Towards Conditional Adversarial Training for Predicting Emotions from Speech

Jing Han, Zixing Zhang, Zhao Ren, Fabien Ringeval, Björn Schuller

TOWARDS CONDITIONAL ADVERSARIAL TRAINING FOR PREDICTING EMOTIONS FROM SPEECH

Jing Han[1]*, Zixing Zhang*[2]*, Zhao Ren*[1]*, Fabien Ringeval*[3]*, Björn Schuller*[1,2]

[1]ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[2]Group on Language, Audio & Music, Imperial College London, London, UK
[3]Laboratoire d'Informatique de Grenoble, Université Grenoble Alpes, France

jing.han@informatik.uni-augsburg.de

ABSTRACT

Motivated by the encouraging results recently obtained by generative adversarial networks in various image processing tasks, we propose a conditional adversarial training framework to predict dimensional representations of emotion, i. e., arousal and valence, from speech signals. The framework consists of two networks, trained in an adversarial manner: The first network tries to predict emotion from acoustic features, while the second network aims at distinguishing between the predictions provided by the first network and the emotion labels from the database using the acoustic features as conditional information. We evaluate the performance of the proposed conditional adversarial training framework on the widely used emotion database RECOLA. Experimental results show that the proposed training strategy outperforms the conventional training method, and is comparable with, or even superior to other recently reported approaches, including deep and end-to-end learning.

Index Terms— Emotion recognition, conditional adversarial training, generative adversarial network

1. INTRODUCTION

Deep learning has shown great potential for Speech Emotion Recognition (SER), i. e., the automatic recognition of emotion from speech. Increasing research efforts have been made during the last years to develop more accurate and robust emotion analysis systems based on deep learning techniques. For example, in [1], Recurrent Neural Networks (RNNs) equipped with Long Short-Term Memory (LSTM) cells were firstly investigated for SER, owing to its powerful capability of learning a long range of temporal-dependent patterns. Henceforth, LSTM-RNNs (or BLSTM for its bi-directional version) is frequently employed for time- and value-continuous SER as it has been found very effective [1–3]. Moreover, in [4], generatively pre-trained Deep Neural Networks (DNNs) were utilised to learn discriminative features of low dimension, and its comparative performance was evaluated on large-scale databases. In [5], deep Convolutional Neural Networks (CNNs) were employed to extract affect salient features directly from the spectrogram, and outperformed several well-established acoustic features. More recently, an attention mechanism was further introduced to identify the contribution of each feature extracted by the network [6].

Apart from these works that intend to explore innovative neural network architectures for SER, various advanced training strategies have been investigated as well. Furthermore, prediction-based learning was introduced in [7] to incorporate the strength of different models, and the reconstruction-error based learning was examined in [8] to compensate the weakness of a neural network itself. In [9], CNNs and LSTM-RNNs were sophisticatedly concatenated into a joint framework, and trained in an end-to-end manner, i. e., by directly learning a suitable representation of the raw signal. Besides, multi-task learning was proposed to take benefits of contextual information that shape emotion, such as age and gender [10, 11].

In this paper we propose a novel network training framework based on *conditional adversarial training*. This framework involves two networks, where the first network tries to best generate predictions as close as possible to the labels, and these predictions are joined with the original features in order to 'cheat' the second network that is responsible for identifying the input source, i. e., either coming from a machine or from a pool of annotators. Compared with the traditional pattern recognition model, the first network is trained with an adversarial feedback from the second network, which helps improving the reliability of the model learned in the first network.

This idea mainly stems from *Generative Adversarial Networks* (GANs) [12], which has recently attracted striking attention in machine learning. However, most efforts related to GANs were made on how to best generate sufficiently realistic samples, such as images and text [13, 14]. Few studies have considered it for pattern recognition [15, 16]. To the best of our knowledge, this is the first tentative work towards this research direction, especially in the field of affective computing.

