EXPLORING SPEECH ENHANCEMENT WITH GENERATIVE ADVERSARIAL NETWORKS FOR ROBUST SPEECH RECOGNITION

Chris Donahue[*]

UC San Diego Department of Music

cdonahue@ucsd.edu

Bo Li, Rohit Prabhavalkar

Google

{boboli,prabhavalkar}@google.com

ABSTRACT

We investigate the effectiveness of generative adversarial networks (GANs) for speech enhancement, in the context of improving noise robustness of automatic speech recognition (ASR) systems. Prior work [1] demonstrates that GANs can effectively suppress additive noise in raw waveform speech signals, improving perceptual quality metrics; however this technique was not justified in the context of ASR. In this work, we conduct a detailed study to measure the effectiveness of GANs in enhancing speech contaminated by both additive and reverberant noise. Motivated by recent advances in image processing [2], we propose operating GANs on log-Mel filterbank spectra instead of waveforms, which requires less computation and is more robust to reverberant noise. While GAN enhancement improves the performance of a clean-trained ASR system on noisy speech, it falls short of the performance achieved by conventional multi-style training (MTR). By appending the GAN-enhanced features to the noisy inputs and retraining, we achieve a 7% WER improvement relative to the MTR system.

Index Terms— Speech enhancement, automatic speech recognition, generative adversarial networks, deep learning

1. INTRODUCTION

Speech enhancement techniques aim to improve the quality of speech by reducing noise. They are crucial components, either explicitly [3] or implicitly [4, 5], in ASR systems for noise robustness. Even with state-of-the-art deep learning-based ASR models, noise reduction techniques can still be beneficial [6]. Besides the conventional enhancement techniques [3], deep neural networks have been widely adopted to either directly reconstruct clean speech [7, 8] or estimate masks [9, 10, 11] from the noisy signals. Different types of networks have also been investigated in the literature for enhancement, such as denoising autoencoders [12], convolution networks [13] and recurrent networks [14].

In their limited history, GANs [15] have attracted attention for their ability to synthesize convincing images when trained on corpora of natural images. Refinements to network architecture have improved the fidelity of the synthetic images [16]. Isola et al. [2] demonstrate the effectiveness of GANs for image "translation" tasks, mapping images in one domain to related images in another. In spite of the success of GANs for image synthesis, exploration on audio has been limited. Pascual et al. [1] demonstrate promising performance of GANs for speech enhancement in the presence of additive noise, posing enhancement as a translation task from noisy signals to clean ones. Their method, speech enhancement GAN (SEGAN), yields improvements to perceptual speech quality metrics over the noisy data and traditional enhancement baselines. Their investigation seeks to improve speech quality for telephony rather than ASR.

In this work, we study the benefit of GAN-based speech enhancement for ASR. In order to limit the confounding factors in our study, we use an existing ASR model trained on clean speech data to measure the effectiveness of GAN-based enhancement. To gauge performance under more-realistic ASR conditions, we consider reverberation in addition to additive noise. We first train a SEGAN model to map simulated noisy speech to the original clean speech in the time domain. Then, we measure the performance of the ASR model on noisy speech before and after enhancement by SEGAN. Our experiment indicates that SEGAN does not improve ASR performance under these noise conditions.

To address this, we refine the SEGAN method to operate on a time-frequency representation, specifically, log-Mel filterbank spectra. With this *spectral feature mapping* (SFM) approach, we can pass the output of our enhancement model directly to the ASR model (Figure 1). While deep learning has previously been applied to SFM for ASR [17, 18, 19], our work is the first to use GANs for this task. Michelsanti et al. [20] employ GANs for SFM, but target speaker verification rather than ASR. Our frequency-domain approach improves ASR performance dramatically, though performance is comparable to the same enhancement model trained with an L1 reconstruction loss. Anecdotally speaking, the GAN-enhanced spectra appear more realistic than the L1-enhanced spectra when visualized (Figure 3), suggesting that ASR models may not benefit from the fine-grained details that GAN enhancement produces.

[*]Work performed as an intern at Google.

