U-GAT-VC: UNSUPERVISED GENERATIVE ATTENTIONAL NETWORKS FOR NON-PARALLEL VOICE CONVERSION

Sheng Shi[12]*, Jiahao Shao*[3]*, Yifei Hao*[4]*, Yangzhou Du*[2]*, Jianping Fan*[12]

[1]Northwest University, Xi'an 710127, P. R. China
[2]AI Lab, Lenovo Research, Beijing 100094, P. R. China
[3]Tsinghua University, Beijing, 100084, P. R. China
[4]University of Southern California, Los Angeles, 90007, CA

ABSTRACT

Non-parallel voice conversion (VC) is a technique of transferring voice from one style to another without using a parallel corpus in model training. Various methods are proposed to approach non-parallel VC using deep neural networks. Among them, CycleGAN-VC and its variants have been widely accepted as benchmark methods. However, there is still a gap to bridge between the real target and converted voice and an increased number of parameters leads to slow convergence in training process. Inspired by recent advancements in unsupervised image translation, we propose a new end-to-end unsupervised framework U-GAT-VC that adopts a novel inter- and intra-attention mechanism to guide the voice conversion to focus on more important regions in spectrograms. We also introduce disentangle perceptual loss in our model to capture high-level spectral features. Subjective and objective evaluations shows our proposed model outperforms CycleGAN-VC2/3 in terms of conversion quality and voice naturalness.

Index Terms— Non-parallel Voice Conversion, Generative Adversarial Network, Inter attention mechanism, Intra attention mechanism, Perceptual loss

1. INTRODUCTION

Voice conversion (VC) is a technique of transferring voice from one style to another while keeping linguistic information unchanged. It has a great potential for various tasks, such as speaking assistance [1], emotional tone conversion [2], and accent conversion. VC methods could be categorized as parallel and non-parallel ones. Parallel VC relies on the availability of temporally-aligned parallel utterance pairs of source and target speech. In this paper, we focus on non-parallel VC that does not rely on a parallel corpus.

Non-parallel VC is challenging owing to the absence of explicit supervision. Earlier methods utilize linguistic information to assist model training [3] [4]. However, deriving such linguistic information needs additional modules and data, which introduces additional costs and limits their application. To avoid such a requirement and achieve non-parallel VC using only acoustic data, deep generative models, such as VAE-based methods [5] and GAN-based methods [6] [7] have been proposed recently. Among them, CycleGAN-VC [7] and its variants, such as CycleGAN-VC2 [8] and StarGAN-VCs [9], have been widely accepted as benchmark methods [10]. However, their application is constrained to mel-cepstrum conversion as their insufficient capacity to capture the time-frequency structure that should be preserved during mel-spectrogram conversion. To overcome these limitations, CycleGAN-VC3 [11] incorporates time-frequency adaptive normalization (TFAN) for mel-spectrogram conversion. However, there is still a gap between real targets and converted speeches. An increased number of model parameters also leads to the slow convergence during model training.

In this paper, we propose a new end-to-end unsupervised framework U-GAT-VC, the main contribution is summarized as three folds:

- An inter- and intra-attention module is proposed to guide the network to capture important information in mel-spectrograms.
- A disentangled perceptual loss is introduced to optimize the generator to obtain high quality synthesized speeches when converting.
- An end-to-end framework is proposed to solve non-parallel VC with faster training time using mel-spectrograms as intermediate input data. Both subjective and objective experimental results show that the proposed method improves the conversion quality compared to CycleGAN-VCs.

2. CYCLEGAN-VCS

CycleGAN-VC [7] is first proposed incorporating the popular CycleGAN [12] architecture to approach voice conversion. This section briefly reviews the framework of CycleGAN-VC3, since it is the latest updated version.

