ROBUST SPEECH RECOGNITION USING GENERATIVE ADVERSARIAL NETWORKS

Anuroop Sriram, Heewoo Jun*, Yashesh Gaur, Sanjeev Satheesh*

Baidu Research, Sunnyvale, CA, USA

ABSTRACT

This paper describes a general, scalable, end-to-end framework that uses the generative adversarial network (GAN) objective to enable robust speech recognition. Encoders trained with the proposed approach enjoy improved invariance by learning to map noisy audio to the same embedding space as that of clean audio. Unlike previous methods, the new framework does not rely on domain expertise or simplifying assumptions as are often needed in signal processing, and directly encourages robustness in a data-driven way. We show the new approach improves simulated far-field speech recognition of vanilla sequence-to-sequence models without specialized front-ends or preprocessing.

Index Terms— automatic speech recognition, speech enhancement, generative adversarial networks

1. INTRODUCTION

Automatic speech recognition (ASR) is becoming increasingly more integral in our day-to-day lives enabling virtual assistants and smart speakers like Siri, Google Now, Cortana, Amazon Echo, Google Home, Apple HomePod, Microsoft Invoke, Baidu Duer and many more. While recent breakthroughs have tremendously improved ASR performance [1, 2] these models still suffer considerable degradation from reasonable variations in reverberations, ambient noise, accents and Lombard reflexes that humans have little or no issue recognizing.

Most of these problems can be mitigated by training the models on a large volume of data that exemplify these effects. However, in the case of non-stationary processes, such as accents, accurate data augmentation is most likely infeasible, and in general, collecting high quality datasets can be expensive and time-consuming. Past robust ASR literature has considered hand-engineered front-ends and data-driven approaches in an attempt to increase the value of relatively parsimonious data with desired effects [3, 4]. While these techniques are quite effective in their respective operating regimes, they do not generalize well to other modalities in practice due to the aforementioned reasons. Namely, it is difficult to model anything beyond reverberation and background noise from the first principles. Existing techniques do not directly induce invariance for ASR or are not scalable. And, due to the sequential nature of speech, alignments are needed to compare two different utterances of the same text.

In this work, we employ the generative adversarial network (GAN) framework [5] to increase the robustness of seq-to-seq models [6] in a scalable, end-to-end fashion. The encoder component is treated as the generator of GAN and is trained to produce indistinguishable embeddings between noisy and clean audio samples. Because no restricting assumptions are made, this new robust training approach can in theory learn to induce robustness without alignment or complicated inference pipeline and even where augmentation is not possible. We also experiment with encoder distance objective to explicitly restrict the embedding space and demonstrate that achieving invariance at the hidden representation level is a promising direction for robust ASR.

The rest of the paper is organized as follows. Related work is documented in Section 2. Section 3 defines our notations and details the robust ASR GAN. Section 4 explains the experimental setup. Section 5 shows results on the Wall Street Journal (WSJ) dataset with simulated far-field effects. Finishing thoughts are found in Section 6.

2. RELATED WORK

A vast majority of work in robust ASR deals with reverberations and ambient noise; [3] provides an extensive survey in this effort. One of the most effective approaches in this variability is to devise a strong front-end such as the weighted prediction error (WPE) speech dereverberation [7, 8] and train the resulting neural network with realistic augmented data [9, 10].

A shift from more traditional signal processing techniques to more modern, data-driven methods was seen when the denoising autoencoder [11] was employed to induce invariance to reverberations [12]. This is novel in that the autoencoder is explicitly trained to predict the original audio features from a perturbed version convolved with an impulse response. While denoising autoencoder models for enhancing speech have been shown to improve perceptual quality of the produced speech, they have not demonstrated significant improvement for the task of speech recognition. This is because autoencoders are trained to reconstruct all aspects of the original audio, including many features that are not im-

* equal contribution.

portant for speech recognition, such as the voice and accent of the speaker, background noises etc. In fact, ASR systems learn to remove such artifacts of the input audio as they can hinder speech recognition performance. [13] proposed multiple rounds of joint denoising and ASR training for each audio sample, but this approach is not scalable for large datasets.

A similar approach in spirit is to minimize the distance in the embedding space between clean and noisy audio. The intuition here is that the embedding distance is a measure of semantic similarity [14]. However, the perturbed speech may have a different time duration than the reference audio; dynamic time warping [15] can be used to approximate the alignment and compare sequences of varying lengths, but there is an increased computational overhead.

