GENERATIVE ADVERSARIAL NETWORK BASED SPEAKER ADAPTATION FOR HIGH FIDELITY WAVENET VOCODER

Qiao Tian, Bing Yang, Jing Chen, Benlai Tang, Shan Liu

Cloud and Smart Industries Group, Tencent Technology Co., Ltd

{briantian, bjarneyang, kevincjchen, benlaitang, shiningliu}@tencent.com

ABSTRACT

Neural networks based vocoders, typically the WaveNet, have achieved spectacular performance for text-to-speech (TTS) in recent years. Although state-of-the-art parallel WaveNet has addressed the issue of real-time waveform generation, there remains problems. Firstly, due to the noisy input signal of the model, there is still a gap between the quality of generated and natural waveforms. Secondly, a parallel WaveNet is trained under a distilled training framework, which makes it tedious to adapt a well trained model to a new speaker. To address these two problems, this paper proposes an end-to-end adaptation method based on the generative adversarial network (GAN), which can reduce the computational cost for the training of new speaker adaptation. Our subjective experiments shows that the proposed training method can further reduce the quality gap between generated and natural waveforms.

***Index Terms*—** Neural Vocoder, Parallel WaveNet, Speaker Adaptation, Generative Adversarial Network

1. INTRODUCTION

Recently, deep learning has made successful progress in the field of speech synthesis. The state-of-the-art approach Tacotron2 [1], which used an end-to-end acoustic model and neural vocoder of WaveNet [2], has achieved high fidelity synthesized audio. Compared with the conventional methods, e.g. long short term memory (LSTM) based statistical parametric speech synthesis [3] using traditional vocoders [4, 5], this approach makes the synthesized speech greatly closer towards natural speech in both speech quality and prosody.

Despite of the high fidelity of speech generated by the WaveNet, there are two issues still need to be addressed in practical applications. The parellel WaveNet [6] is therefore proposed for real-time generation of speech, because of the enormous computational cost of original auto-regressive WaveNet. However, those issues still remains. Firstly, the training pipeline of the parallel WaveNet is relatively tedious since the parallel WaveNet is learned following the model distillation framework [7], a well learned auto-regressive WaveNet is required as the teacher model to guide the training of parallel WaveNet, which is the student model. The Both teacher and student models could take days. In addition, sufficient data are required to train teacher model for the new speaker. Secondly, although generated speech is of good quality, there are still a gap between generated and natural speech. This is due to the input noisy of signal of the parallel WaveNet model. Lots of detailed information are missed in high frequency domain of generated speech.

In this paper, we propose an adaptation framework to adapt a well learned parallel WaveNet to a speaker with few hours of training data. We replaced the distillation component in training framework with a generative adversarial component [8]. The minimax training trick of generative adversarial network (GAN) makes the generated samples undistinguishable from real samples. The contribution of this paper includes: 1) We proposed an end-to-end speaker adaptation for high fidelity neural vocoder based on GAN. The training of the proposed framework is much more efficient than the distillation framework, such as parallel WaveNet and Clarinet [9]. 2) We used the GAN to reduce the gap between generated and natural speech. The discriminator of the GAN can capture the subtle difference between generated waveform and natural ones which is usually neglected by the auto-regressive teacher WaveNet.

This paper is organized as follows: Section 2 will briefly review the basic background of GAN. The proposed method will be given in Section 3. Experimental details and results will be given in Section 4. Lastly in Section 5, some conclusions and potential future research are given.

2. GENERATIVE ADVERSARIAL NETWORK

Generative Adversarial Network is a new framework proposed in recent years, which has been proven to generate impressive samples in the field of computation vision. As showed in Fig. 1, an typical GAN model consists of two sub-networks: a Discriminator network (D) and a Generator network (G). The generator network learns to map a simple noise distribution $p_{\boldsymbol{z}}(\boldsymbol{z})$ to a complex distribution $P_g(\boldsymbol{x})$, where $\boldsymbol{z}$ is the random noise sample and $\boldsymbol{x}$ is the data sam-