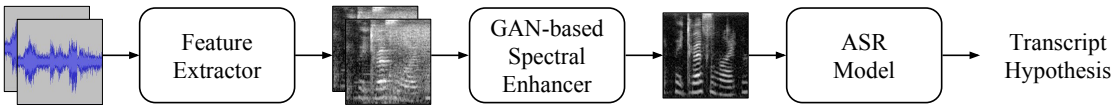


Fig. 1: System overview.

State-of-the-art ASR systems use MTR [21] to achieve robustness to noise at inference time. While this strategy is known to be effective, a resultant model may still benefit from enhancement as a preprocessing stage. To measure this effect, we also use an existing ASR model trained with MTR and compare its performance on noisy speech with and without enhancement. We find that GAN-based enhancement degrades performance of this model, even with retraining. However, retraining the MTR model with both noisy and enhanced features in its input representation improves performance.

2. GENERATIVE ADVERSARIAL NETWORKS

Generative adversarial networks (GANs) are unsupervised generative models that learn to produce realistic samples of a given dataset from low-dimensional, random latent vectors [15]. GANs consist of two models (usually neural networks), a *generator* and a *discriminator*. The generator G maps latent vectors drawn from some known prior p_z to samples: $G : \boldsymbol{z} \mapsto \hat{\boldsymbol{y}}$, where $\boldsymbol{z} \sim p_z$. The discriminator D is tasked with determining if a given sample is real ($\boldsymbol{y} \sim p_{data}$, a sample from the real dataset) or fake ($G(\boldsymbol{z}) \sim p_G$, where p_G is the implicit distribution of the generator when $\boldsymbol{z} \sim p_z$). The two models are pitted against each other in an adversarial framework.

Real-world datasets often contain additional information associated with each example, e.g. the type of object depicted in an image. Conditional GANs (cGANs) [22] use this information $\boldsymbol{x}$ by providing it as input to the generator, typically in a one-hot encoding: $G : \{\boldsymbol{x}, \boldsymbol{z}\} \mapsto \hat{\boldsymbol{y}}$. After training, we can sample from the generator's implicit posterior $p_G(\hat{\boldsymbol{y}} \mid \boldsymbol{x})$ by fixing $\boldsymbol{x}$ and sampling $\boldsymbol{z} \sim p_z$. To accomplish this, G is trained to minimize the following objective, while D is trained to maximize it:

$$\mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}\sim p_{data}}[\log D(\boldsymbol{x}, \boldsymbol{y})] + \mathbb{E}_{\boldsymbol{x}\sim p_{data},\boldsymbol{z}\sim p_z}[\log(1 - D(\boldsymbol{x}, G(\boldsymbol{x}, \boldsymbol{z})))]. \quad (1)$$

Recently, researchers have used full-resolution images as conditioning information. Isola et al. [2] propose a cGAN approach to address image-to-image "translation" tasks, where appropriate datasets consist of matched pairs of images $(\boldsymbol{x}, \boldsymbol{y})$ in two different domains. Their approach, dubbed *pix2pix*, uses a convolutional generator that receives as input an image $\boldsymbol{x}$, a latent vector $\boldsymbol{z}$ and produces $G(\boldsymbol{x}, \boldsymbol{z})$: an image of identical resolution to $\boldsymbol{x}$. A convolutional discriminator is shown pairs of images stacked along the channel axis and is trained to determine if the pair is real $(\boldsymbol{x}, \boldsymbol{y})$ or fake $(\boldsymbol{x}, G(\boldsymbol{x}, \boldsymbol{z}))$.

For conditional image synthesis, prior work [23] demonstrates the effectiveness of combining the GAN objective with an unstructured loss. Noting this, Isola et al. [2] use a hybrid objective to optimize their generator, penalizing it for L1 reconstruction error in addition to the adversarial objective:

$$\min_G \max_D V(G, D) = \mathcal{L}_{cGAN}(G, D) + 100 \cdot \mathcal{L}_{L1}(G),$$
$$\text{where } \mathcal{L}_{L1}(G) = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}\sim p_{data},\boldsymbol{z}\sim p_z}[\|\boldsymbol{y} - G(\boldsymbol{x}, \boldsymbol{z})\|_1]. \quad (2)$$

3. METHOD

We describe our approach to spectral feature mapping using GANs, beginning by outlining the related time-domain SEGAN approach.