[16] uses the generative adversarial networks (GAN) for domain adaptation to make the simulated images look more realistic to improve the task of robotic hand grasping. GAN [5] is an unsupervised learning framework, where the generator network learns to produce increasingly more realistic data in attempt to fool a competing discriminator. Because equilibrium is reached at a saddle point, it is notoriously hard to train. There have been many improvements to this technique. For example, Wasserstein GAN [17] uses the Earth-Mover distance to mitigate optimization issues. It is also less susceptible to architectural choices.

For speech, [18] proposes a GAN based speech enhancement method called SEGAN but without the end goal of speech recognition. SEGAN operates on raw speech samples and hence it is computationally impractical for large scale experiments.

3. ROBUST ASR

3.1. Encoder distance enhancer

As explained in Section 2, denoising reconstruction and perceptual enhancement do not significantly improve ASR. A better approach would be to reconstruct only those aspects of the audio which are important for predicting the text spoken and ignore everything else. We hypothesize that the encoders of well trained ASR systems would learn to retain only this information from the input audio. Based on this idea, we propose a new sequence-to-sequence architecture for robust speech recognition that tries to match the output of the encoder for clean audio and noisy audio.

The system works as follows: the same encoder, g, is applied to the clean audio x and the corresponding noisy audio $\widetilde{x}$ to produce hidden states $z = g(x)$ and $\widetilde{z} = g(\widetilde{x})$. The decoder, h, models the conditional probability $p(y|x) = p(y|z)$ and is used to predict the output text sequence one character at a time. This architecture is described in Figure 1. The entire system is trained end-to-end using a multi-task objective that tries to minimize the cross-entropy loss of predicting y

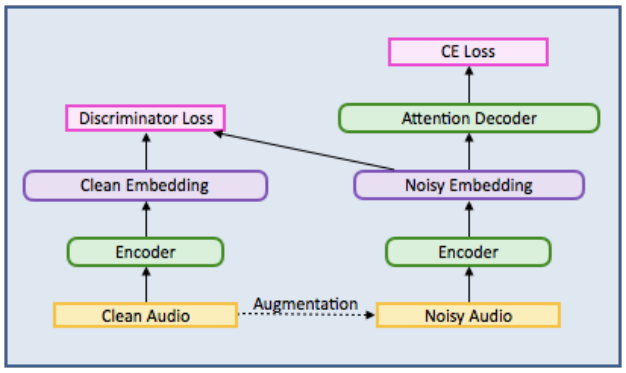

Fig. 1. Architecture of the enhancer models introduced in this paper. The discriminator loss can be L^1-distance or WGAN loss. The entire model is trained end-to-end using both the discriminator loss and the cross-entropy loss. We use RIR convolution to simulate far-field audio. It's also possible to train this model with the same speech recorded in different conditions.

from $\widetilde{x}$ and the normalized L^1-distance between z and $\widetilde{z}$:

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[H(h(\widetilde{z}),y) + \lambda\frac{\|z-\widetilde{z}\|_1}{\|z\|_1 + \|\widetilde{z}\|_1 + \epsilon}\right]. \quad (1)$$

3.2. GAN enhancer

In our experiments, we found the encoder distance penalty to yield excellent results but it has the disadvantage that the encoder content between clean and noisy audio has to match frame for frame. Instead, employing the GAN framework, we can have a discriminator output a scalar likelihood of the entire speech being clean, and train the encoder to generate embeddings that are indistinguishable by the discriminator.

In this paper, Wasserstein GAN (WGAN) [17] is used. Following the notations of WGAN, we parametrize the seq-to-seq and discriminator models with θ and w respectively. The overall architecture depicted in Figure 1 remains the same, but the encoder distance in (1) is now replaced with the dual of Earth-Mover (EM) distance

$$\max_{w\in\mathcal{W}}\left\{\mathbb{E}_x\left[f_w(g_\theta(x))\right] - \mathbb{E}_{\widetilde{x},\varepsilon}\left[f_w(g_\theta(\widetilde{x}+\varepsilon))\right]\right\}. \quad (2)$$

We treat the embedding of the clean input x as real data and the embedding of $\widetilde{x}$, which can either be augmented from x or drawn from a different modality, as being fake. And so, as GAN training progresses, the encoder g_θ should learn to remove extraneous information to ASR to be able to fool the discriminator. In practice, we found that including a random Gaussian noise ε to the input prior of the generator helps improve training. Also, weights in the parameter set $\mathcal{W}$ should be clipped to ensure the duality of (2) holds up to a constant multiple [17]. The adapted WGAN training procedure is detailed in Algorithm 1.

