

A Realistic Face-to-Face Conversation System based on Deep Neural Networks

Ze Zhou Chen
CloudMinds

chenzezhou007@aliyun.com

Zhaoxiang Liu*
CloudMinds

robin.liu@cloudminds.com

Huan Hu
CloudMinds

hans.hu@cloudminds.com

Jinqiang Bai
Beihang University
baijinqiang@buaa.edu.cn

Shiguo Lian
CloudMinds
sg.lian@163.com

Fuyuan Shi
CloudMinds
fuyuan.shi@cloudminds.com

Kai Wang
CloudMinds
kai.wang@cloudminds.com

Abstract

To improve the experiences of face-to-face conversation with avatar, this paper presents a novel conversation system. It is composed of two sequence-to-sequence models respectively for listening and speaking and a Generative Adversarial Network (GAN) based realistic avatar synthesizer. The models exploit the facial action and head pose to learn natural human reactions. Based on the models' output, the synthesizer uses the Pixel2Pixel model to generate realistic facial images. To show the improvement of our system, we use a 3D model based avatar driving scheme as a reference. We train and evaluate our neural networks with the data from ESPN shows. Experimental results show that our conversation system can generate natural facial reactions and realistic facial images.

1. Introduction

Recently, virtual assistants have been playing a more and more important role in our daily life, such as the question-answering assistant Apple Siri and Amazon Alexa, which may be the most popular virtual assistants at now. However, they provide only verbal response, lacking of nonverbal feedback.

Alternatively, some 2D or 3D avatars have been introduced to virtual assistants [14, 23, 5], which can provide both verbal and nonverbal interaction experiences. Nevertheless, few of them can exhibit realistic nonverbal reactions while speaking, e.g., realistic facial expressions. What is more, no natural reactions are exhibited during the listening

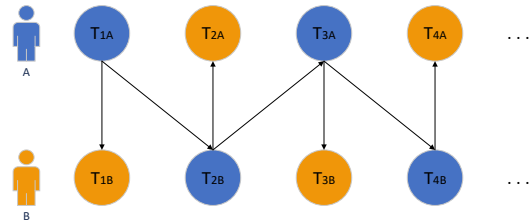


Figure 1. Regularity of human-human conversation. T_{iA} and T_{iB} ($i = 1, 2, 3, \dots$) denotes person A and B is speaking or listening in the i -th time period respectively. Blue blocks indicate that the person is a speaker, while orange blocks a listener. Arrows imply the dependence between different roles in different time periods.

phase when the avatars are talking with human[5]. Generally, they either keep still or randomly react within a series of predefined actions while listening to human. That is far from human's expectation on realistic and nature conversation.

We aim to fill the gap between present virtual assistants and the demand of realistic and natural interaction. Our work is based on the observations of conversational regularity illustrated in Figure 1. In human-human communications, the two roles, listener and speaker, are alternative between the two parties involved in the conversation. The speaker often makes verbal actions (speech) and nonverbal actions (head movements and facial expressions) simultaneously. While listening to the speaker, the listener receives verbal information as well as nonverbal cues and gives nonverbal feedback. Their roles exchange when the previous speaker stops talking. And the new speaker makes verbal and nonverbal responses accordingly. This procedure car-

*corresponding author

ries on repeatedly until the conversation ends.

Motivated by the great success of Sequence-to-Sequence (Seq2Seq) network [33] and Generative Adversarial Network (GAN) [10, 15, 43, 2, 38, 37], we propose a novel face-to-face conversation system which consists of seq2seq based listening and speaking models and a GAN based realistic avatar synthesizer. The listening and speaking models serve as generators of natural facial reactions and head pose. The realistic avatar synthesizer is used to produce a realistic portrait looks like a real human. Generally, the facial action and head pose to some extent imply the intentions and emotions of both speaker and listener. The listening model takes speaker’s facial action and head pose as input and generates facial action and head pose as listener’s nonverbal response. The speaking model takes the verbal response as input and generates facial action and head pose as speaker’s nonverbal accompaniment. The realistic avatar synthesizer takes the outputs of alternative listening model and speaking model as input and produces a realistic image sequence for avatar. Compared to traditional 3D data based realistic face synthesizer, our avatar synthesizer is free from the need for expensive 3D or motion capture data.

The rest of the paper is organized as follows. Some related works are reviewed in Section 2. In Section 3, the proposed conversation system is presented in detail. Experimental results and discussions are given in Section 4. Finally, Section 5 draws the conclusion.

