Speech-Driven Robot Face Action Generation with Deep Generative Model for Social Robots

Chuang Yu[1(✉)], Heng Zhang[2], Zhegong Shangguan[2], Xiaoxuan Hei[2], Angelo Cangelosi[1], and Adriana Tapus[2]

[1] Cognitive Robotics Laboratory, University of Manchester, Manchester, UK
{chuang.yu,angelo.cangelosi}@manchester.ac.uk

[2] Autonomous Systems and Robotics Lab/U2IS, ENSTA Paris, Institut Polytechnique de Paris, Paris, France
{heng.zhang,zhegong.shangguan,xiaoxuan.hei, adriana.tapus}@ensta-paris.fr

Abstract. The natural co-speech facial action as a kind of non-verbal behavior plays an essential role in human communication, which also leads to a natural and friendly human-robot interaction. However, a lot of previous works for robot speech-based behaviour generation are rule-based or handcrafted methods, which are time-consuming and with limited synchronization levels between the speech and the facial action. Based on the Generative Adversarial Networks (GAN) model, this paper developed an effective speech-driven facial action synthesizer, i.e., given an acoustic speech, a synchronous and realistic 3D facial action sequence is generated. In addition, a mapping between the 3D human facial action to the real robot facial action that regulates Zeno robot facial expressions is also completed. The evaluation results show the model has potential for natural human-robot interaction.

Keywords: Social robot · Face action · Human-robot interaction

1 Introduction

Multimodal behavior understanding and generation play an important role in successful human-robot interaction [1–3]. Recently, verbal and non-verbal behavior generation has drawn more and more attention of researchers from many research areas [4], including computing animation and robotics [5–7]. Non-verbal behaviors, including gaze, gestures [8,9], and facial actions [10], can assist verbal expressions in conveying clearer meanings in contrast to speech-only communication and intention. It also can help build trust during real or virtual communication [11]. The co-speech facial action as a non-verbal behavior plays a significant role in human-human communication as they can express rich meanings including the emotional information in the whole facial expression and the verbal content information in the lip or mouth action [12]. In order to make a natural and friendly human-robot interaction, it is necessary to endow a social robot with synchronous

F. Cavallo et al. (Eds.): ICSR 2022, LNAI 13817, pp. 61–74, 2022.
https://doi.org/10.1007/978-3-031-24667-8_6

and realistic facial actions. However, it is very challenging to generate aligned facial actions mapping with speech in long-term human-robot interaction. Most of the previous researches used the handcrafted or rule-based approaches [13] for offline facial action generation based on speech. These methods are time-consuming and have a limited continuity level of successive facial actions.

With the development of deep learning technology, more and more generative models for the time series generation are developed, such as, autoencoder model, seq2seq model [4], the model with the normalizing flows, and GAN (Generative Adversarial Networks) model [14]. Researchers have explored many areas with a generative model for time series generation, for example, human social trajectories generation [15], and gesture generation [5]. These methods also can be used for expressive co-speech facial action generation and simplify the robot facial action generation process as a kind of cross-modal mapping task. The trained generative model for facial action synthesis can be used in real-time and long-term human-robot interaction for a social robot.

In this paper, we built up a temporal GAN framework for a cross-modal mapping task, which can be applied to generate realistic facial action aligned with the speech audio in a real-time human-robot interaction. The basic GAN model is tough to train for cross-modal mapping tasks. Speech-to-facial action generation is not a strict mapping task where humans can conduct many possible natural and simultaneous facial action modes for the same speech, for example, in different emotional states, which makes the GAN more challenging to train towards convergence. To tackle this problem, we built our temporal GAN architecture based on $WGAN-GP$ (Wasserstein Generative Adversarial Networks-Gradient Penalty) [16] model and introduced the L_1 loss in the generator loss function inspired by the $pix2pix$ model [17].

The human face has more than 40 muscles, controlled for facial actions while speaking. However, it is challenging to equip the face of a robot with such a significant number of actuators in order to express rich facial actions. Our research used the Zeno robot, a small humanoid with an expressive face for human-robot interaction. In this paper, we completed the facial action retargeting task from 3D human facial landmarks to the robot facial action with related motor control signals. The pipeline of robot facial action generation is as shown in Fig. 1.

In summary, our contributions in this paper are as follows:

- A temporal GAN architecture with L_1 reconstruction loss was proposed to effectively generate a 3D co-speech facial action sequence, which can be used in long-term human-robot interaction.
- The facial action retargeting task was performed from human 3D face action to the robot face actuators.
- The generated robot facial actions with the related speech were applied to the Zeno robot for human-robot interaction.