Data: n_{critic}, the number of critic per robust ASR updates. c, the clipping parameter. m, the batch size.
while θ has not converged **do**
  for $t = 1, \ldots, n_{\text{critic}}$ **do**
    Sample $\{(x^{(i)}, y^{(i)}) \sim \mathcal{D}\}_{i=1}^{m}$ a batch of labeled speech data.
    Sample $\{\widetilde{x}^{(i)}\}_{i=1}^{m}$ by augmentation or from a different distribution.
    Sample $\{\varepsilon^{(i)}\}_{i=1}^{m}$ a batch of prior noise.
    $g_\theta \leftarrow \nabla_\theta \left[\frac{1}{m}\sum_{i=1}^{m} H(h_\theta(g_\theta(x^{(i)})), y^{(i)})\right]$
    $\theta \leftarrow \theta - \text{Optimizer}(\theta, g_\theta)$
    $g_w \leftarrow \nabla_w \left[\frac{1}{m}\sum_{i=1}^{m} f_w(g_\theta(x^{(i)})) - \frac{1}{m}\sum_{i=1}^{m} f_w(g_\theta(\widetilde{x}^{(i)} + \varepsilon^{(i)}))\right]$
    $w \leftarrow w + \text{RMSProp}(w, g_w)$
    $w \leftarrow \text{clip}(w, -c, c)$
  end
  Sample $\{(x^{(i)}, y^{(i)}) \sim \mathcal{D}\}_{i=1}^{m}$ a batch of labeled speech data.
  Sample $\{\widetilde{x}^{(i)}\}_{i=1}^{m}$ by augmentation or from a different distribution.
  Sample $\{\varepsilon^{(i)}\}_{i=1}^{m}$ a batch of prior noise.
  $g_\theta \leftarrow \nabla_\theta \left[\frac{1}{m}\sum_{i=1}^{m} H(h_\theta(g_\theta(x^{(i)})), y^{(i)}) - \lambda\frac{1}{m}\sum_{i=1}^{m} f_w(g_\theta(\widetilde{x}^{(i)} + \varepsilon^{(i)}))\right]$
  $\theta \leftarrow \theta - \text{Optimizer}(\theta, g_\theta)$
end

Algorithm 1: WGAN enhancer training. The seq-to-seq model was trained using the Adam optimizer in our experiments. If $\widetilde{x}$ can be generated from x, data augmentation can also be used to update the seq-to-seq model.

4. EXPERIMENTAL SETUP

4.1. Corpora and Tasks

We evaluated the enhancer framework on the Wall Street Journal (WSJ) corpus with simulated far-field effects. The dev93 and eval92 sets were used for hyperparameter selection and evaluation respectively. The reverberant speech is generated with room impulse response (RIR) augmentation as in [19], where each audio is convolved with a randomly chosen RIR signal. The clean and far-field audio durations are kept the same with valid convolution so that the encoder distance enhancer can be applied. We collected 1088 impulse responses, using a linear array of 8 microphones, 120 and 192 of which were held out for development and evaluation. The speaker was placed in a variety of configurations, ranging from 1 to 3 meters distance and 60 to 120 degrees inclination with respect to the array, for 20 different rooms. Mel spectrograms of 20 ms samples with 10 ms stride and 40 bins were used as input features to all of our baseline and enhancer models.

4.2. Network Architecture

For the acoustic model, we used the sequence-to-sequence framework with soft attention based on [6]. The architecture of the encoder is described in Table 1. The decoder consisted of a single 256 dimensional GRU layer with a hybrid attention mechanism similar to the models described in [20].

The discriminator network of the WGAN enhancer is described in Table 2. All convolutional layers use leaky ReLU activation [21] with 0.2 slope for the leak, and batch normalization [22].

