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 strong assumptions, 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, robust speech recognition, 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.

Traditional robust ASR literature models the noisy process from first principles, but these hand-engineered front-ends [3, 4] do not generalize well on other modalities in practice. These problems can be solved by training models on a large volume of labeled data with these effects. However, for non-stationary processes, such as accents, high fidelity data augmentation is infeasible, and in general, high quality labeled datasets are expensive and time-consuming to collect. Data driven approaches without strong supervision are ideal for scalable robust training, because the effects can be modeled from the unsupervised data itself.

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. Section 2 documents related work. 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. Section 6 concludes this work.

2. RELATED WORK

Robust ASR has fairly deep roots in signal processing, but these traditional approaches [3] typically have strong priors that make it difficult to incorporate new effects. Methods like the denoising autoencoder (DAE) [7] on the other hand can learn to recover the original audio from a corresponding noisy version [8] without domain knowledge. Such methods have been shown to improve perceptual quality of the produced speech and to a certain extent the final ASR performance [9]. Even though gain in ASR performance from DAE is rather limited given its amount of computation, its data driven nature is very appealing.

The problem with autoencoders is that it attempts to reconstruct all aspects of the original audio, including many features that are not important for the end task, 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.

This problem can be alleviated by training models with an auxiliary objective that measures sensitivity to changes in the bottleneck layer. Intuitively, we want the ASR model to learn robust representations suitable for the end task automatically from data. One simple such heuristic is the embedding distance between clean and noisy speech, but minimizing this requires paired training audio and alignments. Variable speed can make alignments even trickier; expensive methods like dynamic time warping [10] may be needed.

* equal contribution.

The domain adversarial neural network (DANN) [11] solves this problem by minimizing the domain divergence. This involves introducing a secondary task of classifying between source and target domains, and training the feature extractors to produce embeddings that are indistinguishable by the classifier. Because the objective can be computed from a scalar summary of the input and the domain label, such methods can leverage unaligned, unpaired, and unsupervised data. [12] showed this technique indeed improves ASR robustness to ambient noise.

Similarly, the generative adversarial network (GAN) [5] where the generator synthesizes increasingly more realistic data in attempt to fool a competing disctiminator can be used to enable robust ASR. [13] treats encoding speech as a generative process and achieves invariance by confusing the domain critic. Multi-task adversarial learning certainly enhances ASR robustness in a data-driven way, but existing work is applied to a more traditional hybrid speech recognition pipeline. They are unable to take advantage of more recent end-to-end frameworks like sequence-to-sequence models with attention [6].

In general, adversarial methods are quite difficult to train. [14] explains that the Jensen-Shannon divergence's strong topology makes gradients not always useful. Instead, the Wasserstein distance also known as the Earth-Mover distance was proposed to mitigate unstable training. This method was shown to make GAN training more robust to architectural choices and other prior art.

3. ROBUST ASR

3.1. Encoder distance enhancer

As motivated in Section 2, inducing invariant representations to noise via multitask learning naturally improves ASR robustness. The end task objective ensures that only relevant features to recognition are learned, while a sensitivity measure encourages perturbed representations to be similar to those of clean samples. We validate this idea with a straightforward heuristic that measures the distance between clean and noisy encoder embeddings.

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

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

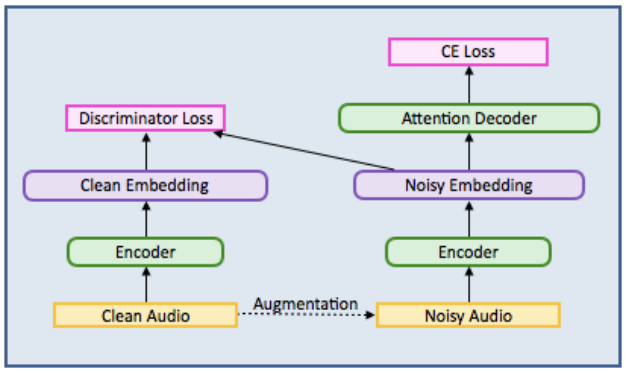

Fig. 1. Architecture of the enhancer models introduced in this paper. The discriminator loss can be L^1-distance or WGAN loss. The entire model is trained end-to-end using both the discriminator loss and the cross-entropy loss. We use RIR convolution to simulate far-field audio. It's also possible to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

3.2. WGAN 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 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, we use the Wasserstein GAN (WGAN) [14]. 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

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

where $\mathcal{W}$ is a set of clipped weights to ensure the duality holds up to a constant multiple [14].

