply face detection to obtain the location of the face. For the subsequent images the location of the face estimated in the previous frame is used as the initial value.

Since AAM is based on PCA, it requires a lot of manually annotated images as training data to construct a model that appropriately represent a variety of faces. As described in 2.1, we manually annotate facial key points in images in the dataset for training AAM. Lucas-Kanade algorithm[19] is used for fitting. For implementation, we used Menpo library[18].

4. Experiments

As we described in previous sections, synthesizing TH from low-level audio feature makes the synthesized TH expressive and accented.

In the experiments, we evaluate the performance of prediction models and quality of synthesized THs. We trained 3 variants of network as described in section 2.2 with the dataset of audio features and facial codes, which is created from SAVEE database. Then we evaluate which network is appropriate for our prediction model by quantitatively comparing the predicted facial codes. Furthermore, we synthesized THs by decoding facial codes predicted by the best prediction model. We quantitatively evaluate the THs through the subjective test by human.

The code for the experiments is available from here https://github.com/f2um2326/talking_heads.

4.1 Details of experimental setup

4.1.1 Dataset

120 speech movies were split into training set, validation set and test set, which consist of 108, 6 and 6 movies respectively. In order to train the AAM, we created an image dataset by manually annotating the 68 facial key points as described in [11] to 5400 images that were chosen at random from the training set. Then all the images in the videos are encoded to facial codes. Note that the cumulative percentage of the variance by the principal components is 98.0% for the shape and 79.9% for the appearance.

4.1.2 Dataset

To the best of our knowledge, there is no directly comparable method. However, networks that take the only phoneme as input without referring information about expression could not learn prediction model of facial codes. Therefore, we created a model that is based on the previous work that uses the only phoneme[2] with labels of expression as additional input. We use that model for comparison. In particular, the input feature vector consists of triphone (i.e. the concatenation of one-hot vectors of phoneme) and the state of phoneme such as:

$$(\underbrace{0,0,...,1}_{I},\underbrace{0,1,...,0}_{I},\underbrace{1,...,0,0}_{I},\underbrace{0,1,0}_{3}), \quad (1)$$

where $I=46$ is the number of phonemes, first $3I$ components correspond to the predecessor, center, successor phoneme respectively. The last 3 components are also one-hot representation that is set to indicate what position the current frame is in a sequence, i.e., first, last or other than those.

However, since SAVEE contains sets of speeches that are an identical sentence but with different emotion, they are indistinguishable by using Eq. (1) as input feature.

Therefore, we incorporate the information of J types of emotion by appending J dimensional one-hot vector indicating the emotion of the movie to the feature (1):

$$(\underbrace{0,0,...,1}_{I},\underbrace{0,1,...,0}_{I},\underbrace{1,...,0,0}_{I},\underbrace{0,...,1,0}_{J},\underbrace{0,1,0}_{3}), \quad (2)$$

where J=7 is the number of emotions. Although there must be uncertainty on phoneme extraction or emotion estimation in practice, in this experiment we use ground truth of phoneme and emotion for comparative method. Note that it is equivalent to we assume phoneme and emotion were perfectly predictable.

4.2 Details of networks and training

We set the equal number of units for each layer in a network. In particular, for the proposed method, the number of units is 312 for FBF, 418 for BFF and 262 for BBF. The number of units are adjusted to match the number of parameters (i.e. about 600,000) among networks. For the comparative methods, the number of units is 624 for FBF, 806 for BFF and 516 for BBF. The number of parameters is about 2,450,000. We applied Dropout[23] with 0.5 of the drop rate to first and second layers. We used RMSprop[20] for optimization with hyperparameter $\gamma=0.99$, $\epsilon=1e-8$ and the learning rate is 0.1 for the proposed method and 0.001 for the comparative method.