Bidirectional GRU (dimension = 256, batch norm)
Pooling (2x1 striding)
Bidirectional GRU (dimension = 256, batch norm)
Pooling (2x1 striding)
Bidirectional GRU (dimension = 256, batch norm)
Pooling (2x1 striding)
Bidirectional GRU (dimension = 256, batch norm)
Bidirectional GRU (dimension = 256, batch norm)
Bidirectional GRU (dimension = 256, batch norm)

Table 1. Architecture of the encoder.

7x2 Convolution, 32 filters, 5x1 striding
3x3 Convolution, 64 filters, 2x1 striding
Bidirectional LSTM (dimension = 32)
3x3 Convolution, 64 filters, 2x1 striding
3x3 Convolution, 96 filters, 1x1 striding
Bidirectional LSTM (dimension = 32)
Linear projection to per-time step scalar
Sigmoid
Mean pool of likelihood scores

Table 2. Architecture of the critic. (feature)×(time).

4.3. Training

To establish a baseline, in the first experiment, we trained a simple attention based seq-to-seq model. All the seq-to-seq networks in our experiments were trained using the Adam optimizer. We evaluate all models on both clean and far-field test sets.

Model	Near-Field		Far-Field	
	CER	WER	CER	WER
seq-to-seq	7.43%	21.18%	23.76%	50.84%
seq-to-seq + far-field Augmentation	7.69%	21.32%	12.47%	30.59%
seq-to-seq + L^1-Distance Penalty	7.54%	20.45%	12.00%	29.19%
seq-to-seq + GAN Enhancer	7.78%	21.07%	**11.26%**	**28.12%**

Table 3. Speech recognition performance on the Wall Street Journal Corpus

To study the effects of data augmentation, we train a new seq-to-seq model with the same architecture and training procedure as the baseline. However this time, in each epoch, we randomly select 40% of the training utterances and apply the train RIRs to them (in our previous experiments we had observed that 40% augmentation results in the best validation performance).

For the enhancer models, λ in Equation 1 was tuned over the dev set by doing a logarithmic sweep in [0.01, 10]. $\lambda = 1$ gave the best performance.

We use Algorithm 1 to train the WGAN enhancer. The clipping parameter was 0.05 and ε was random normal with 0.001 standard deviation. We found that having a schedule for n_{critic} was crucial. Namely, we do not update the encoder parameters with WGAN gradients for the first 3000 steps. Then, we use the normal $n_{\text{critic}} = 5$. We hypothesize that the initial encoder embedding is of poor quality and encouraging invariance at this stage through the critic gradients significantly hinders seq-to-seq training.

5. RESULTS

We present results in Table 3. All of the evaluations were performed using greedy decoding and no language models. To provide context, our near-field result is comparable to the 18.6% WER of [6] obtained with language model beam decoding with 200 beam size. We can see that a seq-to-seq model trained only on near-field audio data performs extremely poorly on far-field audio. This suggests that it is non-trivial for an ASR model to generalize from homogeneous near-field audio to far-field audio.

To overcome this, we train a stronger baseline with simulated far-field audio examples. This model had the same architecture but 40% of the examples that the model was trained on were convolved with a randomly chosen room impulse response during training. We can see from Table 3 that simple data augmentation can significantly improve performance on far-field audio without compromising the performance on near-field audio, implying that seq-to-seq models have a strong ability to learn from far-field examples.

Even with data augmentation, however, there is still a large gap between the WERs on near-field and far-field test sets. The bottom two rows of Table 3 show the performance of the methods introduced in this paper on the same test sets. An L^1-distance penalty can lower the test set WER by 1.32% absolute. Using a GAN enhancer can reduce the WER by an additional 1.07%. Overall, the gap between near-field and far-field performance decreases by almost 27% compared to the model that only uses data augmentation.

An additional benefit of our methods is that the L^1-distance penalty and GAN loss function act as regularizers which reduce generalization error on near field data. The enhancer models have lower WERs even on near-field data compared to the baseline models.

6. CONCLUSION

We introduced a GAN-based framework to train robust ASR models in a scalable, data-driven way, and showed that inducing invariance at the encoder embedding level considerably improves the recognition of simulated far-field speech by vanilla seq-to-seq models. This method has effectively imbued the seq-to-seq encoder with a far-field front-end. We anticipate that coupling the new framework with specialized trainable front-ends, such as WPE, would enhance robustness even more significantly.

7. REFERENCES

[1] Dario Amodei et al., "Deep speech 2 : End-to-end speech recognition in english and mandarin," in *Proceedings of The 33rd International Conference on Machine Learning*, New York, New York, USA, 20–22 Jun 2016, vol. 48 of *Proceedings of Machine Learning Research*, pp. 173–182, PMLR.