2. Related works

Generally, a realistic face-to-face conversation system is able to understand the human’s speech and action, then decide how to respond in verbal and nonverbal manners, and finally drive the avatar. Although few mature end-to-end systems exist till now, there have been many works related to each of the steps.

To understand humans in interaction, some means have been proposed, including head pose estimation [29, 32, 16, 21], gesture recognition [36], gaze tracking [25], and multi-modal fusion [17]. Especially, analyzing facial actions to get emotion information is important for face-to-face conversation. Generally, face landmarks [8] are used for facial analysis, which tells the 2D or 3D positions of key points of the face. Alternatively, action unit has been widely used for face expression analysis or face avatar synthesis [19, 20, 34, 41, 42], which represents face components’ local movements. Different from face landmark, action unit contains only the relative motion of face components while no head pose or face size. Additionally, action unit, as a nonverbal feature, can also be used for verbal analysis, *e.g.*, lip reading [19]. It means that action unit contains both verbal and nonverbal information. Motivated by it, in this paper, we extract action unit and head pose from a face image as face features.

For verbal chatting, various conversation models have been widely used. For example, [35] proposed a translation model based on the encoder-decoder architecture. [33] presented a LSTM [13] based seq2seq structure to solve the training problem of long sequences. However, for nonverbal conversation or verbal-nonverbal combined conversion, there is few mature works. The latest one [35] adapted the seq2seq structure from verbal domain to nonverbal domain, and proved the seq2seq model works for nonverbal scenarios, although it considered only the speaking phase. Inspired by the work, we also use the seq2seq structure to design nonverbal models, including both the speaking model and listening model, which produces facial action units and head pose for avatar.

For avatar synthesis, the action unit based avatar driving technique [36, 5] has often been used, which activates the predefined 2D or 3D avatar by an action unit sequence. For example, the work in [5] also proposed a face-to-face conversation model which could drive a 3D avatar to answer human’s questions. However, avatars are either not realistic enough or difficult to obtain by such means as accurate 3D model scanning. What is more, no natural reactions are exhibited for the 3D avatar during the listening phase in [5]. Fortunately, due to the rapid development of Generative Adversarial Networks (GAN) in image generation [10], it is able to synthesize photo-realistic facial image without complex 3D models. For example, Pix2PixHD [38] and CycleGAN [43] offer appealing and succinct alternative for facial image synthesis. Recently, Chan *et al.* applied Pix2PixHD to learn a motion transformation between two video subjects [2]. Wang *et al.* presented the framework of Vid2Vid [37] to learn mappings between different videos. Motivated by these works [2, 38, 37, 5, 39, 22], we also adopt the GAN to generate the realistic avatar with facial action units and head pose as input.

3. The proposed system

3.1. System overview

In this paper, we propose a novel face-to-face conversation system. As shown in Figure 2, there are two phases in our system: one is for listening, and the other for speaking. When the virtual avatar is communicating with human, the speech audio and face images are input into our system. The face images are processed by face parsing module that produces face information including facial action units and face pose. This information is then passed into the Seq2Seq based listening model whose output is fed into the avatar synthesizer to produce realistic face images as nonverbal reactions when the virtual avatar is listening. The speech audio is transformed into text using the speech recognition algorithm proposed in [12, 11], and then passed into the conversation module for generating response sentence. The

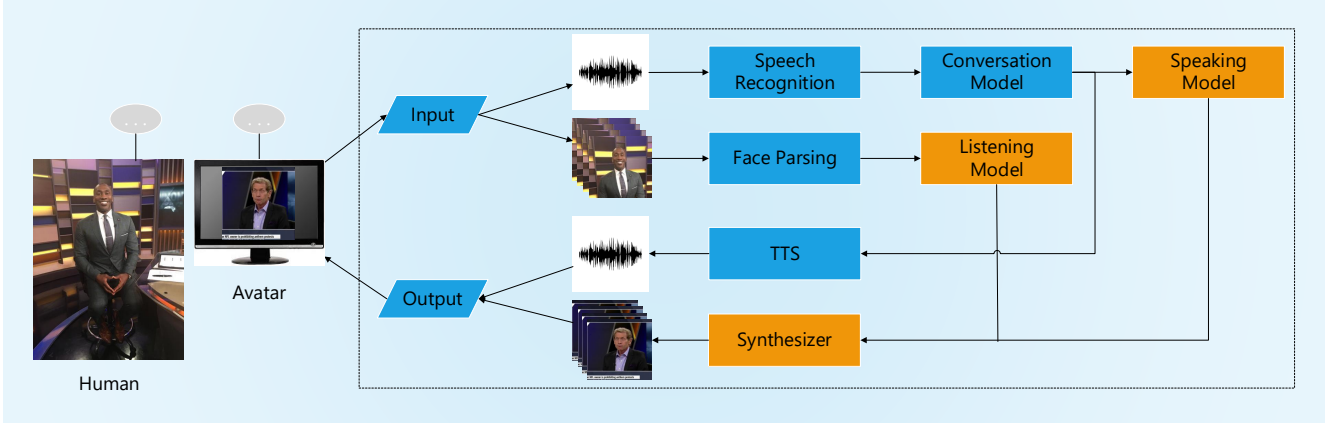


Figure 2. Overview of the proposed realistic face-to-face conversation system. Blocks colored orange indicate our contributions.