4.3 Quantitative evaluation of the prediction models

We evaluate the performance of prediction of facial code by comparing the proposed method that is based on low-level audio feature, and the comparative method that is based on phoneme and emotion label as described in the 4.1.2. For quantitative evaluation, we define a performance metric CORR, which is an average over correlations between sequences of predicted facial codes and ground truth. It is more interpretable than mean

squared error. More precisely, CORR is defined as:

$$CORR = \frac{\sum_{n=1}^{N_{video}} \sum_{d=1}^{D_n} corr(\hat{v}_d^n, v_d^n)}{\sum_{n=1}^{N_{video}} D_n} \quad (3)$$

Where N_{video} is the number of videos, D_n is a number of frames of the n-th video, $\hat{v}_d^n$ and v_d^n are prediction and ground truth of facial codes of d-th frame in the n-th video, $corr(\cdot,\cdot)$ is the normalized correlation. Note that $CORR$ indicates the fitness of predicted facial codes to the ground truth and higher the better, especially maximum and minimum of $CORR$ are 1 and 0. [Table 1] shows the result of quantitative evaluation. The best network in the proposed method was BBF, which $CORR$ was 0.38857. The best network in the comparative method was also BBF, which $CORR$ was 0.17053. The result suggests the proposed method outperformed the comparative method.

4.4 Qualitative evaluation for the synthesized talking heads

We quantitatively evaluated synthesized THs, which are sequences of decoded images from facial codes predicted from audio features by the best prediction models of both proposed method and comparative method. We conducted subjective evaluation test by asking human subjects whether the vocal expression (i.e. emotion or accent) contained in emotional speech is reflected to the synthesized TH as facial expression. We investigated by questionnaire for 62 human subjects. 15 subjects are Europeans and North Americans, and 47 subjects are Asian.

Examples of synthesized THs are shown in [Fig. 4] and [Fig. 5].

The questionnaire items and results are as described below. Note that the subjects do not know which THs come from the proposed method or the comparative method through experiment.

(ⅰ) "Watch a TH without voice (i.e. video only), the answer which emotion does the TH express". See [Fig. 6] for the result.

[Table 1] CORR value of predicted result

	Proposed method	Phoneme-based method
FBF	0.37770	0.14549
BFF	0.38393	0.15364
BBF	**0.38857**	**0.17053**

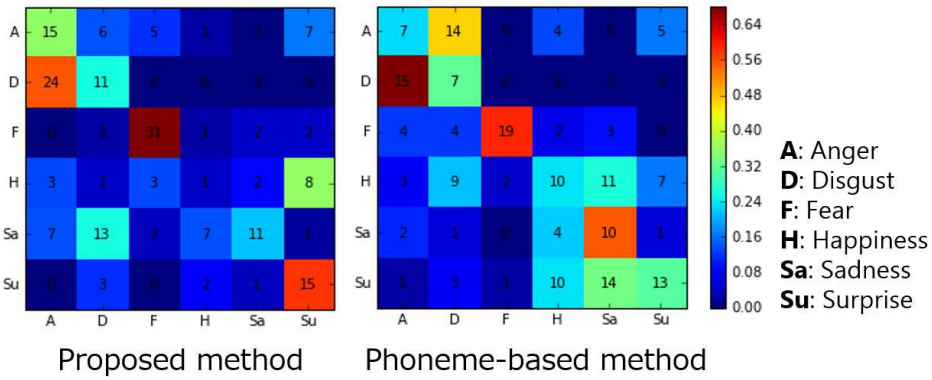
(ⅱ) "Watch two THs with voice (synthesize by proposed and comparative method), then answer which TH is more correctly reproducing the vocal expression". The result is that 77% of subjects answered the proposed method was better than the comparison method.

(ⅲ) "Watch two THs with voice (synthesize by proposed and comparative method), then answer the degree of compatibility between video and voice on a five-point scale". See [Fig. 7] for the result.

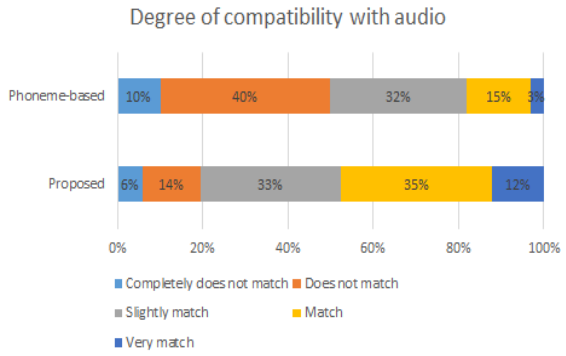
(ⅳ) "Watch two THs with voice (synthesize by proposed and comparative method), then answer the naturalness on a five-point scale". See [Fig. 8] for the result.