Whereas the presented conditional adversarial training framework utilises a cascaded structure, as used in our previous work, i. e., prediction-based learning [7] and reconstruction-error-based learning [8], it includes some specific advantages in comparison to those two approaches. The main idea of prediction-based learning is to take advantage of different models where predictions made by a first model are combined with the original features to learn a second model. Therefore, the two models should be diverse enough to provide complementary views and compensate their respective weaknesses. Whereas the reconstruction-error-based learning aims to explore the model weakness information that can be quantified by the reconstruction error [8], in the assumption that the model could perform better if we explicitly let it be aware of its errors. Thus, the two models should be as similar as possible, so that the weakness information extracted from the first model can be presented for the second one. Both learning strategies are realised in an asynchronously cooperative ways. That is, the well-trained first model provides additional information for the second one to assist in final decision making. The proposed conditional adversarial training framework, however, does not care about the similarity of the two models. Moreover, the two models (networks) can be trained and optimised synchronously in a competitive way, rather than asynchronously in a cooperative way.

2. RELATED WORK

The proposed learning strategy closely relates to GANs, which is a successful alternative to conventional maximum likelihood techniques. GAN was first proposed in [12], where a deep generative model G can be learnt to model the data distribution of the target, while training jointly with another discriminative model D as two players in a minimax game. To be more specific, while the discriminator is trained to estimate the probability that a sample comes from, the real data rather than the output of the generator, the generator learns to maximise the probability of fooling the discriminator. Training of the generator and the discriminator is done iteratively, i. e., weights are updated in turns to compete with each other. Once training is achieved, the generator is able to generate more realistic samples under such an adversarial training strategy. Many researchers in machine learning have reported impressive results with GANs and developed a bulk of variants, for instance, conditional GAN [17], cycle GAN [18], and Wasserstein GAN (WGAN) [19].

In the speech processing domain, however, merely a few works have been reported so far. In [20], GANs were regarded as a back-end filter to compensate the difference between natural and synthesised speech. Similar work has also been done in [21], where GANs were used to generate speech samples with a distribution close to natural speech. GANs were also employed to enhance noisy speech from spectrograms [22], or enrich the volume and diversity of training material used to detect autism spectrum conditions from speech [23]. Overall, one could notice that GANs are mainly applied to generate and synthesise samples of speech, which differs from the goal of our learning framework that is particularly designed for prediction.

3. CONDITIONAL ADVERSARIAL TRAINING FOR PREDICTION

In this section, we firstly have a brief introduction of conditional GANs. Then, we describe the framework of the proposed conditional adversarial training as well as its advanced version.

3.1. Conditional Generative Adversarial Networks

Conditional GAN (CGAN) is a variation of traditional GAN, where both the generator and discriminator are conditioned on certain extra information c [17]. In the generator, the prior input noise variable $p_z(\mathbf{z})$ is combined with the conditional information c as a joint hidden representation. The process of adversarial training then tries to find how this hidden representation is composed. Likewise, in the discriminator, the real data $\mathbf{x}$ or the simulated data from the generator $G(\mathbf{z})$ is further extended with the conditional information c, which are then fed into the network that is responsible for discrimination.

Mathematically, the generator G is trained to minimise the objective function:

$$\mathcal{L}_G = \mathbb{E}[log(D(G(\mathbf{z}), c))], \tag{1}$$

while the discriminator D is trained to maximise the log-likelihood it assigns to the real data:

$$\mathcal{L}_D = \mathbb{E}[log(D(\mathbf{x}, c))] + \mathbb{E}[log(1 - D(G(\mathbf{z}), c))]. \tag{2}$$

Therefore, this learning process is similar to a minimax game between the generator and the discriminator.