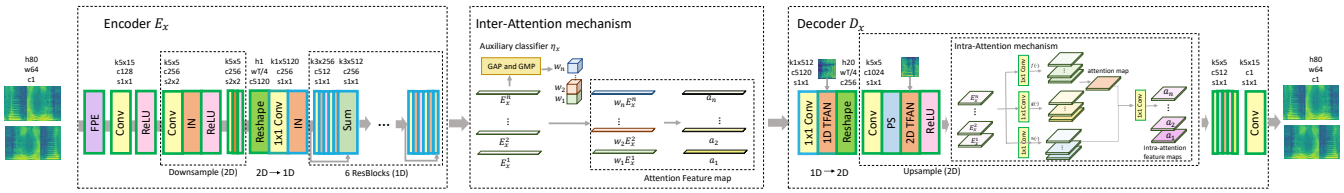


Fig. 1. Network architectures of our generator. h, w, and c denote height, width, and number of channels, respectively. In each convolution layer, k, c, and s denote kernel size, number of channels, and stride size, respectively. FPE, IN, ReLU, and PS indicate frame position embedding, instance normalization, rectified linear units, and pixel shuffler, respectively.

2.1. Training objectives

CylceGAN-VC3 aims at learning a mapping $y = G_{X\to Y}(x)$ from source acoustic features $x \in X$ to target acoustic features $y \in Y$ using non-parallel corpus. To achieve such a goal, it incorporates adversarial loss, cycle-consistency loss [12], second adversarial loss [8] and identity-mapping loss [12]: Adversarial loss is formulated as: $\mathcal{L}_{adv}^{X\to Y} = \mathbb{E}_{y\sim P_Y}[\log D_Y(y)] + \mathbb{E}_{x\sim P_X}[\log(1 - D_Y(G_{X\to Y}(x))]$. The generator $G_{X\to Y}(x)$ attempts to generate a feature that tries to deceive the discriminator D_Y, whereas D_Y attempts to find the best decision boundary of real and converted features by maximizing the adversarial loss. $\mathcal{L}_{adv}^{Y\to X}$ is imposed on the inverse-forward mapping. The adversarial loss is further written as $\mathcal{L}_{adv} = \mathcal{L}_{adv}^{X\to Y} + \mathcal{L}_{adv}^{Y\to X}$. To guarantee the linguistic consistency in conversion, cycle-consistency loss is used as: $\mathcal{L}_{cyc} = \mathbb{E}_{x\sim P_X}[\|G_{Y\to X}(G_{X\to Y}(x)) - x\|_1] + \mathbb{E}_{y\sim P_Y}[\|G_{X\to Y}(G_{Y\to X}(y)) - y\|_1]$. The loss encourages $G_{X\to Y}$ and $G_{Y\to X}$ to identify optimal pseudo pairs through circular conversion. To mitigate the over-smoothing problem caused by L1 loss, CycleGAN-VC2 brings in the additional discriminator $D_X^{'}$ and $D_Y^{'}$ [8]. A second adversarial loss $\mathcal{L}_{adv2}^{X\to Y\to X}$ is imposed on the circularly converted features as: $\mathcal{L}_{adv2}^{X\to Y\to X} = \mathbb{E}_{x\sim P_X}[\log D_X^{'}(x)] + \mathbb{E}_{x\sim P_X}[\log(1 - D_X^{'}(G_{Y\to X}(G_{X\to Y}(x))))]$. The second adversarial loss is written as: $\mathcal{L}_{adv2} = \mathcal{L}_{adv2}^{X\to Y\to X} + \mathcal{L}_{adv2}^{Y\to X\to Y}$. To further encourage the input preservation, identity-mapping loss is formulated as: $\mathcal{L}_{id} = \mathbb{E}_{x\sim P_X}[\|G_{X\to Y}(x) - x\|_1] + \mathbb{E}_{y\sim P_Y}[\|G_{Y\to X}(y) - x\|_1]$. Full training loss is a combination of the four losses above with trade-off parameters λ_{cyc}, λ_{adv2} and λ_{id}, which is defined as: $\mathcal{L}_{full} = \mathcal{L}_{adv} + \lambda_{cyc}\mathcal{L}_{cyc} + \lambda_{adv2}\mathcal{L}_{adv2} + \lambda_{id}\mathcal{L}_{id}$. which the model tries to optimize during training.

2.2. Generator and Discriminator architectures

CycleGAN-VC3 follows *2-1-2D CNN generators* which uses *2D CNNs* in upsampling and downsampling blocks and use *1D CNNs* in residual blocks. CycleGAN-VC3 designed TFAN for 1D and 2D time-frequency features in 2-1-2D CNN [11]. Moreover, CycleGAN-VC3 adapts *PatchGAN* uses convolution at the last layer instead of *FullGAN* which uses a fully-connected layer to stabilize model training [8].

