Improving Neural Silent Speech Interface Models by Adversarial Training

Amin Honarmandi Shandiz[1]([✉]), László Tóth[1], Gábor Gosztolya[2],
Alexandra Markó[4,5], and Tamás Gábor Csapó[3,4]

[1] Institute of Informatics, University of Szeged, Szeged, Hungary
{shandiz,tothl}@inf.u-szeged.hu
[2] MTA-SZTE Research Group on Artificial Intelligence, Szeged, Hungary
ggabor@inf.u-szeged.hu
[3] Department of Telecommunications and Media Informatics,
Budapest University of Technology and Economics, Budapest, Hungary
csapot@tmit.bme.hu
[4] MTA-ELTE "Lendület" Lingual Articulation Research Group, Budapest, Hungary
marko.alexandra@btk.elte.hu
[5] Department of Applied Linguistics and Phonetics, Eötvös Loránd University,
Budapest, Hungary

Abstract. Besides the well-known classification task, these days neural networks are frequently being applied to generate or transform data, such as images and audio signals. In such tasks, the conventional loss functions like the mean squared error (MSE) may not give satisfactory results. To improve the perceptual quality of the generated signals, one possibility is to increase their similarity to real signals, where the similarity is evaluated via a discriminator network. The combination of the generator and discriminator nets is called a Generative Adversarial Network (GAN). Here, we evaluate this adversarial training framework in the articulatory-to-acoustic mapping task, where the goal is to reconstruct the speech signal from a recording of the movement of articulatory organs. As the generator, we apply a 3D convolutional network that gave us good results in an earlier study. To turn it into a GAN, we extend the conventional MSE training loss with an adversarial loss component provided by a discriminator network. As for the evaluation, we report various objective speech quality metrics such as the Perceptual Evaluation of Speech Quality (PESQ), and the Mel-Cepstral Distortion (MCD). Our results indicate that the application of the adversarial training loss brings about a slight, but consistent improvement in all these metrics.

Keywords: Silent speech interface · Generative adversarial network · Adversarial training · Neural vocoder · Articulatory-to-acoustic mapping

1 Introduction

Human speech production is a very complex motor process from speech planning in the brain to the production of the acoustic signal by precisely coordi-

nated respiratory, laryngeal and articulatory movements [1]. This process has three intermediate phases which may be recorded without recording the actual speech signal [1,2]. In the cases of imagined speech and inner speech, there is no articulatory movement, so only the brain activity may be recorded by brain-computer interfaces like functional Magnetic Resonance Imaging (fMRI) or functional Near-Infrared Spectroscopy (fNIRS). The easiest case is when the articulatory movement is performed by the subject, as these articulatory muscle movements can be recorded by several types of devices. The basic concept of silent speech interfaces (SSI) is to reconstruct the speech signal from these articulatory signals. The most important recording methods are Electromagnetic Articulography (EMA) [3,4], Ultrasound Tongue imaging (UTI) [5–7], permanent Magnetic Articulography (PMA) [8,9], and surface electromyography (sEMG) [10,11]. A video of the lip movements can also be considered as input, alone or in a combination with the above methods [7,12]. We envision the application of such SSI systems in two main situations. The first is to aid the speech communication of impaired people, such as patients after laryngectomy. The second case is when talking loudly is not appropriate or impossible because of an extreme background noise, like in certain industrial or military situations.

There are two main approaches to articulatory-to-acoustic conversion, that is, to estimate the speech signal from a recording of the articulatory movement. The two-step approach first converts the signal to text, and then converts the text to speech [3,7,9,10]. Recently, with the invention of very powerful deep learning techniques, the direct approach has become more popular, which converts the articulatory signal to a speech signal without any intermediate steps, typically applying a deep neural network (DNN) [4,6,8,11]. In this paper we follow the second approach using ultrasound tongue imaging videos as the input, and we will experiment with a special DNN configuration to perform the conversion.