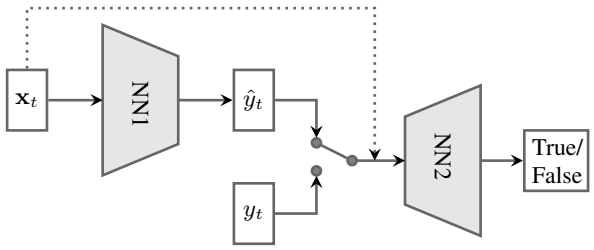

Fig. 1. Framework of conditional adversarial training for prediction: a first model (NN1) predicts time-continuous labels $\hat{y}_t$ from a set of acoustic features $\mathbf{x}_t$, whereas a second model (NN2) infers a binary decision whether the input source comes from the real data y_t or from the first model NN1, given the context $\mathbf{x}_t$.

3.2. Conditional Adversarial Training for Prediction

In contrast to CGAN used for data generation, we propose to exploit the concept of adversarial training to build a predictive model as discussed in Section 1. The framework of the proposed conditional adversarial training is illustrated in Fig. 1. Whereas the structure is analogue to any CGAN with the presence of two networks, the ultimate goal of our system is to obtain an accurate pattern estimation from the generator which is guided by the discriminator.

For this purpose, the algorithm of CGAN is modified. Specifically, we ignore the prior random noise, and consider the features as conditional information, i. e.,

$$c \leftarrow \mathbf{x}, \tag{3}$$

and the predicted emotional values as the generated data, i. e.,

$$G(\mathbf{z}) \leftarrow \hat{y}. \tag{4}$$

The first network (NN1) is thus derived into a 'conventional-like' recognition model, and learns the conditional distribution $P(y_t|\mathbf{x}_t)$ given sequential features $\mathbf{x}_t$ and their labels y_t. Nevertheless, the major difference is that the NN1 is optimised not only through its own prediction error, but also with the aid of adversarial feedback from the second network (NN2).

For NN2, similar to CGAN, we extend the generated one-dimensional predictions ($\hat{y}_t$) or the true labels (y_t) with the auxiliary information (i. e., the original feature vectors, $\mathbf{x}_t$) to obtain a joint representation, i. e., $[\mathbf{x}_t, \hat{y}_t]$ or $[\mathbf{x}_t, y_t]$, and then feed it into the network. The network learns to identify whether the joint representation comes from the generation of the first network (False) or from the original labels (True). Therefore, NN2 is trained to distinguish the joint probability distributions for features and their corresponding 'predicted' (false) labels, i. e., $P_g(\mathbf{x}_t, \hat{y}_t)$, or 'real' labels, i. e., $P_r(\mathbf{x}_t, y_t)$.

More concretely, NN1 is optimised by changing Eq. (1) into

$$\mathcal{L}_{NN1} = \mathbb{E}(|G(\mathbf{x}_t) - y_t|^2) + \lambda * \mathbb{E}(log(D(G(\mathbf{x}_t), \mathbf{x}_t))), \tag{5}$$

where the first item indicates the Mean Square Error (MSE) between the prediction and the label, and λ denotes a hyper-parameter that controls the contribution of the adversarial information from NN2, which is optimised by

$$\mathcal{L}_{NN2} = \mathbb{E}(log(D(y_t, \mathbf{x}_t))) + \mathbb{E}(log(1 - D(G(\mathbf{x}_t), \mathbf{x}_t))). \tag{6}$$

Therefore, intuitively, NN1 is optimised to generate predictions as close as possible to the labels, while fooling NN2 when fed with

joint distributions composed of predictions from NN1 and acoustic features.

3.3. Optimising with Wasserstein Distance

Traditional generative modelling approaches rely on maximising likelihood, or equivalently minimising the Kullback-Leibler (KL) divergence between the realistic data distribution P_r and the generated data distribution P_g [12]. One major issue this approach suffers is the vanishing gradient problem as demonstrated in [24], because the discriminator with infinite ability of separating real from generated samples will lead to a constant Jensen-Shannon (JS) divergence between P_r and P_g when their supports have no or negligible overlap [24]. This results in an impossibility to update the generator accordingly, as the discriminator is quickly trained towards its optimality [24].