3. U-GAT-VC

3.1. Inter-attention mechanism

Timbre perceived as the voice characteristics of a speaker are reflected by the formant frequencies. In a mel-spectrogram, the formants are shown as the salient frequency components of the spectral envelope. Inspired by the attention module of U-GAT-IT [13], we use the class activation maps (CAM) derived by the auxiliary classifier as attention map (the weight of feature map) in mel-spectrograms to distinguish source domain and target domain, which could encourage the decoder to convert to target domain efficiency. By embedding the auxiliary classifier into generator, our model will guide the conversion to focus on more important regions and ignore minor regions.

The conversion model $G_{X\to Y}$ consists of an encoder E_X, a decoder D_X, and an auxiliary classifier η_X, where $\eta_X(x)$ represents the probability that x comes from domain X. The auxiliary classifier is trained to learn the weight of the $k-th$ feature map for the source domain, ω_X^k, by using the global average pooling (GAP) and global max pooling (GMP), i.e.,

$$\eta_X(x) = \sigma(\sum_k \omega_X^k \sum_{i,j} E_X^{k_{i,j}}(x)), \tag{1}$$

where $E_X^{k_{i,j}}(x)$ is the value of the $k-th$ activation map of the encoder at (i, j). By exploiting ω_X^k, we can calculate a set of domain specific attention feature map by multiplying attention map and feature value:

$$a_X(x) = \omega_X * E_X(x) = \{\omega_X^k * {E_X}^k(x) | 1 \le k \le n\}. \tag{2}$$

Then, the conversion model $G_{X\to Y}$ could catch the regions where it needs to improve on the current state by exploiting the information from the auxiliary classifier η_X:

$$L_{cam}^{X\to Y} = -(\mathbb{E}_{x\sim P_X}[\log \eta_X(x)] + \mathbb{E}_{y\sim P_Y}[\log(1 - \eta_X(y))]. \tag{3}$$

Similarly, $L_{cam}^{Y\to X}$ is imposed on the circularly converted features as: $L_{cam} = L_{cam}^{X\to Y} + L_{cam}^{Y\to X}$.

3.2. Intra-attention/self-attention mechanism

Mel-spectrogram is a 2D time-frequency representation of the input speech signal. Our model integrates self-attention

mechanism could capture the temporal cues on the spectral features. Moreover, As CycleGAN-VC3 incorporated TFAN module leads to the slow convergence in model training, self-attention mechanism could improve training efficiency.

The feature map from previous layer x is first transformed into feature embeddings with three feature embedding functions $f(queries)$, $g(keys)$, and $h(values)$. An attention matrix $\omega \in \mathbb{R}^{N\times N}$ is computed as: $\omega_{i,j} = \frac{\exp\left(f(x_i)^\top g(x_j)\right)}{\sum_{i=1}^{N}\exp\left(f(x_i)^\top g(x_j)\right)}$, where element $\omega_{i,j}$ represents the influence of the j^{th} feature point on the i^{th} feature point. The weighted self-attention map a is therefore defined as: $a = v(x)(h(x)\omega^\top)$, where v is the output embedding function. For the final output of the self-attention module, a is multiplied by a scale parameter φ and a residual design is also implied: $y = \varphi a + x$. To apply a progressive learning strategy, φ is initially set to 0. That is, the module first captures the feature associations with only local cues and gradually learns to inspect long-range information of mel-spectrograms.

3.3. Disentangled Perceptual loss

A voice conversion model needs to perform well from two aspects: it needs to transfer non-linguistic information (i.e. style) from source to target. On the other hand, it needs to preserve the linguistic information (i.e. content) as source. One intuition is to explicitly guide the network to follow these two rules at training stage. Therefore we incorporate an additional disentangle perceptual loss into the network training procedure to achieve this goal.