When we utilize DNNs to generate signals such as images or speech, the choice of the optimal loss function is not trivial, as the conventional error functions such as the mean squared error (MSE) do not fully reflect the sense of quality or similarity perceived by humans. Hence, we may expect better results by applying perceptually motivated loss functions, for which we have two options. First, speech technology has defined several objective speech quality metrics, and there were attempts to reformulate these as loss functions for DNN training [13,14]. Alternatively, one may train a second DNN to estimate the perceptual similarity of two images or speech signals [15,16]. This network is trained to discriminate 'real' signals from artificially generated 'fake' signals, so it will be called the discriminator, while the original network will be referred to as the generator. These two networks can be trained in parallel, and this combination of two nets is known as a Generative Adversarial Network (GAN). This work examines the applicability of GANs to improving the quality of speech signals generated by an ultrasound-based neural SSI model. We will compare the GAN-based training strategy with the conventional MSE-based training on Hungarian and English corpora, and we will evaluate the models via objective speech quality metrics.

The paper is organized as follows. In Sect. 2, we introduce the problem of generating speech in SSI systems. In Sect. 3 we explain the GAN framework, and Sect. 4 presents the experimental conditions. The results are presented and discussed in Sect. 5, and the paper is closed with the conclusions in Sect. 6.
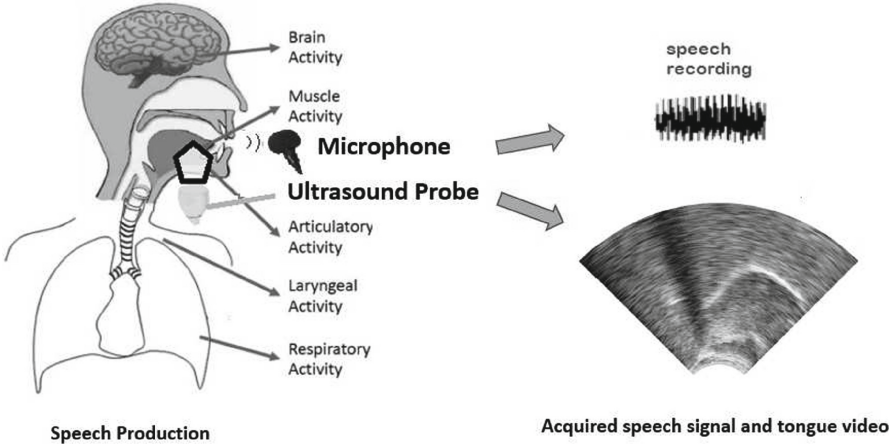

Fig. 1. Illustration of the data acquisition process.

2   Silent Speech Interfaces Built on Ultrasound Tongue Imaging and Neural Vocoders

Here, we worked with an ultrasound-based data acquisition technique. In this framework, an ultrasound probe is placed under the chin of the subject, so it can record images of the tongue movement in a midsagittal orientation (see Fig. 1). The device applied was the "Micro" ultrasound system produced by Articulate Instruments Ltd. To support DNN training, the subjects were asked to speak loud, and their speech was recorded in parallel with the ultrasound output (cf. Fig. 1 again). The ultrasound and speech recordings served as the training inputs and training targets for the DNN, respectively. As the input and output signals were synchronized, we applied conventional network structures which perform the conversion in a time-synchronized, frame-by-frame manner.

Training a network to generate speech signals directly would require a huge amount of training data. Hence, we opted for a two-step solution, which was motivated by speech synthesis. In text-to-speech systems, a popular approach is to first convert the text to a spectrogram, then convert the spectrogram to a speech signal. Nowadays, both steps can be implemented by neural networks [17]. To adjust this approach to our task, only the first network needs to be modified, as our input is an ultrasound video and not a text. This way, we can apply large,

pre-trained networks for the second task, while our network has to estimate only a dense spectral representation instead of the speech waveform itself. Several neural vocoders are available for the speech generation task [18], and we chose to use the WaveGlow model [17], as we applied it successfully earlier [19]. As WaveGlow requires a sequence of 80-dimensional mel-scaled spectral vectors as input, the task of our network was to estimate such a spectral vector form each frame of the video. The next section discusses the details of this step.