The rest of the paper is structured as follows: Sect. 2 describes the related works. Section 3 shows the methodology. Section 4 describes the dataset and related pre-processing operation. Section 5 presents our experiments and results. The conclusions and future work are resumed in Sect. 6.

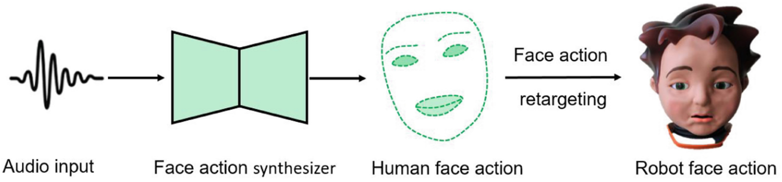

Fig. 1. The pipeline of RS^3 architecture. RS^3 architecture contains a facial action synthesizer from speech and a human-to-robot face action mapping. The facial action synthesizer based on the temporal GAN model takes the acoustic speech as input and outputs the aligned human 3D facial actions. The robot facial actions with control signals of robot facial motors are obtained from the human 3D face action during the facial action retargeting part. These motor control signals can be applied to the Zeno robot face during human-robot interaction.

2 Related Work

2.1 Generative Model

Generative models, including the model based on Naive Bayes [14], Variational Autoencoder (VAE) [14], Generative Adversarial Networks (GAN) [14], and the model based on the normalizing flows technology [18], have been of high interest to researchers on the image generation tasks and time-series data generation tasks. Habibie et al. [7] proposed a recurrent variational autoencoder model to produce human motions given some control signals, which can be applied for the sequence prediction task. In the paper [17], Isola et al. built an image-to-image translation network based on the conditional GAN model for image generation where the generation loss function also took the L_1 distance into consideration in order to obtain better generation results and to simplify the training process of GAN model. Heter et al. [19] came up with a probabilistic and controllable model for motion synthesis using normalising flows technology. The generative architecture as a probabilistic model can achieve a one-to-many mapping given multiple control signals, namely style-controllable generation.

2.2 Facial Image or Animation Generation

Co-speech video or animation generation with facial action is not a new research topic, which has been explored for decades [20,21]. Vougioukas et al. [6] built a temporal GAN model for speech-driven face animation generation. The GAN model used one static image and one speech audio as input and outputted realistic aligned image sequences with the face. In order to improve the randomness of the generated face image sequence, the generative architecture contained a noise generator to produce the noise time series, which was added, respectively to the representation information of each overlapped audio clip in the face generator model of GAN. The generator loss function also considered the L_1 reconstruction loss except for the basic GAN loss, which can improve the generation results.

In the paper [22], Zhou et al. built up an LSTM-based expressive face animation generation model with a self-attention encoder. The model took one unseen speech audio and one static speaker image as inputs. With the help of the disentangled learning skills in the model, the model can achieve the disentanglement of content and style in audio. The model can generate different talking animations with the same speaker style as the one in the input of the static image. Namely, it is speaker-aware speaking head animation generation.

Both methods above produce the co-speech face or head image sequence. There are some other researchers who focus on the face key point (landmark) position generation used to control the virtual face avatar in a simulation environment. Sadoughi et al. [23] proposed a conditional sequential GAN (CSG) model to generate the talking lip actions. The model used the spectral and emotional speech features as conditional input of the generator of GAN to synthesize emotion-aware lips action with key point coordinates, which then were utilized to regulate the virtual face. Abdelazi et al. [24] described a new co-speech facial movement generation structure that can be exploited for the animated face on smart mobile phones. The model jointly used audiovisual information, including the speech audio and one static face image as input to synthesize the aligned 3D facial action. However, these facial action generation models were only applied in animation situations and still do not explore whether it is effective and whether it can make a difference in the real humanoid robot with face-actuated skin.

2.3 Robot Facial Action Generation from Speech

Multimodal robot behaviors with speech, co-speech gestures, and facial actions are essential in a natural and friendly human-robot interaction. Particularly, the co-speech facial action generation is an active research area as facial actions convey more emotional information and speech content information than gestures. Aly et al. [12] built up a multimodal robot behavior synthesis system used on an expressive robot ALICE, in order to imitate natural multimodal human-human interaction. The system can generate speech-related gestures and co-speech facial expressions, which led to an effective narrative human-robot interaction. However, the robot facial action generation method is rule-based and cannot generate natural co-speech facial action. The generated facial action sequences have a limited continuity level in the temporal domain. In the paper [25], the laughter-driven facial motions were generated for a female android robot with facial skin. However, the robot facial action generation was rule-based with limited facial action patterns. It is challenging to generate facial actions in real-time and long-term human-robot interaction.