[2] Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Mike Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig, "Achieving human parity in conversational speech recognition," *arXiv preprint arXiv:1610.05256*, 2016.

[3] Zixing Zhang, Jürgen T. Geiger, Jouni Pohjalainen, Amr El-Desoky Mousa, and Björn W. Schuller, "Deep learning for environmentally robust speech recognition: An overview of recent developments," *CoRR*, vol. abs/1705.10874, 2017.

[4] M. Benzeghiba, R. De Mori, O. Deroo, Stephane Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens, "Automatic speech recognition and speech variability: A review," *Speech Commun.*, vol. 49, no. 10-11, pp. 763–786, Oct. 2007.

[5] 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 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., pp. 2672–2680. Curran Associates, Inc., 2014.

[6] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio, "End-to-end attention-based large vocabulary speech recognition," vol. abs/1508.04395, 2015, http://arxiv.org/abs/1508.04395.

[7] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B. H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, Sept 2010.

[8] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind mimo impulse response shortening," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 10, pp. 2707–2720, Dec 2012.

[9] Bo Li, Tara Sainath, Arun Narayanan, Joe Caroselli, Michiel Bacchiani, Ananya Misra, Izhak Shafran, Hasim Sak, Golan Pundak, Kean Chin, Khe Chai Sim, Ron J. Weiss, Kevin Wilson, Ehsan Variani, Chanwoo Kim, Olivier Siohan, Mitchel Weintraub, Erik McDermott, Rick Rose, and Matt Shannon, "Acoustic modeling for google home," 2017.

[10] 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," 2017, pp. 379–383.

[11] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, pp. 3371–3408, Dec. 2010.

[12] Masato Mimura, Shinsuke Sakai, and Tatsuya Kawahara, "Reverberant speech recognition combining deep neural networks and deep autoencoders augmented with a phone-class feature," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, pp. 62, Jul 2015.

[13] Mirco Ravanelli, Philemon Brakel, Maurizio Omologo, and Yoshua Bengio, "A network of deep neural networks for distant speech recognition," *CoRR*, vol. abs/1703.08002, 2017.

[14] Raia Hadsell, Sumit Chopra, and Yann Lecun, "Dimensionality reduction by learning an invariant mapping," in *In Proc. Computer Vision and Pattern Recognition Conference (CVPR06*. 2006, IEEE Press.

[15] Roland Thiollire, Ewan Dunbar, Gabriel Synnaeve, Maarten Versteegh, and Emmanuel Dupoux, "A hybrid dynamic time warping-deep neural network architecture for unsupervised acoustic modeling.," in *INTERSPEECH*. 2015, pp. 3179–3183, ISCA.

[16] Konstantinos Bousmalis, Alex Irpan, Paul Wohlhart, Yunfei Bai, Matthew Kelcey, Mrinal Kalakrishnan, Laura Downs, Julian Ibarz, Peter Pastor, Kurt Konolige, et al., "Using simulation and domain adaptation to improve efficiency of deep robotic grasping," *arXiv preprint arXiv:1709.07857*, 2017.

[17] Martin Arjovsky, Soumith Chintala, and Léon Bottou, "Wasserstein generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning*, Doina Precup and Yee Whye Teh, Eds. 06–11 Aug 2017, vol. 70 of *Proceedings of Machine Learning Research*, pp. 214–223, PMLR.

[18] Santiago Pascual, Antonio Bonafonte, and Joan Serrà, "SEGAN: speech enhancement generative adversarial network," *CoRR*, vol. abs/1703.09452, 2017.

[19] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael Seltzer, and Sanjeev Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," *ICASSP 2017 (submitted)*, 2017.

[20] Eric Battenberg, Jitong Chen, Rewon Child, Adam Coates, Yashesh Gaur, Yi Li, Hairong Liu, Sanjeev Satheesh, David Seetapun, Anuroop Sriram, and Zhenyao Zhu, "Exploring neural transducers for end-to-end speech recognition," *CoRR*, vol. abs/1707.07413, 2017.

[21] Andrew L. Maas, Awni Y. Hannun, and Andrew Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," 2013.

[22] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *CoRR*, vol. abs/1502.03167, 2015.