3   Generative Adversarial Networks for Articulatory-to-Acoustic Mapping

As we use WaveGlow for the speech synthesis step, the task for our DNN was to convert each ultrasound video frame (64×128 pixels) into a mel-spectral vector (with 80 components). In an earlier paper we found that the best results can be achieved by using a short sequence of video frames as input, rather than just a single frame [5]. We got the lowest error rate with a special 3D convolutional neural network (CNN) structure, and other authors reported similar findings [6]. Here, we apply the same 3D CNN structure that was found the best in our earlier study [5]. The input of this network consists of a 25-frame block of the ultrasound video, and the output is a mel-spectral vector. For more details on this network, we refer the reader to our earlier study.

A crucial parameter of DNN training is the loss function which formalizes the difference between the training targets and the DNN output. In regression tasks, the simplest solution is to use general loss functions such as the mean-squared error (MSE). However, such mathematically motivated loss functions may not conform with our perception. When the network output is a speech signal, such as in speech enhancement, one may try to apply objective speech quality metrics as the loss function [13,14]. Here, we explore another method where a second, discriminator network is trained to judge the quality of the output signal. This approach leads to the subject of Generative Adversarial Networks (GANs).

GANs were first proposed by Goodfellow et al. [20]. A GAN is basically a combination of two networks, a generator and a discriminator. The goal of the generator is to artificially generate data samples as if they were coming from a real distribution, for example, as if they were real images. The distribution to be approximated is not mathematically defined, but it is estimated by the other network, the discriminator. For this purpose, the discriminator is trained to separate real data vectors from fake ones (i.e. those created by the generator), using real and fake data items. The generator and the discriminator are normally trained in parallel, in a sort of minimax game: the discriminator seeks to improve in differentiating real data items from fake ones, so its error is minimized, while the generator tries to generate fake images that are very similar to the real ones, so it seeks to increase the error of the discriminator. This system is called adversarial, as the generator and the discriminator work against each other.

When the goal is not data generation but data transformation, we can modify the original GAN formulas by adding the input to be transformed as a
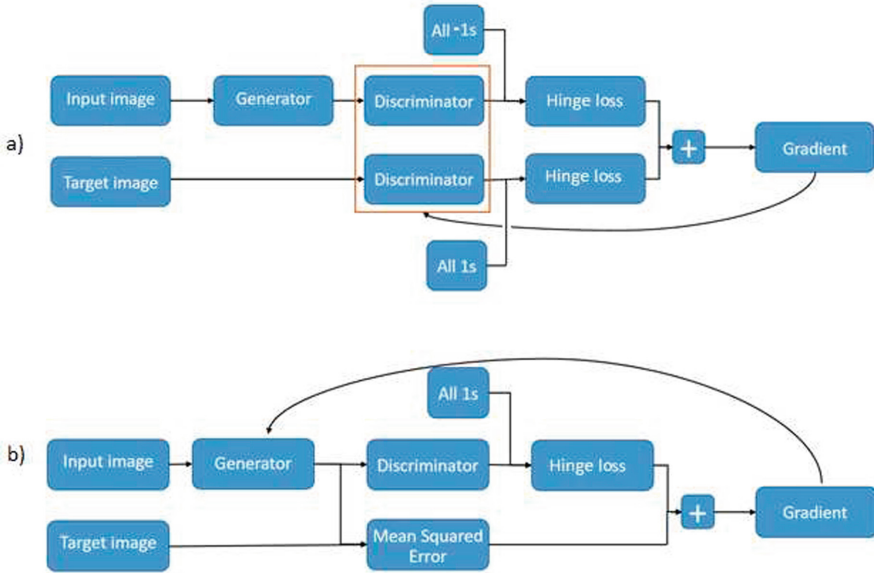
Fig. 2. The error calculation (forward arrows) and weight update (backward arrows) training steps for the discriminator (upper image) and the generator (lower image) networks of the GAN.