response sentence is passed to the Text-to-Speech (TTS) module proposed in [9, 40] to produce synthesized speech. At the same time, the response sentence is passed to the Seq2Seq based speaking model whose output is also input into the avatar synthesizer to produce realistic face images as the accompaniment of speech for the speaking phase. The avatar synthesizer serves listening and speaking alternatively during the whole conversation.

To train our system, we collect 700 videos of ESPN shows (First Take and Undisputed) from youtube that contain the scenes of face-to-face conversation.

3.2. Face parsing

To get the facial feature representation, we have several choices such as facial landmark (FL) [28], action unit [17, 31] and head-pose. FL carries detailed information on facial features and outlines. However, it has three disadvantages. First, FL lacks information on some areas such as forehead and cheeks. Second, FL cannot directly reflect the movement of facial muscles. Third, compared to the action unit, FL is more personalized and we want our high-level features to carry more generalized information. So we combine action-unit and head-pose (AU+POSE) to represent facial features. Compared with the combination of head-pose and text or the combination of text and action-unit, AU+POSE retains sufficient semantic information. In the listening phase, AU+POSE can maintain the smooth of dataflow in the deep network because no extra coding is needed, and this is conducive to the training of the entire system.

The Open-Face [1] can consistently predict the action units using Support Vector Machine (SVM) and Support Vector Regression (SVR) [3] with concatenated HoG feature [6] and facial shape, and accurately estimate head pose by solving PnP problem. Thus, we use it to extract AU+POSE from both the speaker and listener. The

AU+POSE is a 20-D vector which comprises of a 17-D AU and a 3-D head pose. On the acquisition of text data, the sentence is obtained directly from the video with python library AutoSub and TTS.

3.3. Listening model

The listening model takes the AU+POSE of speaker as input and outputs the AU+POSE of listener. We base our listening model on the Seq2Seq [33] architecture (as shown in Figure 3).

Listening encoder: The AU+POSE sequence extracted from the face sequence of speaker is employed as the input of the listening encoder. Let d_1, d_2, \dots, d_n represent the AU+POSE sequence extracted from the consecutive n face images, and h_1, h_2, \dots, h_n represent the hidden transitions of encoder for each input frame. The model is defined as

$$\begin{aligned} i^t &= \sigma(W_i[h^{t-1}, d^t] + b_i) \\ f^t &= \sigma(W_f[h^{t-1}, d^t] + b_f) \\ o^t &= \sigma(W_o[h^{t-1}, d^t] + b_o) \\ c^{-t} &= \text{relu6}(W_c[h^{t-1}, d^t] + b_c) \\ c^t &= f_t \odot c_{t-1} + i_t \odot c^{-t} \\ h_{enc}^t &= o_t \odot \text{relu6}(c_t) \end{aligned} \quad (1)$$

where σ denotes the sigmoid function, relu6 is the piecewise linear function, \odot denotes the element-wise multiplication, i_t, f_t, o_t represent input gate, forget gate, output gate of t -th LSTM unit respectively. c_t and h_t are the t -th cell state and hidden state. W and b are the trainable weights.

Listening decoder: The decoder model in the listening phase maintains the same n serialized outputs as the inputs. The output can be formulated as

$$h_{dec}^l = \text{LSTM}(h_{dec}^{l-1} | h_{enc}, y^{l-1}) \quad (2)$$

where h_{dec} is the hidden state of l -th decoder, and y^{l-1} is the output AU+POSE of the $(l-1)$ -th decoder.

3.4. Speaking model

The speaking model takes the response text of listener as input and outputs the AU+POSE as the accompaniment of speech. We also base our speaking model on the Seq2Seq [33] architecture (as shown in Figure 3).

Speaking encoder: There exist some popular word embedding methods, *e.g.*, Word2Vec [24], GloVe [26], ELMo [27] and BERT [7] to represent the text. Here, for simplicity, we employ a pre-trained Word2Vec model to embed each word into a 200-D vector. Our speaking encoder is defined as

$$h_{text}^n = LSTM(\{x_{text}^{n,l}\}_l) \quad (3)$$

where the LSTM [13] computes the forward sentence encodings, and applies a linear layer on top.