[Fig. 4] TH images of every twentieth frame. Top images are original, middle images are proposed, and bottom images are phoneme-based TH. The proposed method successfully reproduced the mouth movements such as opening widely or puckering his mouth, which exist in the original video

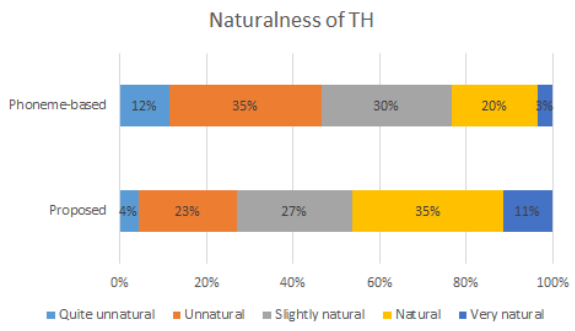
[Fig. 5] TH images of every fifth frames. Top images are original, middle images are proposed, and bottom images are phoneme-based TH. The proposed method produces more dynamic expression than another



[Fig. 6] Emotion estimation result



[Fig. 7] Degree of compatibility of TH with audio (Five levels evaluation)



[Fig. 8] Degree of naturalness of TH (Five levels evaluation)

4.5 Discussion

For quantitative evaluation, BBF performed the best result both in the proposed and comparative method. This result suggests that applying recurrent connection in early layer performs better than in late layer, and multiple recurrent layers also increase prediction accuracy.

For THs synthesized by both methods, we observed that the mouth was moving according to the speech sentence. However, THs of the comparative method lacked dynamics such as opening the mouth wider according to vocal accent. On the other hand, THs of the proposed method successfully reproduced accent or small movement of the mouth.

For qualitative evaluation, human subjects more correctly answered true emotion from TH without voice for the proposed method than the comparative method, as shown in [Fig. 6]. However, the accuracy is indeed not high in both methods. One of the causes of the low accuracy comes from the difficulty of the problem itself. Since facial expression of anger and disgust, or fear and surprise are similar to each other, distinguishing them without voice is difficult even for experienced people.

On the other hand, in terms of compatibility and naturalness for TH with voice, the proposed method achieved better scores.

From these results, we observe that how much vocal accent is appropriately reflected in synthesized TH is important for watcher to make an impression of naturalness. Because in comparative method, true information about emotion is additionally given but information about accent was not used.

In the case treating 7 classes of emotion as categorical information as our proposed or comparative methods, it is feasible to synthesize expressive THs by manually adjusting parameters of emotion. However, it is difficult by hand to adjust appropriate non-symbolic parameters such as accent or intonation. Since our method can incorporate such information in a natural way, THs that further match speech voice can be synthesized.

5. Conclusion

In this paper, we proposed the method to synthesize TH with

expression and accent only from speech voice as input. The proposed method used MFCC, energy, their dynamics and F0 as audio feature, and shape and appearance parameters of AAM as facial code. We constructed a prediction model from a sequence of audio features to a sequence of facial codes using BLSTM as regressor. In qualitative evaluation by human subjects, 77% of subjects answered that THs synthesized by our method are more expressive than comparative method.

References

[1] D.W. Massaro, "Symbiotic value of an embodied agent in language learning," *37th Hawaii International Conference on System Sciences*, Big Island, HI, USA, 2004, doi: 10.1109/HICSS.2004.1265333.

[2] B. Fan, L. Wang, F.K. Soong, and L. Xie, "Photo-real talking head with deep bidirectional LSTM", *International Conference on Acoustics, Speech, and Signal Processing*, Brisbane, QLD, Australia, 2015, doi: 10.1109/ICASSP.2015.7178899.

[3] L. Wang, and F.K. Soong, "HMM trajectory-guided sample selection for photo-realistic talking head," *Multimedia Tools and Applications*, vol. 74, no. 22, pp. 9849–9869, Nov.,|2014.

[4] A. Karpathy, J. Johnson, and L. Fei-Fei, "Visualizing and understanding recurrent networks", arXiv:1506.02078, 2015.

[5] E. Cosatto, and H.P. Graf, "Sample-based synthesis of photo realistic talking heads," *Computer Animation 98*, Philadelphia, PA, USA, USA, pp. 103-110, 1998.

[6] V. Wan, R. Blokland, N. Braunschweiler, L. Chen, B. Kolluru, J. Latorre, R. Maia, B. Stenger, K. Yanagisawa, Y. Stylianou, M. Akamine, M.J.F. Gales, and R. Cipolla, "Photo-Realistic Expressive Text to Talking Head Synthesis," 14th Annual Conference of the International Speech Communication Association, Lyon, France, pp. 2667-2669, 2013.