'condition' of the data distribution, and hence this model is called the conditional GAN or CGAN [21]. GANs and CGANs were shows to produce very good quality images in broad types of vision tasks like image generation [20], image translation [22,23], and text-to-image synthesis [24]. In speech technology, the GAN-based approach is the most successful in speech enhancement [25] and voice conversion [26].

In our case, the role of the generator was played by the network that converts the ultrasound data to mel-spectral data. For this purpose we used the same network as in our previous study [5]. As the baseline system, we trained this network conventionally, using the MSE loss function. Then we created a discriminator which was trained to discriminate real mel-spectrograms from those produced by the generator. As was explained earlier, the generator was trained in an adversarial manner, so it was optimized to create spectrograms that cannot be discriminated from real spectrograms by the discriminator.

The generator and the discriminator was trained in parallel, using a two-step process [25]. As shown in Fig. 2a, the first step trains the discriminator on two types of images: on real spectrograms (lower data path), and on spectrograms created by the generator (upper path). These serve as positive and negative training examples, denoted by the target labels of 1 and −1 in the figure. For the training, we applied the hinge loss function, which is frequently used as a GAN objective [27], and optimized it using the Adam optimizer with a learning

Table 1. The layers of the discriminator network (following Keras' terminology). Albeit not shown, each but the last Conv2D layer is followed by batch normalization.

Layer	Filters	Size	Strides	Padding	Activation
Conv2D	64	(4, 4)	(2, 2)	Same	Relu
Conv2D	128	(4, 4)	(2, 2)	Same	Relu
Conv2D	256	(4, 4)	(2, 2)	Same	Relu
ZeroPadding2D	–	–	–	–	–
Conv2D	512	(2, 2)	(1, 1)	Valid	Relu
ZeroPadding2D	–	–	–	–	–
Conv2D	1	(4, 4)	(1, 1)	Valid	tanh

rate of 0.0002. In this step, only the weights of the discriminator are updated, while the generator weights are frozen.

Figure 2b shows the other training step that updates the generator weights (now the discriminator weights are frozen). We combine two error functions to calculate the loss, and hence the gradient. First, we compare the generator output and the target spectrograms using the MSE loss (lower path). The second loss value is obtained from the discriminator (upper path). Notice that now the discriminator target labels are flipped, as we wish to train the generator in an adversarial manner, to create outputs that look like a real spectrogram.

4   Experimental Set-Up

The goal of the experiments was to compare the performance of the 3D CNN generator network trained conventionally, using the MSE loss function, with the training scheme that applies the GAN-style discriminator network. In the experiments we evaluated our models on two data sets – one of them being recorded from a Hungarian speaker and the other from an English speaker. Here, we shall present the technical details of the experiments.

4.1   Data Sets and Data Preprocessing

Hungarian Data Set: The parallel ultrasound and speech recordings were collected from a Hungarian female subject reading sentences aloud, using the equipment briefly described in Sect. 2. The whole duration of the recordings was about half an hour (438 sentences), from which 310 were used for training, 41 for development and 87 for testing, respectively.

The ultrasound transducer produces an ultrasound video of the tongue movement at a rate of 82 frames per second. One frame of this video has a resolution of 64×946; that is, the device collects 946 data samples along 64 scan lines. As these ultrasound images are very noisy and they contain very few details, we

Table 2. The MSE and R^2 scores of the generator with the MSE and GAN training approaches.