Speaking decoder: The speaking decoder outputs an AU+POSE sequence used to synthesize the face sequence for the speaker (*i.e.*, the previous listener).

3.5. Seq2Seq training

It is intuitive to use Mean Square Error (MSE) loss function to train our listening and speaking models, and the MSE loss can be formulated as

$$\begin{aligned} loss_{mse} = & \frac{1}{n_{au}} \sum_{i=1}^{n_{au}} \|au_{gt}^i - au_{pred}^i\|_2 + \\ & \frac{\gamma}{n_{pose}} \sum_{i=1}^{n_{pose}} \|pose_{gt}^i - pose_{pred}^i\|_2 \end{aligned} \quad (4)$$

where γ denotes the weight parameter, au_{gt}^i and au_{pred}^i denote the action-units vector in the ground truth and predicted sequence respectively, $pose_{gt}^i$ and $pose_{pred}^i$ denote the head-pose vector in the ground truth and predicted sequence. $\|\cdot\|_2$ denotes the Euclidean norm and n is the dimension of action-unit or head-pose.

However, the MSE loss cannot ensure the output continuity which is significant for generating realistic face sequences. To solve this problem, we introduce another loss: continuity loss to assist the Seq2Seq training. The continuity metric loss utilizes the constraints of adjacent frames which can considerably boost the performance of our models. The continuity metric loss can be formulated as:

$$\begin{aligned} loss(i, j) = & \left\| \left(y_{gt}^{i-j} - y_{gt}^{i-j-1} \right) - \left(y_{pred}^{i-j} - y_{pred}^{i-j-1} \right) \right\|_2 \\ loss_{con} = & \frac{1}{n_b} \sum_{i=n_b}^l \max(\{loss(i, j) | j = 1, \dots, n_b - 1\}) \end{aligned} \quad (5)$$

where y_{gt}^i and y_{pred}^i denote the AU+POSE vector in the ground truth and predicted sequence respectively, and n_b

denote the number of the adjacent frames used for calculating continuity metric loss. Thus, combining the MSE with continuity loss, we get the total loss as:

$$loss_{sum} = loss_{mse} + \alpha \cdot loss_{con} \quad (6)$$

where α is a weight used to adjust the two loss components.

Note that the head-pose and each component of the action-unit have different variation ranges. Hence, both the input and output of the listening model are normalized during training, and the output of the speaking model is also normalized during training.

Moreover, all frames in the video clips involving face-to-face conversation are used for training our listening and speaking models to get a smoother and more realistic face sequences.

3.6. Face synthesizer

With the AU+POSE as input, the face synthesizer aims to generate face images of the target person with corresponding postures and expressions. To accomplish this task, we adopt the state-of-the-art generative adversarial network Pix2PixHD [38] to complete this transfer. Specifically, we employed a combined loss widely used in generation tasks:

$$L = L_{GAN}(G, D) + L_{L1}(G(x), y) + L_{VGG}(G(x), y) \quad (7)$$

where $L_{GAN}(G, D)$ is the temporal smoothing adversarial loss presented in paper [2], which modified from Pix2PixHD [38]. When synthesizing the current face image $G(x_t)$, Generator G conditions on its corresponding current AU+POSE image x and the previously synthesized frame $G(x_{t-1})$ to obtain temporally smooth outputs. Discriminator D then attempts differentiate the “real” temporally coherent sequence $(x_{t-1}, x_t, y_{t-1}, y_t)$ from the “fake” sequence $(x_{t-1}, x_t, G(x_{t-1}), G(x_t))$:

$$\begin{aligned} L_{GAN}(G, D) = & E_{(x,y)} [\log D(x_{t-1}, x_t, y_{t-1}, y_t)] + \\ & E_x [\log (1 - D(x_{t-1}, x_t, G(x_{t-1}), G(x_t)))] \end{aligned} \quad (8)$$

where L_{L1} is an $L1$ reconstruction loss that measures the pixel-level deviation between the synthesized image $G(x)$ and ground truth y . L_{VGG} is the perceptual reconstruction loss which measures the $L2$ distance between relu2.2 and relu3.3 features of a pre-trained VGGNet [30].

The full transfer system is shown in the Figure 4. We use corresponding $(x_{t-1}, x_t, y_{t-1}, y_t)$ pairs to learn a mapping G which synthesizes images of the target person given AU+POSE. It is noteworthy that we do not directly feed the AU+POSE feature values into the network. Instead, we fill the feature values into the center of an empty image x , where x has the same size with the ground truth y . This allows our network to have spatial coordinate constraints when training, which can accelerate the convergence.