Paper submitted to IEEE ICASSP 2019

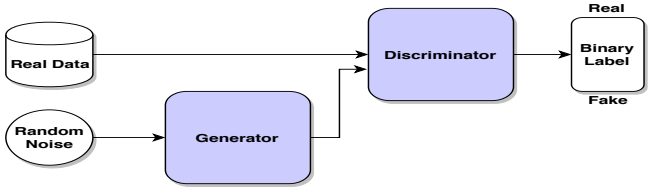


Fig. 1. The architecture of generative adversarial network

ple. The generator is trained to make the generated sample distribution $P_g(\boldsymbol{x})$ undistinguishable from real data distribution $P_d(\boldsymbol{x})$. On the contrary, the discriminator is trained to identify the generated (fake) samples and data (real) samples. Therefore, the adversarial training method plays like a minimax game. For conditional sample generation tasks, such as speech synthesis, the additional condition vector $\boldsymbol{c}$ is usually inputted to both the generator and discriminator, which make the model conditional GAN (cGAN) [10].. The training objective of the cGAN is formulated as

$$\begin{aligned}\min_G \max_D \; V(D,G) &= E_{\boldsymbol{x},\boldsymbol{c}\sim p_d(\boldsymbol{x},\boldsymbol{c})}\left[\log D(\boldsymbol{x},\boldsymbol{c})\right] \\ &+ E_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z}),c\sim p_D(\boldsymbol{c})}\left[\log(1-D(G(\boldsymbol{z},\boldsymbol{c}),\boldsymbol{c})\right]. \quad (1)\end{aligned}$$

The minimax training objective of the original GAN is unstable, difficult to converge, and usually results in mode collapse problem. A lot of tricks are therefore proposed to improve the training and ensure model to learn realistic distribution. In order to address the problem of vanishing gradients caused by sigmoid cross-entropy loss, the least-squares GAN (LSGAN) [11], is proposed by replacing the cross-entropy loss with the least-squares binary coding loss. The training objective for the discriminator of the LSGAN is defined as

$$\begin{aligned}\min_D \; V_{\mathrm{LSGAN}}(D) &= \frac{1}{2}E_{\boldsymbol{x},\boldsymbol{c}\sim p_d(\boldsymbol{x},\boldsymbol{c})}\left[(D(\boldsymbol{x},\boldsymbol{c})-1)^2\right] \\ &+ \frac{1}{2}E_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z}),\boldsymbol{c}\sim p_d(\boldsymbol{c})}\left[D(G(\boldsymbol{z},\boldsymbol{c}),\boldsymbol{c})^2\right], \quad (2)\end{aligned}$$

and the objective for generator is defined as

$$\begin{aligned}\min_G \; V_{\mathrm{LSGAN}}(G) &= \\ &\frac{1}{2}E_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z}),\boldsymbol{c}\sim p_d(\boldsymbol{c})}\left[(D(G(\boldsymbol{z},\boldsymbol{c}),\boldsymbol{c})-1)^2\right]. \quad (3)\end{aligned}$$

The LSGAN has been applied to speech enhancement (SEGAN) [12] which generated clean speech signal conditioned on noisy speech signal. An additional L_1 norm loss is used in learning the parameters of the G network of SEGAN. The SEGAN can benefit from adversarial training to get much clean speech waveform. Therefore, the loss for generator of the SEGAN is defined as follows:

$$\begin{aligned}\min_G \; V_{\mathrm{SEGAN}}(G) &= \lambda||G(\boldsymbol{z},\widetilde{\boldsymbol{x}})-\boldsymbol{x}||_1 + \\ &\frac{1}{2}E_{z\sim p_{\boldsymbol{z}}(\boldsymbol{z}),\widetilde{\boldsymbol{x}}\sim p_d(\widetilde{\boldsymbol{x}})}[((D(G(\boldsymbol{z},\widetilde{\boldsymbol{x}}),\widetilde{\boldsymbol{x}})-1)^2], \quad (4)\end{aligned}$$

in which a hyper parameter λ is used to balance the GAN loss and L_1 loss, and $\tilde{\boldsymbol{x}}$ is the input noisy signal.

3. WAVENET ADAPTATION USING GAN

3.1. Parallel WaveNet Vocoder

The original auto-regressive WaveNet is a model that can generate perfect speech waveform. Different from the conventional vocoders, such as STRAIGHT [5] and WORLD [4], the WaveNet vocoder doesn't depend on the source-filter asumption of speech signal. This makes it a perfect vocoder that can avoid the problems of excitation extraction. However, due to its auto-regressive nature, the waveform generation is unbearably slow (100 times slower than real time or more on a Nvidia Tesla P40 GPU). The parallel WaveNet addressed the inference problem by using the inverse auto-regressive flow (IAF) [13], which can perform 30 times faster than real time on a Nvidia Tesla P40 GPU. However its training is very tricky.