In this light, a Wasserstein (*aka* Earth-Mover) distance was proposed most recently [19], so that the JS distance problem in the classic GAN can be solved by showing that the Wasserstein distance is continuous and differential almost everywhere. Motivated by this work, we further update the NN2 training strategy by maximising

$$\mathcal{L}_{NN2} = \mathbb{E}[D(\mathbf{x}_t)] - \mathbb{E}[D(G(\mathbf{x}_t))], \tag{7}$$

and the NN1 training strategy by minimising

$$\mathcal{L}_{NN1} = \mathbb{E}(|G(\mathbf{x}_t) - y_t|^2) + \lambda * \mathbb{E}[D(G(\mathbf{x}_t))]. \tag{8}$$

In addition, a weight clipping is applied to the NN2 as follows

$$W_{NN2} \leftarrow \text{clip\_by\_value}(W_{NN2}, -0.01, 0.01), \tag{9}$$

in order to improve the stability of training as suggested in [19].

4. EXPERIMENTS AND RESULTS

To evaluate the effectiveness of the proposed training approach for emotion recognition, we conducted extensive experiments on a widely used database in the affective computing community.

4.1. Selected Database and Acoustic Features

The multimodal corpus REmote COLlaborative and Affective interactions (RECOLA) [25] is chosen for our experiments, owing to its widespread examination in continuous emotion recognition [26, 27]. To collect this database, 46 French-speaking participants were asked to process remote collaborative work. Spontaneous and natural interactions were collected during resolving of a collaborative task that was performed in dyads and remotely through video conference. The corpus consists of multimodal signals, i. e., audio, video, Electro-CardioGram (ECG), and Electro-Dermal Activity (EDA), which were recorded continuously and synchronously. In our study, however, only audio signals are utilised for a tentative evaluation.

It is worth mentioning that, these participants have different mother tongues, i. e., French, Italian, and German, which provides further diversity in the encoding of affect. In order to ensure speaker-independence, the corpus was almost equally divided into three partitions, i. e., training (16), development (15), and test (15), by approximate balancing the gender, age, and mother tongue of the participants.

To annotate the corpus, value- and time-continuous dimensional affect ratings in terms of *arousal* and *valence* were performed by six French-speaking raters (three females) for the first five minutes of all recording sequences. The obtained labels were then resampled at a constant frame rate of 40 ms, and averaged over all raters by considering inter-evaluator agreement, to provide a 'gold standard' [25].

To extract acoustic features from the speech recordings, we took our open-source openSMILE toolkit to extract 13 Low-Level Descriptors (LLDs), i. e., Mel Frequency Cepstral Coefficients (MFCCs) 0-12 and logarithmic energy, with a frame window size of 25 ms and a step of 10 ms. The arithmetic mean and the coefficient of variance were then computed over the sequential LLDs at a rate of 40 ms – to match the granularity of the annotation – using overlapping windows of 8 s length, resulting in 26 statistical features per analysis window. The total numbers of segments in the train, development, and test partitions are 120.0 k, 112.5 k, and 112.5 k, respectively.

4.2. Experimental Setup and Evaluation Metric

We implemented the framework with LSTM-RNN, since it has been frequently examined to be effective in capturing long-range context information for sequential pattern recognition tasks, for example, continuous emotion recognition in our case [1, 27]. For the sake of fair comparison, we took the same network structure as the one used in [7] for both NN1 and NN2. That is, the number of hidden layers was set to be two and the number of nodes per hidden layer to be 20. To accelerate the training process, the network weights were updated after running every mini-batch of eight sequences for computation in parallel.

To train the networks, an on-line standardisation was applied to the development and test partitions by using the means and variations of the training partition. Besides, annotation delay compensation was also performed to compensate for the temporal delay between the observable cues shown by the participants, and the corresponding emotion reported by the annotators. We set this delay to be four seconds (as suggested in our previous experiments [7]) which was duly compensated, by shifting the gold standard back in time with respect to the features for both arousal and valence in all of our experiments.