3.1. SEGAN

Pascual et al. [1] propose SEGAN, a technique for enhancing speech in the time domain. The SEGAN method is a 1D adaptation of the 2D pix2pix [2] approach. The fully-convolutional generator receives second-long (16384 samples at 16 kHz) windows of noisy speech as input and is trained to output clean speech. During inference, the generator is used to enhance longer segments of speech by repeated application on one-second windows without overlap.

The generator's encoder consists of 11 layers of stride-2 convolution with increasing depth, resulting in a feature map at the bottle-neck of 8 timesteps with depth 1024. Here, the authors append a latent noise vector $\boldsymbol{z}$ of the same dimensionality along the channel axis. The resultant 8x2048 matrix is input to an 11-layer upsampling decoder, with skip connections from corresponding input feature maps. As a departure from pix2pix, the authors remove batch normalization from the generator. Furthermore, they use 1D filters of width 31 instead of 2D filters of size 4x4. They also substitute the traditional GAN loss function with the least squares GAN objective [24].

In agreement with observations from [25], we found that the SEGAN generator learned to ignore $\boldsymbol{z}$. We hypothesize that latent vectors may be unnecessary given the presence of noise in the input. We removed the latent vector from the generator altogether; the resultant deterministic model demonstrated improved performance in our experiments.

3.2. FSEGAN

It is common practice in ASR to preprocess time-domain speech data into time-frequency *spectral* representations.

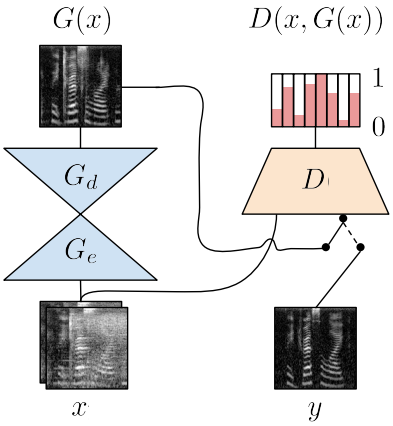


Fig. 2: Time-frequency FSEGAN enhancement strategy. G (composition of encoder G_e and decoder G_d) maps stereo, noisy spectra $\boldsymbol{x}$ to enhanced $G(\boldsymbol{x})$. D receives as input either $(\boldsymbol{x}, \boldsymbol{y})$ or $(\boldsymbol{x}, G(\boldsymbol{x}))$ and decides if the pair is real or enhanced.

Phase information of the signal is typically discarded; hence, an enhancement model only needs to reconstruct the magnitude information. Because frequency-domain representations are used in the discriminative setting, an enhancer targeting improved ASR may benefit from operating in this domain. With this motivation, we propose a frequency-domain SEGAN (FSEGAN), which performs spectral feature mapping using an approach similar to pix2pix [2]. FSEGAN ingests time-windowed spectra of noisy speech and is trained to output clean speech spectra.

The fully-convolutional FSEGAN generator contains 7 encoder and 7 decoder layers (4x4 filters with stride 2 and increasing depth), and features skip connections across the bottleneck between corresponding layers. The final decoder layer has linear activation and outputs a single channel. As with SEGAN, we exclude both batch normalization and latent codes $\boldsymbol{z}$ from our generator, resulting in a deterministic model. The discriminator contains 4 convolutional layers with 4x4 filters and a stride of 2. A final 1x8 layer (stride 1 with *sigmoid* nonlinearity) aggregates the activations from 8 frequency bands into a single decision for each of 8 timesteps. We train FSEGAN with the objective in Equation 2. Other architectural details are identical to pix2pix [2].[1] The FSEGAN approach is depicted in Figure 2.

4. EXPERIMENTS

4.1. Dataset

We use the Wall Street Journal (WSJ) corpus [26] as our source of clean speech data. Specifically, we train on the 16 kHz, speaker-independent SI-284 set (81 hours, 284 speakers, 37k utterances). We perform validation on the dev93 set and evaluate on the eval92 set.

[1]We modify the following open-source implementation: `https://github.com/affinelayer/pix2pix-tensorflow`

We use large, stereo datasets of musical and ambient signals as our additive noise sources for MTR. This data is collected from YouTube and recordings of daily life environments. During training, we use discrete mixtures ranging from 0 dB to 30 dB SNR, averaging 11 dB. At test time, the SNRs are slightly offset, ranging from 0.2 dB to 30.2 dB.