IAF is a method that enables the model to convert the input noise signal into speech waveform. Using noise signal as the input enables the model to compute in parallel. This is key to real-time generation. However, IAF is difficult to optimize directly because of the requirement of auto-regressive ly computed log-likelihood loss. [6] proposed a probability density distillation method to distill the student WaveNet efficiently from auto-regressive WaveNet with mixture of logistic (MoL) output distribution [14]. Therefore, the student WaveNet can generate audios whose fidelity can be very close to that of the auto-regressive WaveNet. On the other hand, the training of the model starts from training a time consuming teacher auto-regressive WaveNet.

3.2. WaveNet Adaptation

Speaker adaptation is a commonly adopted method for fast building of acoustic models for speech synthesis and speech recognition, especially for cases where training data are limited. Speaker adaptation can also be applied directly to the parallel WaveNet model, which means that speaker adaptation need to be applied to both teacher and student models. This makes the adaptation training of a new speaker slow and tedious.

Therefore in this paper, we proposed to apply the generative adversarial training to the adaptation of parallel WaveNet. As showed in Fig. 2, we adapt parallel WaveNet to a new speaker directly based on the adversarial training method by

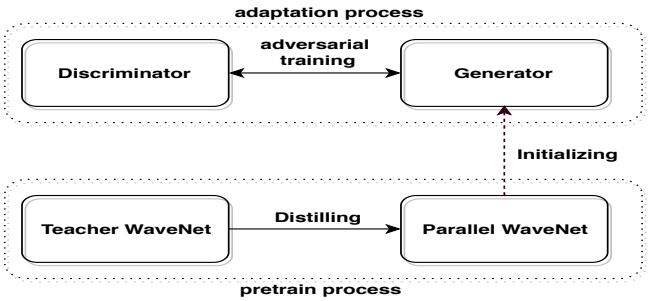


Fig. 2. The architecture of speaker adaptation of parallel WaveNet using GAN.

replacing the distilling teacher model with a discriminator from a GAN.

Specifically for the adaptation GAN (AGAN), a parallel WaveNet of one speaker was pre-trained in advance. We used the same model architecture as reported in [6]. Different from [6], the mean square loss of log-scale STFT-magnitude (log-mag loss) [15] were adopted instead of the power-loss.

At the adaptation phase, the pre-trained model was used to initialize the parameters of the generator of the GAN. We used the LSGAN to stablize the training of adaptation. Similar to SEGAN, a secondary components was added to the loss of generator, in order to better capture the detailed information in frequency domain of the new speaker. As using a single adversarial loss is difficult achieve waveform with high fidelity, which is similar to parallel WaveNet with single Kullback-Leibler (KL) loss [6]. We therefore use an additional log-mag loss which is proven to be effective in capturing spectral details. The log-mag loss is defined as:

$$L_{\log-\mathrm{mag}}(\boldsymbol{x}, \boldsymbol{x}') = || \log(|\mathrm{STFT}(\boldsymbol{x})| + \epsilon) - \log(|\mathrm{STFT}(\boldsymbol{x}')| + \epsilon)||_1 \quad (5)$$

where L_1 norm is used and ϵ is a very small number to ensure the positive of spectral magnitude.

We utilized a discriminator network with a similar architecture with a non-autoregressive WaveNet, which is a non-causal dilated convolution network [16], to identify the generated (fake) waveform against the recored (real) waveform. Similar to conditional GAN, the mel-domain spectrograms were also used as conditional input as well as the generator. We adopted a discriminator network with 10 dilated convolution layers without sacrificing discrimination performance at sample scale.

In detail, for each sentence, we sampled the waveform $\boldsymbol{x}'$ from the output distribution of the generator network. $\boldsymbol{x}'$ was then fed to the discriminator network to evaluate the D loss against samples from real data distribution. The loss of the generator is defined as

$$\min_G V_{\mathrm{AGAN}}(G) = L_{\log-\mathrm{mag}}(G(\boldsymbol{z}, \boldsymbol{c}), \boldsymbol{x}) + \frac{\lambda}{2} E_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z}), \boldsymbol{c}\sim p_d(\boldsymbol{c})}[((D(G(\boldsymbol{z}, \boldsymbol{c}), \boldsymbol{c}) - 1)^2]. \quad (6)$$