3 Methodology

3.1 Problem Formulation

Speech-Driven Facial Action Generation: It is a crossmodal translation task with time series both as input and output. Given one speech audio $S^m =$

$[s_t^m]_{t=1:T}$ as input, the model attempts to produce one 3D facial action sequence $A^m = [A_t^m]_{t=1:T'}$. Namely, the generative model tries to learn a relation function $F_{mapping}$ to maximize the conditional probability $P(A^m|S^m)$ to generate the natural and aligned facial action sequence. Here, T and $T^{'}$ are the time steps of the speech audio as input and the facial action sequence as output, respectively, and they are different from each other because the digital speech audio and the facial action sequence have different sampling rates. m in the model means m^{th} mapping task.

$$\mathbf{A^m} = F_{map}\left(\mathbf{S^m}\right) \tag{1}$$

Facial Action Retargeting: The problem is to map the human facial actions A^m with the 3D positions of the face key points to the robot facial action sequence $C^m = [c_t^m]_{t=1:T'}$ with the face motors' control signals. The mapping task for face action retargeting consists of getting a function that finds the relation between human facial action and robot facial action. The final function can make the appearance of the human facial and the appearance of robot face as similar as possible at each time step.

$$\mathbf{C}^m = F_{retarget}\left(\mathbf{A}^m\right) \tag{2}$$

3.2 Facial Action Synthesizer from Speech

This section describes our novel proposed facial action synthesizer from speech with temporal GAN. The speech-to-face-action GAN ($S2FGAN$) architecture is shown in Fig. 2. $S2FGAN$ model is made of a generator and a discriminator. The generator with a sequence model takes the temporal representation of speech audio as input and outputs the mapping gesture. The discriminator is employed to differentiate whether the speech and the facial action match each other.

The generator comprises two layers of GRU (Gated Recurrent Unit) and MLP (Multilayer Perceptron) layer. Firstly, Mel-frequency Cepstral Coefficients (MFCCs) as audio representation are extracted from the overlapped audio clips. The MFCC feature sequence is input into the batch normalization layer following two layers GRU of the generator. The following MLP layer takes the latent representation of the former GRU layer to generate the synchronous 3D facial action sequence mapping with the speech audio. And each frame of the facial action sequence contains 3D positions of 68 face landmarks. The discriminator works to distinguish whether the facial action sequence and the speech audio match with each other. The audio clip representations and the facial action sequence are input into two MLP layers and decoded to 100-dimensional features and 50-dimensional features each time step, respectively. The following concatenation layer fusions the two modal features in each time step, whose output is input to a GRU layer. In the final GRU cell, an MLP layer is followed to classify whether the speech audio matches the 3D facial action sequence.

The loss function of our $S2FGAN$ model comprises two parts, namely L_1 loss part and the standard conditional GAN loss part, actually the Wasserstein

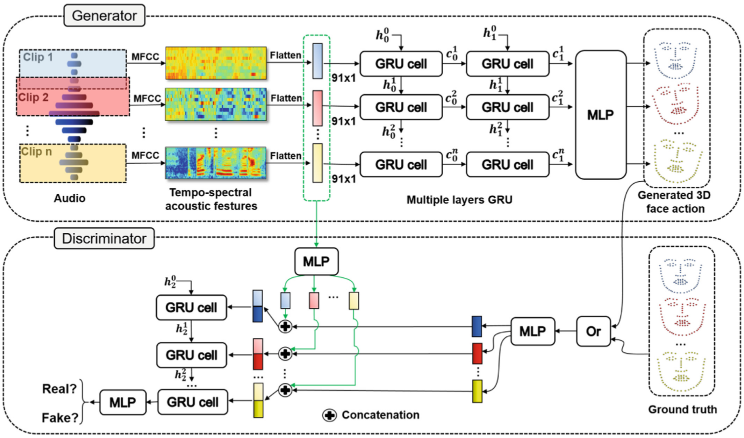

Fig. 2. The *S2FGAN* architecture. The model has a generator and a discriminator. The generator takes the spectral features of speech as input and outputs the synchronous 3D facial action data. The discriminator with the speech audio and generated/real facial action sequence as inputs try to classify whether the speech and the facial action sequence align in the temporal domain.