As a source of reverberation for MTR, we use a room simulator as described in [27]. The simulator randomizes the positions of the speech and noise sources, the position of a virtual stereo microphone, the T60 of the reverberation, and the room geometry. Through this process, our monaural speech data becomes stereo. Room configurations for training and testing are drawn from distinct sets; they are randomized during training and fixed during testing.

4.2. ASR Model

We train a monaural listen, attend and spell (LAS) model [28] on the clean WSJ training data as described in Section 4.1, performing early stopping by the WER of the model on the validation set. To compare the effectiveness of GAN-based enhancement to MTR, we also train the same model using MTR as described in Section 4.1, using only one channel of the noisy speech. We refer to the clean-trained model as ASR-Clean and the MTR-trained model as ASR-MTR.

To preprocess the time-domain data, we first apply the short-time Fourier transform with a window size of 32 ms and a hop size of 10 ms. We retain the magnitude spectrum of the output and discard the phase. Then, we calculate triangular windows for a bank of 128 filters, where filter center frequencies are equally spaced on the Mel scale between 125 Hz and 7500 Hz. After applying this transform to the magnitude spectrum, we take the logarithm of the output and normalize each frequency bin to have zero mean and unit variance.

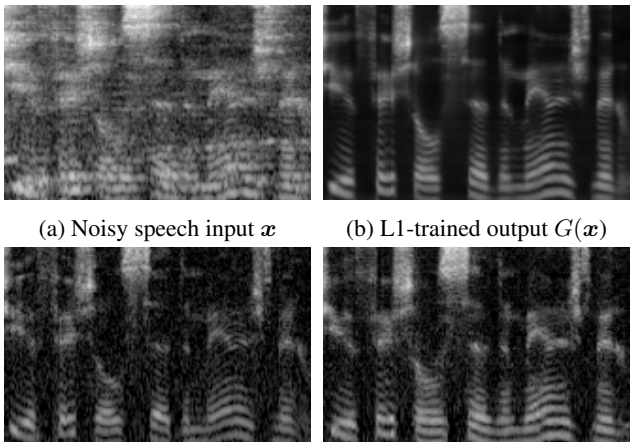
To process these features, our LAS encoder contains two convolutional layers with filter sizes: 1) 3x5x1x32, and 2) 3x3x32x32. The activations of the second layer are passed to a bidirectional, convolutional LSTM layer [29, 30], followed by three bidirectional LSTM layers. The decoder contains a unidirectional LSTM with additive attention [31] whose outputs are fed to a *softmax* over characters.

4.3. GAN

For our GAN experiments, we generate multi-style, matched pairs of noisy and clean speech in the manner described in Section 4.1. For our FSEGAN experiments, we transform these pairs into time-frequency spectra in a manner identical to that of the ASR model described in Section 4.2. We frame the pairs into 1.28 s windows with 50% overlap and train with random minibatches of size 100. The resultant SEGAN inputs are 20480 samples long and the FSEGAN inputs are 128x128. In alignment with [1], we use no overlap during evaluation. We perform early stopping based on the WER of ASR-Clean on the enhanced validation set.

Test Set	Enhancer	ASR-Clean WER	ASR-MTR WER
Clean	None	11.9	14.3
MTR	None	72.2	20.3
	SEGAN	80.7	52.8
	FSEGAN	33.3	25.4

Table 1: Results of GAN enhancement experiments.



(a) Noisy speech input $\boldsymbol{x}$ (b) L1-trained output $G(\boldsymbol{x})$

(c) Clean speech target $\boldsymbol{y}$ (d) FSEGAN output $G(\boldsymbol{x})$

Fig. 3: Noisy utterance enhanced by FSEGAN.

5. RESULTS

We compute the WER of ASR-Clean and ASR-MTR on both the clean and MTR test sets. We also compute the WER of both models on the MTR test set enhanced by SEGAN and FSEGAN. Results are shown in Table 1. While the WER of ASR-Clean (11.9%) is not state-of-the-art, we focus more on relative performance with enhancement. Our previous work [32] has shown that LAS can approach state-of-the-art performance when trained on larger amounts of data.