4. EXPERIMENTS

4.1. Data Set

We used an open source LJSpeech dataset [17] to evaluate the performance of the proposed speaker adaptation GAN. The initial parallel WaveNet model was pre-trained on our internal speech dataset, which contains 12 hours of mandarin records by a female speaker. The audio for pre-train is re-sampled at 24 kHz, while 22.05 kHz sampling rate of LJSpeech was reserved in the adaptation. The LJSpeech contains 13,100 short audio clips of public domain English speech data from a speaker. The lengths of audio clips range from 1 to 10 seconds and the total length is approximately 24 hours. We randomly selected 2000 audio clips, which is about 3 hours in total, as our training data. Another 30 sentences were used as our test set for subjective evaluation.

4.2. Model setup

Following the configuration of acoustic analysis in Tacotron2 [18], we extracted mel spectrograms as the local acoustic condition for neural vocoders with a frame shift of 256 points and frame length of 2048 points. The initial parallel WaveNet was trained with 1500k steps with a teacher MoL WaveNet trained on the same dataset. At adaptation time, we adopted the Adam optimizer [19] for the AGAN. The noam scheme [20] for learning rate was used with 4k warm-up step. The AGAN model was trained with a batch size of 4 clips and max sample length 24000. For comparison, another parallel WaveNet was adapted by distilling using data of the target speaker.

The discriminator of AGAN was trained by a random initialization. The architecture of discriminator was a non-causal WaveNet with 10 dilated convolution layers, which used filter size 3 and max dilation rate 10. We added a Leaky ReLU [21] activation function with $\alpha = 0.2$ after each layer of convolution except the last output layer. The discriminator also used mel spectrograms as local condition. We used 4-layer de-convolution network to up-sample frame scale mel spectrograms to sample scale. During training, the model converged at 150k steps. The learning rate of the generator and discriminator were set to 0.005 and 0.001 respectively. In the first 50k steps, we froze the parameters of the generator in order to better learn a discriminator. After 50k steps, the generator and discriminator were adversarial trained simultaneously. We found that the value of adversarial loss can, to some extend, reflect the fidelity of generating waveform. We achieved a relatively good result by setting λ to 1.5. Another model with $\lambda = 0.05$ was used for comparison.

Method	Subjective 5-scale MOS
Parallel WaveNet (baseline)	4.53 ± 0.17
AGAN (λ=0.05)	4.50 ± 0.20
AGAN (λ=1.50)	$\mathbf{4.58 \pm 0.16}$
Ground-truth	4.63 ± 0.14

Table 1. Mean Opinion Score(MOS) with 95% confidence intervals for different adaptation method.

4.3. Experimental results

We adopted the commonly used Mean Opinion Score (MOS) to subjectively evaluate our proposed speaker adaptation framework based on GAN. In order to ensure that the results are convincing enough, 30 sentences, which are not included in training set, were randomly selected for fair comparison. Three models were compared, including a parallel WaveNet adaptively trained with the conventional distillation framework, two AGAN models with $\lambda = 0.05$ and $\lambda = 1.5$ respectively. The ground-truth recordings were used in comparison. 63 professional English listeners participated in the listening test. Since it is a neural vocoder, we focused on the fidelity (quality) of speech samples in our experiment.

The results of the subjective MOS evaluation were presented in Table 1. As we can see, our best model (AGAN with $\lambda = 1.5$) performed better than conventional adaptation approach (baseline). Although the absolute improvement, which is 0.05, is not huge enough, the proposed method reduced the gap between baseline model and ground truth by 50%. The proposed model has achieved speech generation with human level high fidelity [1].

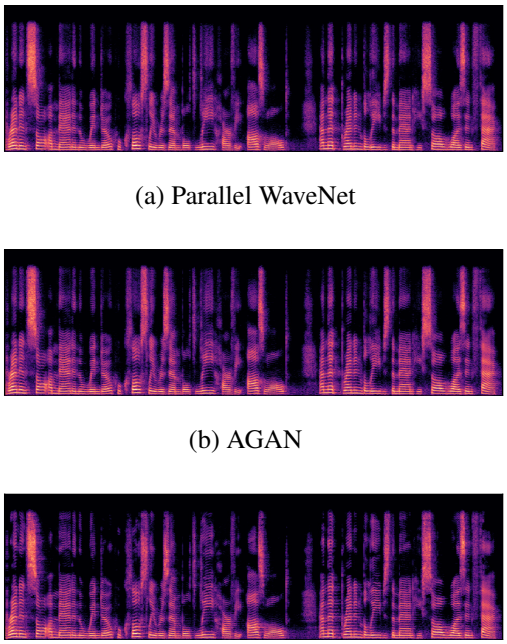
We also investigated the importance of adversarial loss in AGAN by setting different values to λ. It is obviously observed from the results in Table 1 that the performance of AGAN significantly degraded by decreasing weight of adversarial loss, even worse than the baseline model. Anyway, more experiments are still needed to further investigate the relation between speech quality and adversarial loss.