loss and the gradient penalty used in the WGAN-GP model. The basic GAN model often experiences the training instability problem. The Wasserstein GAN (WGAN) model makes a more stable training than basic GANs [26]. WGAN can also produce samples with low quality and suffer from convergence problems during the training process. WGAN introduces a weight clipping skill to enforce a Lipschitz constraint on the discriminator (namely, the critic named in WGAN) to address these problems, which also can result in gradient explosion/vanishing without careful tuning of the weight clipping parameter. In WGAN-GP [16], the authors proposed an alternative skill to the weight clipping, namely, adding the gradient penalty to the discriminator loss, which leads to a more stable training process. The WGAN-GP loss contains the generator loss $\mathcal{L}_G$ and the discriminator loss $\mathcal{L}_D$, as shown in Eq. 3 and Eq. 4, respectively. Where, the sample S^n from the sampling uniformly along straight lines between pairs of points sampled from the data distribution of real facial action sequences and the generator distribution.

$$\mathcal{L}_G = -\mathbb{E}_{S^m}[D(S^m, G(S^m))] \tag{3}$$

$$\begin{aligned}\mathcal{L}_D =& \mathbb{E}_{S^m}[D(S^m, G(S^m))] - \mathbb{E}_{S^m,A^m}[D(S^m, A^m)] + \\ &\lambda \mathbb{E}_{A^n}[(\|\nabla_{A^n} D(A^n)\|_2 - 1)^2]\end{aligned} \tag{4}$$

Inspired by the *pix2pix* model [17], we introduced a L_1 reconstruction loss to improve the realistic co-speech facial action generation. The L_1 loss is pixel-wise in the image translation task with *pix2pix* model, while we used the frame-wise L_1 loss for the facial action sequence, as shown in the Eq. 5. The final discriminator loss keeps same, and the L_1 loss is added to the generator loss to get the final generator loss $\mathcal{L}_{G-all}$ as shown in Eq. 6. Where λ is an empirical hyperparameter during $S2FGAN$ model training, which is to balance how much contribution $\mathcal{L}_{L1}$ or $\mathcal{L}_G$ make for all the loss.

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{S^m,A^m}\left[\|A^m - G(S^m)\|_1\right] \tag{5}$$

$$\mathcal{L}_{G-all} = \mathcal{L}_G + \lambda\mathcal{L}_{L1}(G) \tag{6}$$

3.3 Facial Action Retargeting

Our facial action retargeting task is a mapping from human facial action to robot facial action. The mapping objective is to approximate the human face appearance with a limited number of robot face actuators. In this paper, we use a Zeno robot to present the synchronous generated facial action sequence with the speech audio. There are four motors for skin-based face appearance regulation. The four motors can control the eyebrows/forehead up or down (one motor), the eyelids open (one motor), the mouth open, and the left and right corner for the smile (one motor). Each motor's control signal of the Zeno robot is a continuous value ranging from 0 to 1. The retargeting process from human facial action to robot facial action is as shown in Fig. 3. The human facial action includes 3D positions of 68 landmarks, and four motors of Zeno with skin regulate the robot's facial expression.

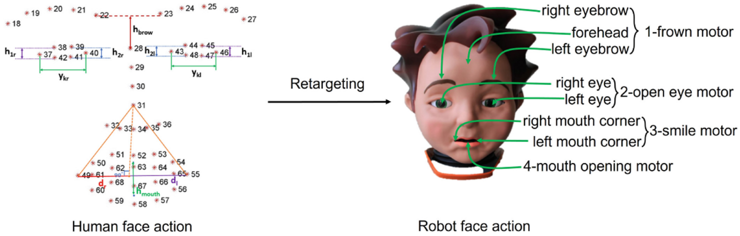

Fig. 3. Facial action retargeting overview. The human face contains 68 human landmarks in each frame. The robot face has four motors controlling the eye, the forehead, the mouth, and the mouth corners for a smile.

We name the distance between the 38th landmark and the 42nd landmark as $\mathrm{h_{1r}}$, the 39th landmark and the 41st landmark as $\mathrm{h_{2r}}$. The right eye wide $\mathrm{y_{kr}}$ is the distance between the 37th landmark and the 40th landmark, which is used to normalize the open degree of the eye as different persons have the different eye sizes. Apply the same rule for the left eye to get the $\mathrm{h_{1l}}$, $\mathrm{h_{2l}}$ and $\mathrm{y_{kl}}$. Because Zeno has only one motor to control two eyelids, we calculate the average for the two eyes. Then, we can get the scale for the eyelid motor as shown in Eq. 7.