[7] M. Schuster, and K.K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, Nov., 1997.

[8] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber, "Gradient flow in recurrent nets: the difficulty of learning long-term dependencies," *A field guide to dynamical recurrent neural networks*, IEEE Press, 2001.

[9] A. Graves, M. Liwicki, S. Fernandez, R. Bertolami, H. Bunke, and J. Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, May 2009.

[10] H. Sanaul and P.J.B. Jackson, "Multimodal Emotion Recognition," W. Wang ed., *Machine Audition: Principles, Algorithms and Systems*, Hershey, PA: IGI Global, 2011, pp. 398-423, doi: 10.4018/978-1-61520-919-4.ch017.

[11] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image and Vision Computing*, vol. 28, no. 5, pp. 807-813, May, 2010.

[12] O.-W. Kwon, K. Chan, J. Hao, and T.-W. Lee, "Emotion Recognition by speech signals," *8th European Conference on Speech Communication and Technology*, Geneva, Switzerland, pp. 125-128, 2003.

[13] Y. Pan, P. Shen, and L. Shen, "Speech emotion recognition using support vector machine," *International Journal of Smart Home*, vol. 6, no. 2, pp. 101-107, April, 2012.

[14] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 1, pp. 52-59, Feb., 1986.

[15] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X.A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK book*, Microsoft Corporation, 1995.

[16] S. Imai, T. Kobayashi, K. Tokuda, T. Masuko, K. Koishida, S. Sako, and H. Zen, Speech signal processing toolkit (SPTK), [Online], http://sp-tk.sourceforge.net/, Accessed: Feb. 14, 2018.

[17] I. Matthews and S. Baker, "Active appearance models revisited," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 135-164, Nov., 2004.

[18] J. Alabort-i-Medina, E. Antonakos, J. Booth, P. Snape, and S. Zafeiriou, "Menpo: a comprehensive platform for parametric image alignment and visual deformable models," *22nd ACM international conference on Multimedia*, Orlando, Florida, USA, pp. 679-682, 2014.

[19] B.D. Lucas, and T. Kanade, "An iterative image registration technique with an application to stereo vision," *1981 DARPA Image Understanding Workshop*, pp. 121-130, April 1981.

[20] T. Tieleman, and G. Hinton, "Lecture 6.5-rmsprop: divide the gradient by a running average of its recent magnitude", *COURSERA: Neural Networks for Machine Learning*, vol. 4, pp. 26-30, 2012.

[21] W. Han, L. Wang, F. Soong, and B. Yuan, "Improved minimum converted trajectory error training for real-time speech-to-lips conversion," *2012 IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, 2012, doi: 10.1109/ICASSP.2012.6288921.

[22] D. W. Massaro, J. Beskow, M. M. Cohen, C. L. Fry, and T. Rodgriguez, "Picture My Voice: Audio to Visual Speech Synthesis using Artificial Neural Networks", *Auditory-Visual Speech Processing*, Santa Cruz, CA, USA, pp. 133-138, 1999.

[23] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research*, Vol. 15, pp. 1929-1958, Jun., 2014.

Ryuhei Sakurai

2004 Engineering, Ritsumeikan University. (BE)

2008 Engineering, Ritsumeikan University. (ME)

2012 ~ present Ritsumeikan University Assistant

Interests: Computer Vision, Machine Learning

Taiki Shimba

2013 Information Science and Engineering, Ritsumeikan Univ. (BE)

2015 Information Science and Engineering, Ritsumeikan Univ. (ME)

Interests: Computer Vision, Machine Learning, System Integration

Hirotake Yamazoe

2000 Engineering Science, Osaka University. (BE)

2002 Engineering Science, Osaka University. (ME)

2005 Engineering Science, Osaka University. (Ph.D)

2015 ~ present Lecturer, Ritsumeikan University

Interests: Computer Vision, Human Behavior Analysis, Wearable Computing

Joo-Ho Lee

1993 Electrical Engineering, Korea University. (BE)

1995 Electrical Engineering, Korea University. (ME)

1999 Electrical Engineering, University of Tokyo. (PhD)

2004 ~ present Ritsumeikan University Professor

Interests: Robotics, Computer Vision, Machine Learning, Intelligent Space