On the other hand, we conducted a case study on the spectrograms of the models. As shown in Figure 3, we found that the proposed model can better capture the detailed spectral information of the target speaker, especially in high frequency domain. We can found obvious harmonic structures in spectrogrames of samples of AGAN and ground truth, while they are missing in that of baseline.

4.4. Adaptation cost of training

The time cost of adaptation training is vital for deployment in practical applications. In order to investigate the adaptation efficient of the proposed model, we evaluate the training time of adaptation methods on an Nvidia Tesla P40 GPU. It took about 36 hours to perform a adaptation training for the baseline parallel WaveNet, which contains adaptation training of both teacher and student WaveNet model. On the other hand, the adaptation training of AGAN on the same data set took about 12 hours. This is not only because of the efficiency of AGAN training, but also since its independence on a teacher model. Although a discriminator is required in AGAN, its training is quite fast and stable.

[1]Audio samples can be found at `https://agan-demo.github.io/`.



(a) Parallel WaveNet

(b) AGAN

(c) Ground-Truth

Fig. 3. STFT spectrograms of samples from parallel WaveNet, **AGAN (ours)** and Ground-Truth recording.

5. CONCLUSIONS

In this work, we proposed a speaker adaptation framework for parallel WaveNet vocoder based on GAN (AGAN). Comparing with conventional retrain based model adaptation, AGAN can effectively perform adaptation on a relatively small amount of a new speaker data and achieve the ability of generating speech with higher perceptual quality. Our experiments indicated that the proposed method can further reduce the gap between speech samples from recording and proposed model. In addition, as a future work, it is straightforward that the proposed method can also be applied to optimize an IAF directly based parallel WaveNet model from scratch without the requirement of auto-regressive teacher model.

6. REFERENCES

[1] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al., "Tacotron: A fully end-to-end text-to-speech synthesis model," *arXiv preprint*, 2017.

[2] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[3] Heiga Zen and Haşim Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4470–4474.

[4] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[5] Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain De Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds1," *Speech communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[6] Aaron van den Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George van den Driessche, Edward Lockhart, Luis C Cobo, Florian Stimberg, et al., "Parallel wavenet: Fast high-fidelity speech synthesis," *arXiv preprint arXiv:1711.10433*, 2017.

[7] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[9] Wei Ping, Kainan Peng, and Jitong Chen, "Clarinet: Parallel wave generation in end-to-end text-to-speech," *arXiv preprint arXiv:1807.07281*, 2018.

[10] Augustus Odena, Christopher Olah, and Jonathon Shlens, "Conditional image synthesis with auxiliary classifier gans," *arXiv preprint arXiv:1610.09585*, 2016.

[11] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley, "Least squares generative adversarial networks," in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2813–2821.

[12] Santiago Pascual, Antonio Bonafonte, and Joan Serra, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[13] Diederik P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling, "Improved variational inference with inverse autoregressive flow," in *Advances in Neural Information Processing Systems*, 2016, pp. 4743–4751.

[14] Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P Kingma, "Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications," *arXiv preprint arXiv:1701.05517*, 2017.

[15] Sercan O Arik, Heewoo Jun, and Gregory Diamos, "Fast spectrogram inversion using multi-head convolutional neural networks," *arXiv preprint arXiv:1808.06719*, 2018.

[16] Dario Rethage, Jordi Pons, and Xavier Serra, "A wavenet for speech denoising," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5069–5073.

[17] Keith Ito, "The lj speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[18] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al., "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[19] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[20] Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan N Gomez, Stephan Gouws, Llion Jones, Lukasz Kaiser, Nal Kalchbrenner, Niki Parmar, et al., "Tensor2tensor for neural machine translation," *arXiv preprint arXiv:1803.07416*, 2018.

[21] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. icml*, 2013, vol. 30, p. 3.