$$S_{eye} = \frac{\frac{h_{1l}+h_{2l}}{2y_{kl}} + \frac{h_{1}r+h_{2r}}{2y_{kr}}}{2} \tag{7}$$

To obtain the eyebrows motor scale, we need to calculate the distance between the midpoint of the 22$^{\mathrm{nd}}$ landmark and the 23$^{\mathrm{rd}}$ landmark and the 28$^{\mathrm{th}}$ landmark, $\mathrm{h}_{\mathrm{brow}}$. So, the scale for eyebrows is shown in Eq. 8. Here, we divide $(\mathrm{y}_{\mathrm{kl}} + \mathrm{y}_{\mathrm{kr}})/2$ is to reduce the influence of different face sizes of people on the results of the mapping task from 3D facial action to robot motor action.

$$S_{\mathrm{brow}} = \frac{h_{\mathrm{brow}}}{\left(\frac{y_{kl}+y_{kr}}{2}\right)} \tag{8}$$

The scale for mouth motor can be obtained as shown in Eq. 9. Here, $\mathrm{h}_{\mathrm{mouth}}$ is the distance between the midpoint of the 52$^{\mathrm{nd}}$ landmark and the 63$^{\mathrm{rd}}$ landmark and the midpoint of the 67$^{\mathrm{th}}$ landmark and the 58$^{\mathrm{th}}$ landmark.

$$S_{\mathrm{mouth}} = \frac{h_{\mathrm{mouth}}}{\left(\frac{y_{kl}+y_{kr}}{2}\right)} \tag{9}$$

The scales of the smile motor controlling the left and right corners of the mouth can be obtain from Eq. 10, Eq. 11, and Eq. 12. The d_1 is the distance between the 55$^{\mathrm{th}}$ landmark and the foot of the perpendicular through the 31$^{\mathrm{st}}$ landmark. Similarly, the d_r is the distance between the 49$^{\mathrm{th}}$ landmark and the foot of the perpendicular through the 31$^{\mathrm{st}}$ landmark. Furthermore, d_l and d_r can be calculated based on the law of cosines. Because there is only one smile motor to control mouth corners, the mean value of S_{smile_l} and S_{smile_r} is used to get the scale of the smile motor, namely S_{smile}.

$$S_{smile_l} = \frac{d_l}{\left(\frac{y_{kl}+y_{kr}}{2}\right)} \tag{10}$$

$$S_{smile_r} = \frac{d_r}{\left(\frac{y_{kl}+y_{kr}}{2}\right)} \tag{11}$$

$$S_{smile} = \frac{S_{smile_l} + S_{smile_r}}{2} \tag{12}$$

Since robot motor control signals in the Zeno system range from 0 to 1, normalization operation for the scales should be done as shown in Eq. 13. That is to say, find the maximum and minimum of every scale which are applied to get the final control signal of face motors. Where, $\mathrm{s} \in \{S_{eye}, S_{\mathrm{brow}}, S_{\mathrm{mouth}}, S_{smile_r}, S_{smile_l}\}$.

$$\mathrm{norm}(s) = \frac{s - s_{\min}}{s_{\max} - s_{\min}} \tag{13}$$

4 Dataset and Preprocessing

4.1 Dataset

In this paper, we used the open database-Biwi 3D Audiovisual Corpus of Affective Communication dataset [27], which was developed at ETH Zurich. The corpus contains 1109 sentences uttered by 14 native English speakers, including six males and eight females, aged between 21 and 53 (average age of 33.5). A real-time 3D scanner and a professional microphone were utilized to obtain the speakers' facial action and synchronous speech audio during the data recording process. The dense dynamic face scans were obtained with a sampling rate of 25 frames per second. Moreover, the RMS error in the 3D reconstruction is about 0.5 mm, which is good enough for our facial action generation task. For the dataset development, the participants imitate the forty short English sentences extracted from film clips. For each sentence, the subject should speak two times, one with a neutral state and one with an emotional state, which is the same as the film clip's emotion. In this paper, the speech audio contains intrinsically emotional information, so we did not take the emotion label into consideration for the co-speech facial action generation task.

4.2 Pre-processing

The pre-processing step includes speech audio spectral feature extraction with MFCC [28], face landmarks extraction from 3D face images in the database with Dlib [29] library, and how to align the speech and facial action in the temporal domain and so on.