First proposed in the image translation [14], a perceptual loss is defined with two parts: *content loss* is defined as the squared Euclidean distance between feature representations: $\ell_{cont}^{\phi,j}(y,x) = \frac{1}{C_j H_j W_j}\|\phi_j(G(x)) - \phi_j(x)\|_2^2$, where ϕ_j is the j^{th} activations of the perceptual network ϕ with respect to the input. C_j, H_j, W_j denotes the channels, heights and width of the j^{th} feature maps of ϕ, respectively. A *style loss* is proposed [14] by first defining a Gram matrix, which effectively captures the associations of features activating together. Then the *style loss* is computed as the squared Frobenius norm of the differences between the Gram matrices of the output and target mel-spectrograms: $\ell_{sty}^{\phi,j}(y,y') = \|Gram_j^\phi(G(x)) - Gram_j^\phi(y)\|_2^2$.

Inspired by the work in [15], a four-layer perceptual network ϕ is embedded in the first four layers of the discriminator D. The third layer of ϕ is used to extract content representations and all four layers of ϕ are used when computing style losses. The framework of ϕ as well as the discriminator is shown in figure 2.

Full perceptual loss. As described above, given the content target x, style target y and the generated mel-spectrograms $G(x)$, a *content loss* is computed at the third layer of ϕ, which can be written as: $\mathcal{L}_{cont} = \ell_{cont}^{\phi,3}(G(x),x)$. A *style loss* is computed at all four layers of ϕ, which can be formulated as: $\mathcal{L}_{sty} = \sum_{j=1}^{4}\ell_{sty}^{\phi,j}(G(x),y)$. The full Perceptual loss is defined as the sum of the two sub-losses: $\mathcal{L}_{perc} = \mathcal{L}_{cont} + \mathcal{L}_{sty}$.

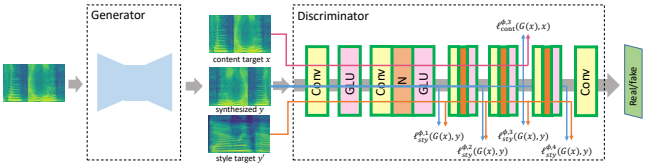


Fig. 2. Framework of our discriminator, the four-layer perceptual network is embedded into the discriminator. The third layer is used to extract Content loss and all four layers are used to compute Style loss

3.4. Others

Considering the time-frequency structure of mel-spectrogram, we also introduce frame position embeddings (FPEs) to help generator capture the overall relationship along with the feature direction while preserving the temporal structure. Moreover, GLUs [16] which used in CycleGAN-VC needs to learn much more parameters which greatly increase the computational complexity. In our model, we replace GLUs whose contribution are overridden by attention mechanism with ReLUs to decrease the number of converter parameters, thus increase convergence rate during model training.

3.5. Model architectures

Updated training objectives. With the newly introduced CAM loss and disentangled perceptual loss, we update the training objective $\mathcal{L}_{full}$ from CycleGAN-VC3 described in section 2 as: $\mathcal{L}'_{full} = \mathcal{L}_{full} + \lambda_{CAM}\mathcal{L}_{CAM} + \lambda_{perc}\mathcal{L}_{perc}$.

Fig 1 shows the architecture of our generator with the following modifications compared with CycleGAN-VC3: we start with an FPE block in the beginning of the encoder. An inter attention module is followed after the feature map of the encoder. At decoding process, we incorporate intra-attention after 2D TFAN since it brings in much more new temporal-structure information. Finally, we replace GLUs with ReLUs for faster convergence.

4. EXPERIMENTS

4.1. Experimental settings

Dataset. We evaluate our model on the non-parallel VC task of VCC 2018 dataset [17]. Considering both inter-gender and intra-gender VC, we select a subset of speakers: VCC2SF3 (*SF*), VCC2SM3 (*SM*), VCC2TF1 (*TF*), and VCC2TM1 (*TM*), where S, T, F and M represent source, target, female, and male, respectively. In our experiments, audio clips were downsampled to 22.05 kHz. We extract an 80-dimensional

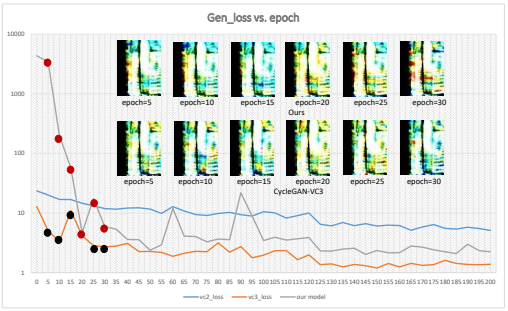


Fig. 3. The generative loss and the visualization of the weighted feature map

log mel-spectrogram with a window length of 1024 and hop length of 256 samples.