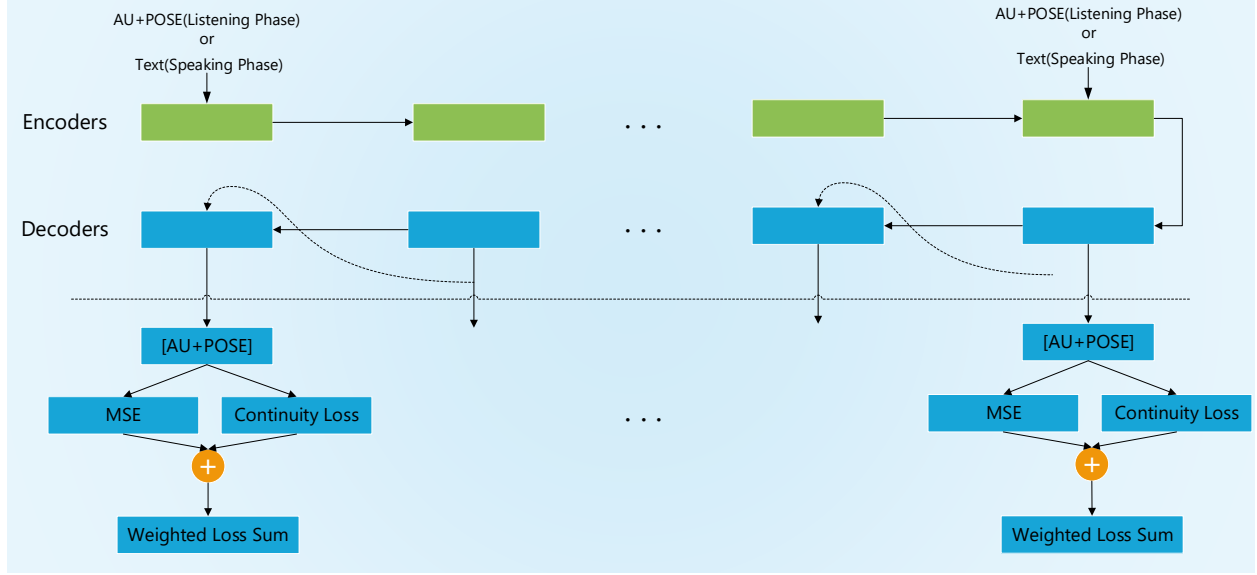


Figure 3. Architecture of listening model and speaking model. In the listening phase, the encoder takes the AU+POSE as input. And in speaking phase, the encoder takes the text as input. In both phases, decoder outputs AU+POSE. MSE and Continuity Losses are employed to supervise the training for both phases.

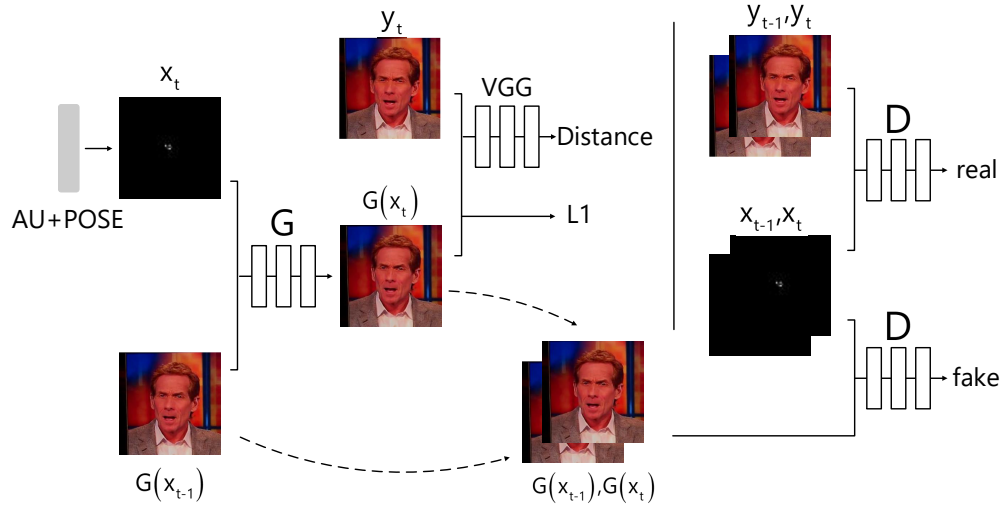


Figure 4. Architecture of the face synthesizer. We first map AU+POSE to image x . In the training phase, we use pair $(x_t, G(x_{t-1}))$ to learn the mapping G to generate $G(x_t)$. Discriminator D then attempts to differentiate the “real” temporally coherent sequence $(x_{t-1}, x_t, y_{t-1}, y_t)$ from the “fake” sequence $(x_{t-1}, x_t, G(x_{t-1}), G(x_t))$.