Training method	Hungarian corpus				English corpus			
	Dev		Test		Dev		Test	
	MSE	Mean R^2	MSE	Mean R^2	MSE	Mean R^2	MSE	Mean R^2
MSE	0.327	0.68	0.326	0.677	0.316	0.683	0.317	0.68
GAN	0.287	0.719	0.29	0.713	0.287	0.715	0.293	0.71

decreased the image size to 64×128 by applying a bicubic interpolation. The pixel intensities were min-max scaled to the $[-1, 1]$ range.

The speech signals were recorded in parallel with the ultrasound video at a sampling rate 22050 Hz. The speech and ultrasound signals were synchronized using the software tool provided by Articulate Instruments.

English Data Set: As the English data set, we used the TAL corpus [28]. It contains parallel speech, tongue ultrasound and lip video recordings from 81 speakers. Here, we just used the TAL1 subset of the corpus, which contains the recordings of a single trained native English speaker. The recording conditions were very similar to that of the Hungarian data set and, after division, the train, validation and test sets contained 1015, 50 and 24 utterances, respectively. We applied exactly the same preprocessing steps as for the Hungarian data.

4.2   DNN Configuration and Training

As the generator we used the 3D CNN from our earlier study [5], with the slight modification that instead of one target vector, here we specified 5 consecutive spectral vectors as targets. We hoped this modification would improve the performance of the discriminator.

As the discriminator, we applied the so-called Patch-GAN method [22,23]. Instead of a two-class decision for the whole image, the discriminator of a patch-GAN returns a set of output values for different regions of the image. We applied a fully convolutional CNN for this purpose, with an output vector of 10 components. The actual network configuration and its parameters are shown in Table 1.

When synchronized training targets are available, as in our case, adversarial training can be combined with conventional MSE-training. We got the best results when the MSE loss function was combined with the adversarial training loss using a weighting ratio of 0.75-0.25 for the two loss functions, respectively.

5   Results and Discussion

Evaluating the quality of the generator is not trivial, as its output is a spectrogram, which is converted to a speech signal by WaveGlow. Our main goal is to increase the quality of the synthesized speech, and we have two ways to measure this quality. First, we can perform subjective listening tests such as MUSHRA [29]. Unfortunately, this would require a lot of human subjects, hence

Table 3. Objective quality scores for the two training methods and the two corpora.

Corpus and training method	STOI	ESTOI	PESQ	SISDR	SDR	PMSQE	MCD
Hun - MSE	0.7050	0.456	1.282	−39.363	−18.021	2.797	4.627
Hun - GAN	**0.7067**	**0.4673**	**1.311**	**−37.868**	**−16.893**	**2.777**	**4.558**
Eng - MSE	0.612	0.385	1.464	−36.176	−18.31	2.827	3.38
Eng - GAN	**0.623**	**0.405**	**1.503**	**−36.031**	**−17.672**	**2.735**	**3.229**

it would be slow and troublesome, which we wished to avoid here. Instead, we can evaluate objective, formally defined metrics. As the simplest of these, we report two metrics that are popular in machine learning for regression tasks, namely the mean squared error and the correlation-based mean R^2 score (the former has to be minimized, while the latter should be maximized). As can be seen in Table 2, all these metrics gave a slight improvement when the GAN-style training was applied, both for the Hungarian and English subjects, and both for the development and the test sets.