To evaluate the systems, we considered the official metric for the AVEC in 2015 [26] and 2016 [27] in which a subset of the corpus was featured, namely *Concordance Correlation Coefficient* (CCC) [26]:

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}, \tag{10}$$

where ρ presents the *Pearson's Correlation Coefficient* (PCC) between two time series (e. g., prediction and gold-standard), μ_x and μ_y denote the means of each time series, and σ_x^2 and σ_y^2 are the corresponding variances. Compared with PCC, CCC takes not only the linear correlation, but also the bias and variance between the two compared series into consideration. For continuous emotion recognition, we are often interested in not only the prediction trend but also the absolute value/degree of personal emotional state. Therefore, the metric of CCC fits better for continuous emotion recognition than PCC. Particularly, the value of CCC is in the range of [-1, 1], where $+1$ represents total concordance, -1 total discordance, and 0 no concordance at all.

To refine the obtained predictions, we further performed a chain of *post-processing*, including median filtering, centring, scaling, and time-shifting as suggested in [26, 27]. All the post-processing parameters were determined on the development set, and then applied to the test set.

To estimate the statistical significance level of performance improvement, *Fisher's r-to-z transformation* [28] was carried out over

Table 1. Performance in terms of Concordance Correlation Coefficient (CCC) of the proposed conditional adversarial training approaches, as well as its variation (+ Wasserstein distance), for both *arousal* and *valence* regressions, evaluated on the *dev*elopment and *test* partitions.

approaches	arousal		valence	
	dev	test	dev	test
baseline				
LSTM-RNN (2 layers)	.777	.718	.491	.435
LSTM-RNN (4 layers)	.761	.723	.487	.390
state of the art				
CCC-objective [29]	.412	.350	.242	.199
end-to-end [9]	.741	.686	.325	.261
reconstruction-error-based [8]	.785	.729	.378	.360
prediction-based [7]	.774	**.744**	.412	.377
proposed				
conditional adversarial training	.780	.732	**.501**	**.455**
+ Wasserstein Distance	**.797**	.737	.474	.444

the whole predictions between the proposed and the baseline approaches. Unless stated otherwise, a p value less than 0.05 indicates significance.

4.3. Results and Discussion

In the network training process, we alternatively trained NN1 and NN2, and repeated this process (runs). In each learning run, we conducted more training times on NN2 than NN1. More specially, we experientially carried out 10 steps on NN1, and 50 steps on NN2 for arousal prediction, and 15 steps on NN2 for valence prediction. This operation is twofold: (i) NN2 is required to be superior enough [12], otherwise it is vulnerable to be 'cheated' by the predictions, and could not provide sufficient challenges to advance NN1; (ii) valence prediction is normally considered as a harder task than arousal prediction. Thus, NN1 needs relatively more training steps for valence compared with arousal in each run. Besides, we optimised the hyper-parameter of λ in Eq. (5) and (8) by a grid search in the range of [.01, .02, .05, 0.1, .2, .5] on the development set.

Table 1 displays the result performance in terms of CCC obtained from the systems by using conditional adversarial training approaches, on the development and test partitions of RECOLA from speech signals. For comparison, we conducted traditional training approaches without conditional adversarial training as baselines. The networks with two hidden layers and four hidden layers were respectively evaluated. It should be noted that the benchmarks are slightly different from the ones presented in previous work [7], which might be mainly due to the change of experimental platforms (from CURRENNT to Tensorflow).

Compared with the baseline, it can be seen that the system performance is significantly improved for both arousal and valence predictions (via a Fisher's r-to-z transformation as outlined in Section 4.2), when performing conditional adversarial training. Specifically, on the test set, the CCC values increase to .732 for arousal predictions, and .455 for valence predictions. The performance gain indicates that adversarial training with NN2 brings benefit to NN1 to further ameliorate its predictions to some extent.