4. Experimental results

4.1. Implementation details

We implement our seq2seq-based listening and speaking models by using TensorFlow framework [4]. In listening phase, we construct a 4-layer encoder-decoder model using LSTM as shown in Figure 3. We substantiate all these

LSTMs with a 1000-d LSTM cell on top of a 1000-d embedding layer, followed by a linear layer with piecewise-linear relu6 to compute the final encoding. At each stage, both the decoder and the encoder have 10 cells cascaded, that is, corresponding consecutive 10 frames of data as input. The learning rate is set to 0.0001, the batch size is 16, and the ADAM [18] is used as the optimizer. The training process

stops after 80,000 iterations.

In speaking phase, the model uses only two layers of 1000-dimensional LSTM cell nested structure with relu6 on the activation function. The learning rate, batch size and optimizer selection are consistent with the listening phase, the γ in Eq. 4 is set to 8.1, the α in Eq. 6 is set to 0.1, and the training stops after 40,000 iterations. At each stage, the encoder has 25 cells cascaded and the decoder has 370 cells cascaded. We first pre-train the encoder and decoder on a small subset of our training set, after that, the encoder is frozen and the decoder is trained. After these tasks are completed, we finally fine-tune the whole encoder-decoder network.

We train our face generation network for 100 epochs using the ADAM optimizer [18] with lr=0.0002.

4.2. Evaluation on Seq2Seq models

We use MSE as well as cosine similarity in Eq. 9 to evaluate the performance of the listening model and the speaking model separately on the test set. Correspondingly, test data are divided into two parts. In the listening test set, each sample is comprised of a sequence of AU+POSE of the speaker as input, and a sequence of AU+POSE of the listener as ground truth. In the speaking test set, the input is just a sequence of words spoken by the speaker, and the ground truth is a sequence of the speaker’s AU+POSE. We compare our approach to the baseline: The classic Seq2Seq [33] method that uses MSE loss only.

$$\begin{aligned} d_{mse} &= \frac{1}{n} \sum_{i=1}^n l_{mse}^i \\ l_{mse} &= \frac{1}{n_{au}} \sum_{i=1}^{n_{au}} \|au_{gt}^i - au_{pred}^i\|_2 + \\ &\quad \frac{1}{n_{pose}} \sum_{i=1}^{n_{pose}} \|pose_{gt}^i - pose_{pred}^i\|_2 \\ d_{con} &= \frac{1}{n} \sum_{i=1}^n \frac{y_{gt}^i \cdot y_{pred}^i}{\|y_{gt}^i\| \cdot \|y_{pred}^i\|} \end{aligned} \quad (9)$$

Method	Metric	
	MSE	Cosine
Baseline	0.0565	0.983
Ours	0.0540	0.992

Table 1. Evaluation of training methods in listening phase.

Method	Metric	
	MSE	Cosine
Baseline	0.195	0.9635
Ours	0.141	0.980

Table 2. Evaluation of training methods in speaking phase.

In the listening phase as shown in Table 1, our method performs a little better than the baseline method in both

MSE metric and cosine similarity metric. In the speaking phase as shown in Table 2, our method also outperforms the baseline method. It is noteworthy that the gap between our method and baseline in speaking phase is much larger than that in listening phase. One reason is that, when human is speaking, the diversity of facial action units and head poses is much more than when human is listening. So the variance of training data in speaking phase is much larger than that in listening phase. Consequently, the speaking model benefits more than the listening model from our proposed metric learning. The other reason is that, compared to the speaking model, the listening model is much deeper. Hence, the listening model has stronger learning ability which makes the gap between the baseline and our method in listening phase smaller than that in speaking phase.

4.3. Evaluation on avatar synthesizer

We run facial action units and head pose extractor on the output of our avatar synthesizer, and compare these reconstructed action units and poses to the counterpart of the original input video. To assess the performance of our model, we employ the average error defined by

$$\begin{aligned} d_{au} &= \frac{1}{m \cdot n_{au}} \sum_{j=1}^m \sum_{i=1}^{n_{au}} \|au_{gt}^{ij} - au_{reconstructed}^{ij}\|_1 \\ d_{pose} &= \frac{1}{m \cdot n_{pose}} \sum_{j=1}^m \sum_{i=1}^{n_{pose}} \|pose_{gt}^{ij} - pose_{reconstructed}^{ij}\|_1 \end{aligned} \quad (10)$$

where au_{gt} , $au_{reconstructed}$, $pose_{gt}$, $pose_{reconstructed}$ denote the ground truth of action unit, the reconstructed action unit, ground truth of pose, the reconstructed pose and n denotes the number of frames in test set.