As we mentioned in Sect. 1, our output is a speech signal, and in this case the above simple metrics may not perfectly reflect human perception. Fortunately, several objective measures have been developed in speech technology and in telecommunications to compare the quality or intelligibility of speech recordings. These measures include the Short-Time Objective Intelligibility (STOI) [14] and its extended version (ESTOI), the perceptual evaluation of speech quality (PESQ) method [14], and its extended version known as perceptual metric for speech quality evaluation (PMSQE) [14]. While the first three measure quality, and hence a higher value means a better performance, PMSQE was designed to be applicable as a loss function in DNN training, so in this case a lower value indicates better quality. We also evaluated the signal-to-distortion ratio (SDR) and its extended, scale invariant version (SI-SDR) [30]. As their name suggests, for these metrics a higher value means better quality. Finally, we calculated the mel-cepstral distortion (MCD) [14], which is a distortion metric, so it should be minimized to increase speech quality. These results are summarized in Table 3 for the test sets of the two databases. To aid readability, the two metrics that are to be minimized were placed in the rightmost columns, and the best scores are highlighted for both corpora. As the results indicate, extending the MSE training criterion with GAN-style adversarial training led to a consistent improvement in all the evaluated metrics and for both corpora. Although in certain cases the improvement is only slight, the results clearly justify the utility of generative adversarial networks for the articulatory-to-acoustics mapping task. In comparison, Ribeiro et al. reported an MCD score of 2.99 for the English dataset using a much more sophisticated endocer-decoder neural architecture [28].

6   Conclusions

The application of the GAN framework has already proved successful in speech enhancement and voice conversion tasks, and here we made the first attempts to apply it to the articulatory-to-acoustic mapping task of ultrasound-based silent speech interfaces. As the baseline, we trained the generator network conventionally, using the MSE loss, and then we extended its training with adversarial training by means of a discriminator. We applied our method on two data sets, a Hungarian corpus and an English corpus, and in both cases we found that the quality of the generated speech signals improved, according to several objective speech quality metrics. In the future, we plan to extend our framework to the direct generation of speech, such as by using MelGAN [27], and we also plan to study the effect of incorporating perceptual speech quality metrics into the training loss [14].

References

1. Schultz, T., Wand, M., Hueber, T., Krusienski, D.J., Herff, C., Brumberg, J.S.: Biosignal-based spoken communication: a survey. IEEE/ACM Trans. ASLP **25**(12), 2257–2271 (2017)
2. Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J.M., Brumberg, J.S.: Silent speech interfaces. Speech Commun. **52**(4), 270–287 (2010)
3. Kim, M., Cao, B., Mau, T., Wang, J.: Speaker-independent silent speech recognition from flesh-point articulatory movements using an LSTM neural network. IEEE/ACM Trans. ASLP **25**(12), 2323–2336 (2017)
4. Taguchi, F., Kaburagi, T.: Articulatory-to-speech conversion using bi-directional long short-term memory. In: Proceedings of Interspeech, pp. 2499–2503 (2018)
5. Tóth, L., Shandiz, A.H.: 3D convolutional neural networks for ultrasound-based silent speech interfaces. In: Rutkowski, L., Scherer, R., Korytkowski, M., Pedrycz, W., Tadeusiewicz, R., Zurada, J. (eds.) ICAISC 2020. LNCS, vol. 12415, pp. 159–169. Springer, Heidelberg (2020)
6. Saha, P., Liu, Y., Gick, B., Fels, S.: Ultra2speech – a deep learning framework for formant frequency estimation and tracking from ultrasound tongue images. In: Martel, A., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M., Kevin Zhou, S., Racoceanu, D., Joskowicz, L. (eds.) MICCAI 2020. LNCS, vol. 12263, pp. 473–482. Springer, Heidelberg (2020)
7. Hueber, T., Benaroya, E.L., Chollet, G., Dreyfus, G., Stone, M.: Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips. Speech Commun. **52**(4), 288–300 (2010)