When implementing the Wasserstein distance into the objective function of NN2, the obtained results are shown in the last row of Table 1. One can observe that the performance of the system for arousal prediction is further enhanced, i. e., from .780 to .797 of CCCs on the development set, and from .732 to .737 of CCCs on the test set; whereas for valence, one cannot get a similar observation. This implicitly suggests that for arousal, it is somewhat effortless for NN2 to distinguish the input sources.

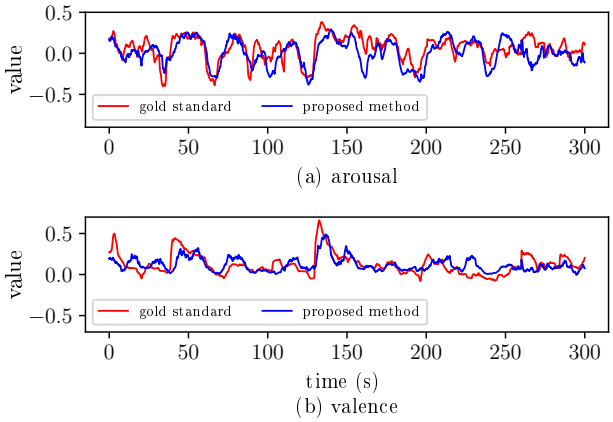

Fig. 2. Automatic predictions of arousal (a) and valence (b) obtained by conducting conditional adversarial training, for a randomly selected subject from the test partition on RECOLA database.

Furthermore, one might notice that our systems is comparable to the state-of-the-art systems as listed in Table 1 for arousal prediction, and outperforms all other systems for the valence prediction, which yields the best results to date on the RECOLA database from speech.

To intuitively present the system performance, we illustrate the arousal and valence predictions on a randomly selected subject from the test partition in Fig. 2 (a) and (b), respectively. From the figure, it is clear to observe that the predictions (blue lines) and the corresponding gold standards (red lines) have a high correlation.

5. CONCLUSIONS

In contrast to previous works that use adversarial training for generation, in this paper, we tentatively examined the performance of conditional adversarial training in the application of Speech Emotion Recognition (SER). To stablise the learning process, we further modified the objective function by using Wasserstein distance. A set of experiments have been conducted on RECOLA to assess the training performance, and we find that conditional adversarial training is helpful to improve the system performance for SER.

Future work includes more experimental evaluations on other prediction tasks with a larger size of data [30]. Moreover, it is interesting to perform an end-to-end structure to automatically extract salient features for emotion prediction [9], rather than the hand-crafted features that were employed in this present framework.

6. ACKNOWLEDGEMENTS

This work was supported by the European Union's Horizon 2020 Programme through the Innovative Action No. 645094 (SEWA) and the UK's Economic & Social Research Council through the research Grant No. HJ-253479 (ACLEW).

7. REFERENCES

[1] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie, "Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies," in *Proc. INTERSPEECH*, Brisbane, Australia, 2008, pp. 597–600.

[2] A. Metallinou, M. Wöllmer, A. Katsamanis, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive learning for enhanced audiovisual emotion classification," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 184–198, Jan. 2012.

[3] F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J.-P. Thiran, T. Ebrahimi, D. Lalanne, and B. Schuller, "Prediction of Asynchronous Dimensional Emotion Ratings from Audiovisual and Physiological Data," *Pattern Recognition Letters*, vol. 66, pp. 22–30, November 2015.

[4] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, H. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: Raising the benchmarks," in *Proc. ICASSP*, Prague, Czech Republic, 2011, pp. 5688–5691.

[5] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2203–2213, Dec. 2014.

[6] M. Neumann and N. T. Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," *CoRR*, vol. abs/1706.00612, June 2017.

[7] J. Han, Z. Zhang, F. Ringeval, and B. Schuller, "Prediction-based learning for continuous emotion recognition in speech," in *Proc. ICASSP*, New Orleans, LA, 2017, pp. 5005–5009.