The SEGAN method degrades performance of ASR-Clean on the MTR test set by 12% relative. To verify the accuracy of this result, we also ran an experiment to remove only additive noise with SEGAN: the conditions in the original paper. Under that condition, we found that SEGAN improved performance of ASR-Clean by 21% relative, indicating that SEGAN struggles to suppress reverberation.

In contrast, our FSEGAN method improves the performance of ASR-Clean by 54% relative. While this is a dramatic improvement, it does not exceed the performance achieved with MTR training (33% vs. 20% WER). Furthermore, FSEGAN degraded performance for ASR-MTR, consistent with observations in [33].

We show a visualization of FSEGAN enhancement in Figure 3. The procedure appears to reduce both the presence of additive noise and reverberant smearing. Despite this, the procedure degrades performance of ASR-MTR. We hypothesize that the enhancement process may be introducing hitherto-unseen distortions that compromise performance.

Model	WER (%)
MTR Baseline *	20.3
+ Stereo	19.0
MTR + FSEGAN Enhancer *	25.4
+ Retraining	21.0
+ Hybrid Retraining	17.6
MTR + L1-trained Enhancer *	21.4
+ Retraining	18.0
+ Hybrid Retraining	17.1

Table 2: Results of ASR-MTR retraining. Rows marked with * are the same model under different enhancement conditions.

5.1. Retraining Experiments

Hoping to improve performance beyond that of MTR training alone, we retrain ASR-MTR using FSEGAN-enhanced features. To examine the effectiveness of the adversarial component of FSEGAN, we also experiment with training the same enhancement model using only the L1 portion of the hybrid loss function ($\mathcal{L}_{L1}(G)$ from Equation 2).

Considering that the model may benefit from knowledge of both the enhanced *and* noisy features, we also train a model to ingest these two representations stacked along the channel axis. We initialize this new hybrid model from the existing ASR-MTR checkpoint, setting the additional parameters to zero to ensure identical performance at the start of training. To ensure that the hybrid model is not *strictly* benefiting from increased parametrization, we train an LAS model from scratch with stereo MTR input. Results for these experiments appear in Table 2.

Retraining ASR-MTR with FSEGAN-enhanced features improves performance by 17% relative to naively feeding them, but still falls short of MTR training. Hybrid retraining with both the original noisy and enhanced features improves performance further, exceeding the performance of stereo MTR training alone by 7% relative. Our results indicate that training the same enhancer with the L1 objective achieves better ASR performance than an adversarial approach, suggesting limited usefulness of GANs in this context.

6. CONCLUSIONS

We have introduced FSEGAN, a GAN-based method for performing speech enhancement in the frequency domain, and demonstrated improvements in ASR performance over a prior time-domain approach. We provide evidence that, with retraining, FSEGAN can improve the performance of existing MTR-trained ASR systems. Our experiments indicate that, for ASR, simpler regression approaches may be preferable to GAN-based enhancement. FSEGAN appears to produce plausible spectra and may be more useful for telephonic applications if paired with an invertible feature representation.

7. REFERENCES

[1] Santiago Pascual, Antonio Bonafonte, and Joan Serrà, "SEGAN: Speech Enhancement Generative Adversarial Network," in *Proc. INTERSPEECH*, 2017.

[2] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. CVPR*, 2017.

[3] Jingdong Chen, Jacob Benesty, Yiteng Arden Huang, and Eric J Diethorn, "Fundamentals of noise reduction," in *Springer Handbook of Speech Processing*, pp. 843–872. Springer, 2008.

[4] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal processing letters*, vol. 21, no. 1, pp. 65–68, 2014.

[5] Bo Li, Tara N Sainath, Ron J Weiss, Kevin W Wilson, and Michiel Bacchiani, "Neural Network Adaptive Beamforming for Robust Multichannel Speech Recognition," in *Proc. INTERSPEECH*, 2016.

[6] Jinyu Li, Yan Huang, and Yifan Gong, "Improved cepstra minimum-mean-square-error noise reduction algorithm for robust speech recognition," in *Proc. ICASSP*. IEEE, 2017.