8. Gonzalez, J.A., Cheah, L.A., Gomez, A.M., Green, P.D., Gilbert, J.M., Ell, S.R., Moore, R.K., Holdsworth, E.: Direct speech reconstruction from articulatory sensor data by machine learning. IEEE/ACM Trans. ASLP **25**(12), 2362–2374 (2017)
9. Fagan, M.J., Ell, S.R., Gilbert, J.M., Sarrazin, E., Chapman, P.M.: Development of a (silent) speech recognition system for patients following laryngectomy. Med. Eng. Phys. **30**(4), 419–425 (2008)
10. Wand, M., Schultz, T., Schmidhuber, J.: Domain-adversarial training for session independent EMG-based speech recognition. In: Proceedings of Interspeech, pp. 3167–3171 (2018)
11. Janke, M., Diener, L.: EMG-to-speech: direct generation of speech from facial electromyographic signals. IEEE/ACM Trans. ASLP **25**(12), 2375–2385 (2017)
12. Wand, M., Koutník, J., Schmidhuber, J.: Lipreading with long short-term memory. In: Proceedings of ICASSP, pp. 6115–6119. IEEE (2016)
13. Zhao, Y., Xu, B., Giri, R., Zhang, T.: Perceptually guided speech enhancement using deep neural networks. In: Proceedings of ICASSP, pp. 5074–5078 (2018)
14. Martín-Doñas, J., Gomez, A., Gonzalez Lopez, J., Peinado, A.: A deep learning loss function based on the perceptual evaluation of the speech quality. IEEE Signal Process. Lett. **25**(11), 1680–1684 (2018)
15. Pihlgren, G.G., Sandin, F., Liwicki, M.: Improving image autoencoder embeddings with perceptual loss. In: Proceedings of IJCNN, pp. 1–7. IEEE (2020)
16. Dosovitskiy, A., Brox, T.: Generating images with perceptual similarity metrics based on deep networks. arXiv preprint arXiv:1602.02644 (2016)
17. Prenger, R., Valle, R., Catanzaro, B.: WaveGlow: a flow based generative network for speech synthesis. In: Proceedings of ICASSP, pp. 3617–3621 (2019)
18. Govalkar, P., Fisher, J., Zalkov, F., Dittmar, C.: A comparison of recent neural vocoders for speech signal reconstruction. In: Proceedings of ISCA Speech Synthesis Workshop (2019)
19. Csapó, T.G., Zainkó, C., Tóth, L., Gosztolya, G., Markó, A.: Ultrasound-based articulatory-to-acoustic mapping with WaveGlow speech synthesis. In: Proceedings of Interspeech 2020, pp. 2727–2731 (2020)
20. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. arXiv preprint arXiv:1406.2661 (2014)
21. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
22. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of Conference on Computer Vision and Pattern Recognition, pp. 1125–1134 (2017)
23. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of International Conference on Computer Vision, pp. 2223–2232 (2017)
24. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: International Conference on Machine Learning, pp. 1060–1069 (2016)
25. Zhang, Z., Deng, C., Shen, Y., Williamson, D.S., Sha, Y., Zhang, Y., Song, H., Li, X.: On loss functions and recurrency training for GAN-based speech enhancement systems. In: Proceedings of Interspeech, pp. 3266–3270 (2020)
26. Kaneko, T., Kameoka, H., Tanaka, K., Hojo, N.: StarGAN-VC2: rethinking conditional methods for StarGAN-based voice conversion. In: Proceedings of Interspeech, pp. 679–683 (2019)

27. Kumar, K., Kumar, R., de Boissiere, T., Gestin, L., Teoh, W.Z., Sotelo, J., de Brébisson, A., Bengio, Y., Courville, A.C.: MelGAN: generative adversarial networks for conditional waveform synthesis. Adv. Neural. Inf. Process. Syst. **32**, 14910–14921 (2019)
28. Ribeiro, M.S., Sanger, J., Zhang, J.X., Eshky, A., Wrench, A., Richmond, K., Renals, S.: TaL: a synchronised multi-speaker corpus of ultrasound tongue imaging, audio, and lip videos. arXiv preprint arXiv:2011.09804 (2020)
29. ITU-R: ITU-R recommendation BS.1534: method for the subjective assessment of intermediate audio quality (2001)
30. Roux, J.L., Wisdom, S., Erdogan, H., Hershey, J.R.: SDR - half-baked or well done? In: Proceedings of ICASSP (2019)