[8] ——, "Reconstruction-error-based learning for continuous emotion recognition in speech," in *Proc. ICASSP*, New Orleans, LA, 2017, pp. 2367–2371.

[9] G. Trigeorgis, F. Ringeval, R. Bruckner, E. Marchi, M. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? End-to-end speech emotion recognition using a Deep Convolutional Recurrent Network," in *Proc. ICASSP*, Shanghai, China, 2016, pp. 5200–5204.

[10] R. Xia and Y. Liu, "A multi-task learning framework for emotion recognition using 2D continuous space," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 3–14, Jan. 2017.

[11] B. Zhang, E. M. Provost, and G. Essl, "Cross-corpus acoustic emotion recognition with multi-task learning: Seeking common ground while preserving differences," *IEEE Transactions on Affective Computing*, vol. PP, no. 99, pp. 1–1, Mar. 2017.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NIPS*, Montreal, Canada, 2014, pp. 2672–2680.

[13] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text-to-image synthesis," in *Proc. ICML*, New York, NY, 2016, pp. 1060–1069.

[14] L. Yu, W. Zhang, J. Wang, and Y. Yu, "SeqGAN: Sequence generative adversarial nets with policy gradient." in *Proc. AAAI*, San Francisco, CA, 2017, pp. 2852–2858.

[15] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. ICLR*, San Juan, PR, 2016, no pagination.

[16] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGan: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. NIPS*, Barcelona, Spain, 2016, pp. 2172–2180.

[17] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *CoRR*, vol. abs/1411.1784, June 2014.

[18] J. Zhu, T. Park, P. Isola, and A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *CoRR*, vol. abs/1703.10593, Mar. 2017.

[19] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," *CoRR*, vol. abs/1701.07875, Mar. 2017.

[20] T. Kaneko, H. Kameoka, N. Hojo, Y. Ijima, K. Hiramatsu, and K. Kashino, "Generative adversarial network-based postfilter for statistical parametric speech synthesis," in *Proc. ICASSP*, New Orleans, LA, 2017, pp. 4910–4914.

[21] S. Yang, L. Xie, X. Chen, X. Lou, X. Zhu, D. Huang, and H. Li, "Statistical parametric speech synthesis using generative adversarial networks under a multi-task learning framework," *CoRR*, vol. abs/1707.01670, July 2017.

[22] D. Michelsanti and Z.-H. Tan, "Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification," in *Proc. INTERSPEECH*, Stockholm, Sweden, 2017, pp. 2008–2012.

[23] J. Deng, N. Cummins, M. Schmitt, K. Qian, F. Ringeval, and B. Schuller, "Speech-based diagnosis of autism spectrum condition by generative adversarial network representations," in *Proc. the 2017 International Conference on Digital Health*, London, UK, 2017, pp. 53–57.

[24] M. Arjovsky and L. Bottou, "Towards principled methods for training generative adversarial networks," *CoRR*, vol. abs/1701.04862, Jan. 2017.

[25] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *Proc. EmoSPACE (FG)*, Shanghai, China, 2013, pp. 1–8.

[26] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic, "AV+EC 2015: The first affect recognition challenge bridging across audio, video, and physiological data," in *Proc. the 5th AVEC Workshop*, Brisbane, Australia, 2015, pp. 3–8.

[27] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "AVEC 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proc. the 6th AVEC Workshop*, Amsterdam, The Netherlands, 2016, pp. 3–10.

[28] J. Cohen, P. Cohen, S. G. West, and L. S. Aiken, *Applied multiple regression/correlation analysis for the behavioral sciences*. Abingdon, UK: Routledge, 2013.

[29] F. Weninger, F. Ringeval, E. Marchi, and B. Schuller, "Discriminatively trained recurrent neural networks for continuous dimensional emotion recognition from audio," in *Proc. IJCAI*, New York, NY, 2016, pp. 2196–2202.

[30] Z. Zhang, N. Cummins, and B. Schuller, "Advanced data exploitation for speech analysis – An overview," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 107–129, July 2017.