[7] Xue Feng, Yaodong Zhang, and James Glass, "Speech feature denoising and dereverberation via deep autoencoders for noisy reverberant speech recognition," in *Proc. ICASSP*. IEEE, 2014.

[8] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on ASLP*, vol. 23, no. 1, pp. 7–19, 2015.

[9] DeLiang Wang and Jitong Chen, "Supervised speech separation based on deep learning: an overview," *arXiv:1708.07524*, 2017.

[10] Bo Li and Khe Chai Sim, "A spectral masking approach to noise-robust speech recognition using deep neural networks," *IEEE/ACM Transactions on ASLP*, vol. 22, no. 8, pp. 1296–1305, 2014.

[11] Arun Narayanan and DeLiang Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. ICASSP*. IEEE, 2013.

[12] Xugang Lu, Yu Tsao, Shigeki Matsuda, and Chiori Hori, "Speech enhancement based on deep denoising autoencoder.," in *Proc. INTERSPEECH*, 2013.

[13] Szu-Wei Fu, Yu Tsao, and Xugang Lu, "Snr-aware convolutional neural network modeling for speech enhancement.," in *Proc. INTERSPEECH*, 2016.

[14] Felix Weninger, Hakan Erdogan, and *et al.*, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *Proc. International Conference on Latent Variable Analysis and Signal Separation*, 2015.

[15] Ian Goodfellow, Jean Pouget-Abadie, and *et al.*, "Generative adversarial nets," in *Proc. NIPS*, 2014.

[16] Alec Radford, Luke Metz, and Soumith Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. ICLR*, 2016.

[17] Andrew L Maas, Quoc V Le, and *et al.*, "Recurrent neural networks for noise reduction in robust ASR," in *Proc. INTERSPEECH*, 2012.

[18] Kun Han, Yanzhang He, Deblin Bagchi, Eric Fosler-Lussier, and DeLiang Wang, "Deep neural network based spectral feature mapping for robust speech recognition," in *Proc. INTERSPEECH*, 2015.

[19] Arun Narayanan and DeLiang Wang, "Improving robustness of deep neural network acoustic models via speech separation and joint adaptive training," *IEEE/ACM transactions on ASLP*, vol. 23, no. 1, pp. 92–101, 2015.

[20] Daniel Michelsanti and Zheng-Hua Tan, "Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification," in *Proc. INTERSPEECH*, 2017.

[21] R Lippmann, Edward Martin, and D Paul, "Multi-style training for robust isolated-word speech recognition," in *Proc. ICASSP*. IEEE, 1987.

[22] Mehdi Mirza and Simon Osindero, "Conditional generative adversarial nets," *arXiv:1411.1784*, 2014.

[23] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros, "Context encoders: Feature learning by inpainting," in *Proc. CVPR*, 2016.

[24] Xudong Mao, Qing Li, and *et al.*, "Least squares generative adversarial networks," *arXiv:1611.04076*, 2016.

[25] Michael Mathieu, Camille Couprie, and Yann LeCun, "Deep multi-scale video prediction beyond mean square error," in *Proc. ICLR*, 2016.

[26] Douglas B Paul and Janet M Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.

[27] Chanwoo Kim, Ananya Misra, Kean Chin, Thad Hughes, Arun Narayanan, Tara Sainath, and Michiel Bacchiani, "Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in Google Home," in *Proc. INTERSPEECH*, 2017.

[28] William Chan, Navdeep Jaitly, Quoc V Le, and Oriol Vinyals, "Listen, attend and spell," *arXiv:1508.01211*, 2015.

[29] Yu Zhang, William Chan, and Navdeep Jaitly, "Very deep convolutional networks for end-to-end speech recognition," in *Proc. ICASSP*. IEEE, 2017.

[30] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. NIPS*, 2015.

[31] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. ICLR*, 2015.

[32] Chung-Cheng Chiu, Tara N Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J Weiss, Kanishka Rao, Katya Gonina, et al., "State-of-the-art speech recognition with sequence-to-sequence models," in *Proc. ICASSP (accepted)*. IEEE, 2018.

[33] Arun Narayanan and DeLiang Wang, "Joint noise adaptive training for robust automatic speech recognition," in *Proc. ICASSP*. IEEE, 2014.