Alignment Between Speech and Facial Action Sequence. In this paper, we used the same time size for each speech audio to simplify the training process. From the distribution of audio length, we know that most audios are longer than 2.5 s. There are 1096 files in our database, of which 1095 are longer than 1 s, 1072 are longer than 2 s, 926 are longer than 3 s, and 645 are longer than 4 s. We chose 3 s as the time size for $S2FGAN$ training to use as many samples as possible. The audios longer than 3 s were cut into 3 s, and the samples with audio size shorter than 3 s were deleted from the database. Meanwhile, the sampling rate of the face image is 25 fps. In addition, we also deleted some samples whose audio size was more than 3 s but the face images less than 3 s. Finally, we got 788 samples from the original dataset for $S2FGAN$ training.

Because the speech audio and facial action series have different sampling rates, 44100 Hz for audio and 25 fps for facial action, the whole speech audio was divided into audio clips to align the facial action and audio in the temporal domain. Namely, one frame corresponds to 1764 audio frames. Considering the facial action time series's temporal dependence, we used the overlapped audio clips with 3528 audio frames centered on the related facial action frame, and the stride of the overlapped audio clips was 1764.

Speech Audio Feature Extraction. MFCC (Mel-Frequency Cepstral Coefficients) is often used for acoustic speech representation in speech recognition and other related speech audio tasks. MFCC feature of speech audio is the one in the frequency domain using the Mel scale based on the human ear scale. MFCCs, as frequency domain features, are much more accurate than time-domain features in the recognition task [30]. So, we used the MFCC as the overlapped audio clips from the whole speech audio in this paper. Each audio clip corresponding to one face frame extracted an MFCC feature with size of $7 * 13$.

3D Face Landmarks Detection. To get the 3D face landmarks, we first got the 2D face landmarks from the 2D face image frame by frame using the Dlib library. The pre-trained face landmark detector inside the Dlib library can extract the location of the 68 face landmarks (x, y)-coordinates that map to face structures on the face. The indexes of the 68 coordinates can be seen in Fig. 3. In our case, the triangle mesh texture recorded in the database is the corresponding RGB file. Firstly, we have detected 2D face landmarks located in the RGB face image. The detected landmark location is the same as the landmark location in the texture image. The relation between texture image and 3D mesh can be learned from the depth image. Then from the 2D position, we can directly extract the 3D positions of the 68 landmarks.

5 Experiments and Results

5.1 S2FGAN Training

The conditional GAN *pix2pix* model with L_1 loss explored multiple cross-modal translation tasks with the small dataset with 400 images or less, and it got the receivable testing results finally [17]. Like the *pix2pix* model, our speech-face database for $S2FGAN$ training contains 788 samples (600 samples for training, 90 samples for validation, and 98 for testing) with the speech audios and the 3D facial action sequences. During the training, the batch size was 30, and the time steps of 3D facial action were 75 as the audio time size was fixed to 3 s during the $S2FGAN$ training. The Standardization operation was employed on the 3D facial action data before inputting the $S2FGAN$ model, and the batch normalization procedure was applied in the generator of $S2FGAN$, which both can effectively reduce the overfitting problem during model training based on the tuning experiments. We did not use the batch normalization layer in the discriminator because the layer can lead to a convergence failure during WGAN training [16]. Both generator and discriminator of $S2FGAN$ model used the Adam algorithm [31] for optimization during training with the learning rate = 0.0002, the parameter $\beta_1 = 0.5$, $\beta_2 = 0.999$, and $\epsilon = 10^{-7}$. Moreover, the dropout setting of GRU is 0.1. The number of discriminator iterations per generator iteration is five during training. The model is developed with Tensorflow 2.3, and the training with 10000 epochs was done on an NVIDIA Quadro P1000 GPU for about four days.

5.2 Results and Evaluation

During the testing part, the speech-driven 3D facial action sequences were generated using the trained $S2FGAN$ model. Then, the generated facial actions were transferred to the control signals of the robot face motors, which finally were presented on the Zeno robot facial actions with the aligned speech audios.

Applying the generated co-speech robot facial action to the Zeno robot, we recorded some videos with the speech audio and the synchronous facial action, and some frames in a generated co-speech robot facial action sequence are as shown in Fig. 4.

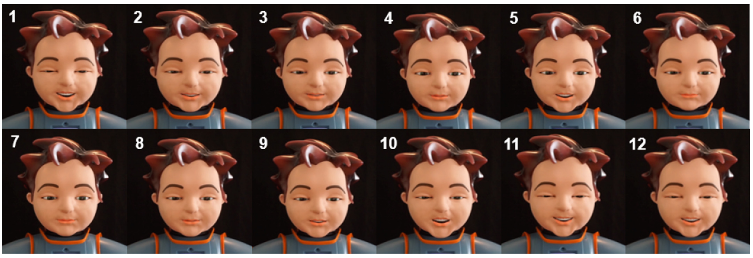

Fig. 4. The generated facial action on the robot Zeno in one example. Twelve frames were sampled with the same interval sampling from one sample of generated face action series from speech. The number in the figure is to show the order of the sampled frames.