Phase	Metric	
	AU	Pose
Listening	0.046	0.038
Speaking	0.105	0.073

Table 3. Evaluation of avatar synthesizer.

As shown in Table 3, in both the listening and speaking phases, the difference between the synthesized images and the original images is minor in terms of action unit and pose.

4.4. Evaluation on End-to-End system

Here we employ a 3D model based avatar directly driven by facial action unit and head pose as comparison. Four examples (two for listening and another two for speaking) are illustrated in Figure 5 and Figure 6 to show the comparison between ground truth and predictions of our models, as well as the results of 3D avatar and our synthesizer. As can be seen, the facial images generated by our synthesizer are more similar to the ground-truth, and more realistic than the 3D model based avatar.



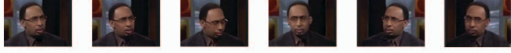
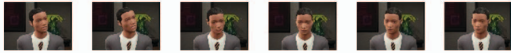
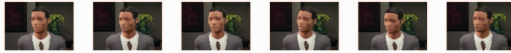


Listening Phase				
Scene				
Speaker-Video & Test				
Listener-Ground Truth				
Listener-Synthesis On Avatar				
Listener-Synthesized Images				

Figure 5. Experimental results of our models in listening phase. Two scenes are shown here, each of which includes the conversation scene (the first row), the speaker’s images (second row), the listener’s ground truth images (third row), the listener’s 3d avatar (fourth row), and the listener’s synthesized images (last row).





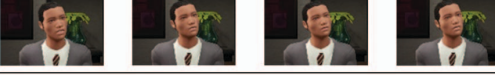



Speaking Phase				
Scene				
Pre-speaker-text	Tom Brady is not your MVP?		Not my fault.	
Speaker-text	I’ve been on his bus since mid-October		Your Dallas Cowboys are dead. How do you feel.	
Speaker-ground truth				
Speaker-synthesis on avatar				
Speaker-synthesized images				

Figure 6. Experimental results of our models in speaking phase. Two scenes are shown, each of which includes the conversation scene (the first row), the verbal cues of two persons in the scene (second row for the previous speaker and third row the present speaker), the speaker’s ground truth images (fourth row), the speaker’s 3d model based avatar (fifth row), and the speaker’s synthesized images (last row).

5. Conclusion

We present a realistic face-to-face conversation system based on deep neural networks. In listening phase, the speaker’s face features including action unit and head pose are firstly extracted, then transformed into the avatar’s face features by a seq2seq model, and finally used to generate realistic face images (avatar) by a pix2pix based GAN network. In speaking phase, the previous speaker’s words are turned into word features by word2vec, then transformed

into the avatar’s face features by another seq2seq model, and finally used to generate realistic face images (avatar) by the pix2pix based GAN network. The model is trained and evaluated with the video data from the ESPN shows. For comparison, both the realistic avatar generated by the proposed models and the one driven by traditional 3D model based method are given. It shows that the proposed conversation system can generate natural and realistic avatars. In the future, multi-modal information will be considered to produce more personalized conversation avatars.