Conversion process and training settings. For a fair comparison, we use the same training settings, conversion and synthesis process as CycleGAN-VC3 [11]. For pre-processing, we normalize the mel-spectrograms using the training set statistics. The batch size was set to 1, where each training sample consisted of 64 randomly cropped frames (approximately 0.75s). λ_{cyc}, λ_{id}, λ_{cam} and λ_{prec} were set to 10, 5, 1000 and 1 respectively, and λ_{id} was used for only the first 10k iterations.

4.2. Objective evaluation

We evaluate our model from three perspectives: visualization of the attention maps, objective evaluation metrics (i.e. Frechet DeepSpeech Distance (FDSD) and Kernel DeepSpeech Distance (KDSD) [18]) and convergence rate.

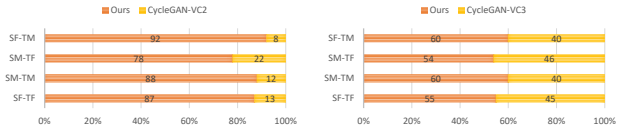
Figure 3 shows the generative loss curve as the number of training epoch increases (only the first 200 epochs is shown, and λ_{cam} is set to 1000). First row lists the attention map of our model for every five epochs (epoch=5,10,...,30), and for comparison, the second row lists weighted feature map, while equal weights are assigned in CycleGAN-VC3. It is notable that our model could quickly capture salient frequency components, which is hard for CycleGAN-VC3.

We used two evaluation metrics FDSD and KDSD. FDSD and KDSD are proposed for speech generation based on FID and KID, where they replace the Inception image recognition network with the DeepSpeech audio recognition network, and are shown to be well correlated with human judgement [19]. For the two metrics, the smaller the value, the better the performance. Table 1 shows our model outperforms both VC2 and VC3 in the inter-gender and intra-gender tasks.

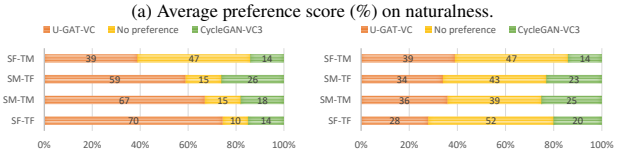
We further conduct a comparative analysis on model parameters. VC3 has a 48% increase in parameters compared with VC2 due to introducing TFAN into the network, yet our model achieves a decrease in parameters with 34.69% compared to VC3 by replacing activation function GLUs with ReLUs. Moreover, as the attention mechanism could guide the module to focus on important regions, our model can reduce the training time by 34.33% under same training settings of VC3.

Table 1. Comparison of FDSD and KDSD

Metric	Method	Intra-gender		Inter-gender	
		SF-TF	SM-TM	SF-TF	SM-TF
FDSD	VC2	9.4114	11.5487	10.4418	10.8824
	VC3	7.2870	10.6996	9.6075	9.7211
	Ours	**6.8689**	**4.8537**	**7.1213**	**8.1470**
KDSD	VC2	0.0084	0.0106	0.0095	0.0105
	VC3	0.0047	0.0096	0.0067	0.0088
	Ours	**0.0033**	**0.0016**	**0.0023**	**0.0050**



(a) Average preference score (%) on naturalness.



(b) Average preference score (%) on speaker similarity.

Fig. 4. Subjective evaluation results

4.3. Subjective evaluation

We conducted listening tests to evaluate the differences in perceptual quality. We investigated the comparative performance between our model and VC2/3 using two choice preference tests. Ten well-educated English speakers are chosen to participate the evaluation. In AB test on naturalness and intelligibility, each listener is presented with two speech samples and asked to choose preferred one ("A" or "B"). In XAB test on speaker similarity, "X" is target speech, each listener is presented with three speech samples and is asked to select their preferred one ("A" or "B" or "No preference"). Ten sentences with length range from 2s and 5s are randomly selected from the evaluation sets. The results are shown in Figure 4. The results indicate that our method outperforms CycleGAN-VC2 and CycleGAN-VC3.