From Fig. 4, we can see that the Zeno face driven by the generated robot facial action has noticeable movement in the mouth area and the eye area. The forehead area has limited change as the human forehead's noticeable movement often happens in intensely emotional expression instead of the common human co-speech facial actions. Besides, the subjects' forehead mostly remains still during the speech as present in the database.

Quantitative Evaluation. The quantitative evaluation of speech-driven facial action is challenging [32] as the mapping between speech and facial action (or gesture) sequence is a weaker correlation than the image-to-image translation in *pix2pix*, which is a rigid one-to-one mapping. In this paper, we explored the quantitative evaluation for the generated speech-driven facial action sequence with an Average Position Error (APE) [32] as shown in Eq. 14, where T is the time steps of the robot facial action, equal to 75; S is the number of testing samples, equal to 98; $f_{real}(s,t)$ and $f_{generated}(s,t)$ are the real robot facial motor control signal action and the generated one of sample s at time step t, respectively.

$$APE = \frac{1}{S \times T} \sum_{s=1}^{S} \sum_{t=1}^{T} \left| f_{\text{real}}(s,t) - f_{\text{generated}}(s,t) \right| \tag{14}$$

The APE validation result is 0.409 for the eye-opening motor, 0.190 for the eyebrow motor, 0.187 for the mouth-opening motor, and 0.189 for the smile motor. The eye-opening motor has the biggest APE 0.409 because the degree of the eye blink has a weak correlation with speech, and the action is primarily random to human speech. Hence, the generated eye blink action has a slight fluctuation. For example, in some generated samples, the robot face keeps squinting because this one-to-many mapping from the speech to the eye blink makes the alignment model fit to the average value of eye blink (around zero) when the model cannot find the suitable mode. Other motor APEs perform better when the related face actions strongly relate to the speech. The result looks like it still has some space to improve. However, it is still receivable for this kind of one-to-many mapping task with weak correlation. In the future, the generated robot facial action should be applied to the real human-robot interaction where the participants are asked to validate whether the generated robot facial action is synchronous with the speech, whether the facial action is natural, and whether the speech-face interaction is better than the speech-only interaction.

6 Conclusions and Future Work

In this paper, we built an effective temporal GAN architecture, namely $S2FGAN$, with losses of WGAN-GP and L_1 loss for co-speech facial action generation, which is promising to be used for other cross-modal mapping tasks with time series as input and output. The trained $S2FGAN$ model can generate realistic and synchronous facial action sequence with speech audio. The facial action retargeting from human face landmarks to robot facial action was completed. The robot facial action series were presented on the real Zeno robot in human-robot interaction. Finally, the generated facial action series was assessed with the qualitative evaluation and the quantitative evaluation. In the future, we will do user experiments to explore the long-term human-robot interaction environment with the generated face action presented on the Zeno robot. In addition, we will take the emotion label into consideration to explore emotional facial action generation for the robot's face.

Acknowledgement. We thank Cognitive Robotics Lab, Department of Computer Science, the University of Manchester, and Autonomous Systems and Robotics Lab, U2IS, ENSTA Paris, Institute Polytechnic de Paris. And we thank the UKRI Node on Trust (EP/V026682/1) for support.

References

1. Yu, C., Tapus, A.: Interactive robot learning for multimodal emotion recognition. In: Salichs, M.A., et al. (eds.) ICSR 2019. LNCS (LNAI), vol. 11876, pp. 633–642. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-35888-4_59
2. Noda, K., Arie, H., Suga, Y., Ogata, T.: Multimodal integration learning of robot behavior using deep neural networks. Robot. Autonom. Syst. **62**(6), 721–736 (2014)