References

- [1] T. Baltrusaitis, P. Robinson, and L. Morency. Openface: an open source facial behavior analysis tool. In *Proc. Workshop on Applications of Computer Vision (WACV)*, IEEE, 2016. 3
- [2] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros. Everybody dance now. *arXiv preprint arXiv:1808.07371*, 2018. 2, 4
- [3] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27, 2011. 3
- [4] H.-Y. Chen et al. Tensorflow—a system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016. 5
- [5] H. Chu, D. Li, and S. Fidler. A face-to-face neural conversation model. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7113–7121, 2018. 1, 2
- [6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *international Conference on computer vision & Pattern Recognition (CVPR'05)*, volume 1, pages 886–893. IEEE Computer Society, 2005. 3
- [7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 4
- [8] T. Devries, K. Biswaranjan, and G. W. Taylor. Multi-task learning of facial landmarks and expression. In *Computer & Robot Vision*, 2014. 2
- [9] Y. Fan, Y. Qian, F.-L. Xie, and F. K. Soong. Tts synthesis with bidirectional lstm based recurrent neural networks. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014. 3
- [10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2
- [11] A. Graves, A.-r. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. IEEE, 2013. 2
- [12] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, B. Kingsbury, et al. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal processing magazine*, 29, 2012. 2
- [13] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2, 4
- [14] L. Hu, S. Saito, L. Wei, K. Nagano, J. Seo, J. Fursund, I. Sadeghi, C. Sun, Y.-C. Chen, and H. Li. Avatar digitization from a single image for real-time rendering. *ACM Transactions on Graphics (TOG)*, 36(6):195, 2017. 1
- [15] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2
- [16] M. Johansson, G. Skantze, and J. Gustafson. Head pose patterns in multiparty human-robot team-building interactions. In *International conference on social robotics*, pages 351–360. Springer, 2013. 2
- [17] M. Katzenmaier, R. Stiefelhofen, and T. Schultz. Identifying the addressee in human-human-robot interactions based on head pose and speech. In *Proceedings of the 6th international conference on Multimodal interfaces*, pages 144–151. ACM, 2004. 2, 3
- [18] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 5, 6
- [19] J. J. Lien, T. Kanade, J. F. Cohn, and C.-C. Li. Automated facial expression recognition based on face action units. In *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 390–395. IEEE, 1998. 2
- [20] J. J.-J. Lien, T. Kanade, J. F. Cohn, and C.-C. Li. Detection, tracking, and classification of action units in facial expression. *Robotics and Autonomous Systems*, 31(3):131–146, 2000. 2
- [21] Z. Liu, Z. Chen, J. Bai, S. Li, and S. Lian. Facial pose estimation by deep learning from label distributions. *arXiv preprint arXiv:1904.13102*, 2019. 2
- [22] Z. Liu, H. Hu, Z. Wang, K. Wang, J. Bai, and S. Lian. Video synthesis of human upper body with realistic face. *arXiv preprint arXiv:1908.06607*, 2019. 2
- [23] Y. Matsuyama, A. Bhardwaj, Z. Ran, O. Romeo, S. Akoju, and J. Cassell. Socially-aware animated intelligent personal assistant agent. In *Meeting of the Special Interest Group on Discourse & Dialogue*, 2016. 1
- [24] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *Computer Science*, 2013. 4
- [25] K. Nickel and R. Stiefelhofen. Visual recognition of pointing gestures for human-robot interaction. *Image and vision computing*, 25(12):1875–1884, 2007. 2
- [26] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 4
- [27] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. *CoRR*, abs/1802.05365, 2018. 4
- [28] E. Sangineto. Pose and expression independent facial landmark localization using dense-surf and the hausdorff distance. *IEEE transactions on pattern analysis and machine intelligence*, 35(3):624–638, 2013. 3
- [29] E. Seemann, K. Nickel, and R. Stiefelhofen. Head pose estimation using stereo vision for human-robot interaction. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.*, pages 626–631. IEEE, 2004. 2
- [30] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4
- [31] R. Stiefelhofen, C. Fugen, R. Giesemann, H. Holzapfel, K. Nickel, and A. Waibel. Natural human-robot interaction using speech, head pose and gestures. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566)*, volume 3, pages 2422–2427. IEEE, 2004. 3
- [32] R. Subramanian, Y. Yan, J. Staiano, O. Lanz, and N. Sebe. On the relationship between head pose, social attention and

- personality prediction for unstructured and dynamic group interactions. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 3–10. ACM, 2013. 2
- [33] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014. 2, 3, 4, 6
- [34] Y.-I. Tian, T. Kanade, and J. F. Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 23(2):97–115, 2001. 2
- [35] O. Vinyals and Q. V. Le. A neural conversational model. *CoRR*, abs/1506.05869, 2015. 2
- [36] S. Waldherr, R. Romero, and S. Thrun. A gesture based interface for human-robot interaction. *Autonomous Robots*, 9(2):151–173, 2000. 2
- [37] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*, 2018. 2
- [38] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, 2018. 2, 4
- [39] Z. Wang, Z. Liu, Z. Chen, H. Hu, and S. Lian. A neural virtual anchor synthesizer based on seq2seq and gan models. *arXiv preprint arXiv:1908.07262*, 2019. 2
- [40] H. Zen and H. Sak. Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4470–4474. IEEE, 2015. 3
- [41] L. Zhang. Automatic adaptation of a face model using action units for semantic coding of videophone sequences. *IEEE Transactions on Circuits and Systems for Video Technology*, 8(6):781–795, 1998. 2
- [42] Y. Zhang, M. Brady, and S. Smith. Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE transactions on medical imaging*, 20(1):45–57, 2001. 2
- [43] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017. 2