5. CONCLUSIONS

We propose an end-to-end unsupervised framework U-GAT-VC, which incorporates a new inter- and intra-attention mechanism to guide the conversion to focus on more important regions. Also, we introduce disentangled perceptual loss to optimize the network to obtain high quality synthesized speeches when converting. Objective and Subjective evaluation show that our model boosts the model performance of its baseline CycleGAN-VCs while achieving faster convergence.

6. REFERENCES

[1] Alexander B. Kain, John-Paul Hosom, Xiaochuan Niu, Jan P.H. van Santen, Melanie Fried-Oken, and Janice Staehely, "Improving the intelligibility of dysarthric speech," *Speech Communication*, vol. 49, no. 9, pp. 743–759, 2007.

[2] Zeynep Inanoglu and Steve Young, "Data-driven emotion conversion in spoken english," *Speech Communication*, vol. 51, no. 3, pp. 268–283, 2009.

[3] Athanasios Mouchtaris, Jan Van Spiegel, Paul Mueller, Athanasios Mouchtaris, Jan Van Der Spiegel, and Paul Mueller, "Nonparallel training for voice conversion based on a parameter adaptation approach," *IEEE Trans. Audio, Speech and Language Processing*, pp. 952–963, 2006.

[4] Feng-Long Xie, Frank Soong, and Haifeng Li, "A kl divergence and dnn-based approach to voice conversion without parallel training sentences," 09 2016, pp. 287–291.

[5] Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," 2016.

[6] Takuhiro Kaneko and Hirokazu Kameoka, "Parallel-data-free voice conversion using cycle-consistent adversarial networks," 2017.

[7] Takuhiro Kaneko and Hirokazu Kameoka, "Parallel-Data-Free Voice Conversion Using Cycle-Consistent Adversarial Networks," *arXiv e-prints*, Nov. 2017.

[8] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo, "Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019.

[9] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo, "Stargan-vc: Non-parallel many-to-many voice conversion with star generative adversarial networks," *CoRR*, vol. abs/1806.02169, 2018.

[10] J. Zhang, Z. Ling, and L. Dai, "Non-parallel sequence-to-sequence voice conversion with disentangled linguistic and speaker representations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 540–552, 2020.

[11] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo, "Cyclegan-vc3: Examining and improving cyclegan-vcs for mel-spectrogram conversion," in *Proceedings of the Annual Conference of the International Speech Communication Association*, 2020.

[12] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2242–2251.

[13] Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwanghee Lee, "U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation," 2020.

[14] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*, 2016.

[15] Yan Zhang Yang Wang Yanping Li, Dongxiang Xu and Binbin Chen, "Non-parallel many-to-many voice conversion with psr-stargan," in *Interspeech2020*, 2020.

[16] Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier, "Language modeling with gated convolutional networks," *CoRR*, vol. abs/1612.08083, 2016.

[17] Jaime Lorenzo-Trueba, Junichi Yamagishi, Tomoki Toda, Daisuke Saito, Fernando Villavicencio, Tomi Kinnunen, and Zhenhua Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," 06 2018, pp. 195–202.

[18] Mikołaj Bińkowski, Jeff Donahue, Sander Dieleman, Aidan Clark, Erich Elsen, Norman Casagrande, Luis C. Cobo, and Karen Simonyan, "High fidelity speech synthesis with adversarial networks," 2019.

[19] Dario Amodei, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Erich Elsen, Jesse Engel, Linxi Fan, Christopher Fougner, Tony Han, Awni Hannun, Billy Jun, Patrick LeGresley, Libby Lin, Sharan Narang, Andrew Ng, Sherjil Ozair, Ryan Prenger, Jonathan Raiman, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Yi Wang, Zhiqian Wang, Chong Wang, Bo Xiao, Dani Yogatama, Jun Zhan, and Zhenyao Zhu, "Deep speech 2: End-to-end speech recognition in english and mandarin," 2015.