3. Yu, C., Tapus, A.: Multimodal emotion recognition with thermal and RGB-D cameras for human-robot interaction. In: Companion of the ACM/IEEE International Conference on Human-Robot Interaction, vol. 2020, pp. 532–534 (2020)
4. Yu, C., Changzeng, F., Chen, R., Tapus, A.: First attempt of gender-free speech style transfer for genderless robot. In ACM/IEEE International Conference on Human-Robot Interaction, vol. 2022, pp. 1110–1113 (2022)
5. Yoon, Y., et al.: Speech gesture generation from the trimodal context of text, audio, and speaker identity. ACM Trans. Graph. **39**, 6 (2020)
6. Vougioukas, K., Petridis, S., Pantic, M.: End-to-end speech-driven realistic facial animation with temporal gans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 37–40 (2019)
7. Habibie, I., Holden, D., Schwarz, J., Yearsley, J., Komura, T.: A recurrent variational autoencoder for human motion synthesis. In: 28th British Machine Vision Conference (2017)
8. Yu, C., Tapus, A.: Srg 3: Speech-driven robot gesture generation with GAN. In: 2020 16th International Conference on Control, Automation, Robotics and Vision (ICARCV), pp. 759–766. IEEE (2020)
9. Zhang, H., Yu, C., Tapus, A.: Why do you think this joke told by robot is funny? The humor style matters. In: 2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), pp. 572–577. IEEE (2022)
10. Yu, C.: Robot behavior generation and human behavior understanding in natural human-robot interaction. Ph.D. dissertation, Institut Polytechnique de Paris (2021)
11. Lee, J., Marsella, S.: Nonverbal behavior generator for embodied conversational agents. In: Gratch, J., Young, M., Aylett, R., Ballin, D., Olivier, P. (eds.) IVA 2006. LNCS (LNAI), vol. 4133, pp. 243–255. Springer, Heidelberg (2006). https://doi.org/10.1007/11821830_20
12. Aly, A., Tapus, A.: Multimodal adapted robot behavior synthesis within a narrative human-robot interaction. In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2986–2993. IEEE (2015)
13. Park, J.W., Lee, H.S., Chung, M.J.: Generation of realistic robot facial expressions for human robot interaction. J. Intell. Robot. Syst. **78**(3–4), 443–462 (2015)
14. Foster, D.: Generative deep learning: teaching machines to paint, write, compose, and play. O'Reilly Media (2019)
15. Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., Alahi, A.: Social gan: Socially acceptable trajectories with generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2255–2264 (2018)
16. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. In: Advances in Neural Information Processing Systems, pp. 5767–5777 (2017)
17. Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134 (2017)
18. Rezende, D., Mohamed, S.: Variational inference with normalizing flows. In: International Conference on Machine Learning, pp. 1530–1538 (2015)
19. Henter, G.E., Alexanderson, S., Beskow, J.: Moglow: Probabilistic and controllable motion synthesis using normalising flows. arXiv preprint arXiv:1905.06598 (2019)
20. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, pp. 187–194 (1999)

21. Egger, B., et al.: 3d morphable face models-past, present, and future. ACM Trans. Graph. **39**(5), 1–38 (2020)
22. Zhou, Y., Han, X., Shechtman, E., Echevarria, J., Kalogerakis, E., Li, D.: Makelttalk: Speaker-aware talking-head animation. ACM Trans. Graph. **39**(6), 1–15 (2020)
23. Sadoughi, N., Busso, C.: Speech-driven expressive talking lips with conditional sequential generative adversarial networks. IEEE Trans. Affect. Comput. (2019)
24. Hussen Abdelaziz, A., Theobald, B.-J., Dixon, P., Knothe, R., Apostoloff, N., Kajareker, S.: Modality dropout for improved performance-driven talking faces. In: Proceedings of the 2020 International Conference on Multimodal Interaction, pp. 378–386 (2020)
25. Ishi, C.T., Minato, T., Ishiguro, H.: Analysis and generation of laughter motions, and evaluation in an android robot. APSIPA Trans. Signal Inf. Process. **8** (2019)
26. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: Proceedings of the 34th International Conference on Machine Learning, vol. 70, pp. 214–223 (2017)
27. Fanelli, G., Gall, J., Romsdorfer, H., Weise, T., Van Gool, L.: A 3-d audio-visual corpus of affective communication. IEEE Trans. Multim. **12**(6), 591–598 (2010)
28. Sahidullah, M., Saha, G.: Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition. Speech Commun. **54**(4), 543–565 (2012)
29. King, D.E.: Dlib-ml: A machine learning toolkit. J. Mach. Learn. Res. **10**, 1755–1758 (2009)
30. Dave, N.: Feature extraction methods LPC, PLP and MFCC in speech recognition. Int. J. Adv. Res. Eng. Technol. **1**(6), 1–4 (2013)
31. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
32. Hasegawa, D., Kaneko, N., Shirakawa, S., Sakuta, H., Sumi, K.: Evaluation of speech-to-gesture generation using bi-directional LSTM network. In: Proceedings of the 18th International Conference on Intelligent Virtual Agents, pp. 79–86 (2018)