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 of the generator helps improve training. This is most likely because there are a limited number of impulse responses and a small perturbation prevents the discriminator from easily memorizing all augmentation patterns. Also, weights in the parameter set $\mathcal{W}$ should be clipped to ensure the duality of (2) holds up to a constant multiple [14]. 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, m', the batch sizes.

1 **while** θ has not converged **do**
2 **for** $t = 1, \ldots, n_{\text{critic}}$ **do**
3 Sample $\{(x^{(i)}, y^{(i)}) \sim \mathcal{D}\}_{i=1}^{m}$ a batch of labeled speech data.
4 Sample $\{\widetilde{x}^{(i)}\}_{i=1}^{m'}$ by augmentation or from a noisy dataset
5 Sample noise $\{\varepsilon^{(i)}\}_{i=1}^{m'}$.
6 $g_\theta \leftarrow \nabla_\theta \left[\frac{1}{m}\sum_{i=1}^{m} H(h_\theta(g_\theta(x^{(i)})), y^{(i)})\right]$
7 $\theta \leftarrow \theta - \text{Adam}(\theta, g_\theta)$
8 $g_w \leftarrow \nabla_w \left[\frac{1}{m}\sum_{i=1}^{m} f_w(g_\theta(x^{(i)})) - \frac{1}{m'}\sum_{i=1}^{m'} f_w(g_\theta(\widetilde{x}^{(i)} + \varepsilon^{(i)}))\right]$
9 $w \leftarrow w + \text{RMSProp}(w, g_w)$
10 $w \leftarrow \text{clip}(w, -c, c)$
11 **end**
12 Sample $\{(x^{(i)}, y^{(i)}) \sim \mathcal{D}\}_{i=1}^{m}$ a batch of labeled speech data.
13 Sample $\{\widetilde{x}^{(i)}\}_{i=1}^{m'}$ by augmentation or from a noisy dataset
14 Sample noise $\{\varepsilon^{(i)}\}_{i=1}^{m'}$.
15 $g_\theta \leftarrow \nabla_\theta \left[\frac{1}{m}\sum_{i=1}^{m} H(h_\theta(g_\theta(x^{(i)}), y^{(i)}) - \lambda\frac{1}{m'}\sum_{i=1}^{m'} f_w(g_\theta(\widetilde{x}^{(i)} + \varepsilon^{(i)}))\right]$
16 $\theta \leftarrow \theta - \text{Adam}(\theta, g_\theta)$
17 **end**

Algorithm 1: WGAN enhancer training. Adam and RMSProp were used to update the seq-to-seq and critic models. If data augmentation is possible, training the seq-to-seq model with augmentation in lines 6 and 15 can further improve results.

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 [15], 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 [16].

The discriminator network of the WGAN enhancer is described in Table 2. All convolutional layers use leaky ReLU activation [17] with 0.2 slope for the leak, and batch normalization [18].

Bidirectional GRU (dimension = 256, batch norm)
Bidirectional GRU (dimension = 256, batch norm)
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)
Mel spectrogram

Table 1. Encoder architecture (feature)×(time).

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

Table 2. Critic architecture (feature)×(time).

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

4.3. Training

To establish a baseline, in the first experiment, we trained a simple attention based seq-to-seq model [6]. All the seq-to-seq networks in our experiments with the exception of WGAN critic were trained using the Adam optimizer. We evaluate all models on both clean and far-field test sets.

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

All of the evaluations in Table 3 were performed using greedy decoding. To provide context, our near-field result is comparable to the 18.6% word error rates (WER) of [6] obtained from 200 beam decoding. Our results show that seq-to-seq models trained only on near-field data perform extremely poorly on far-field speech. This suggests that it is non-trivial for a seq-to-seq ASR model to generalize from homogeneous near-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 too much, implying that our seq-to-seq model is capable of modeling far-field speech to a certain extent.

Even with data augmentation, however, there is still a fairly large gap between near- and far-field test performance. The L^1-distance penalty lowers the test set WER by 1.32% absolute. GAN enhancer reduces the error rate by an additional 1.07%. Overall, the gap decreases by almost 27% relative compared to the model that only uses data augmentation.

A benefit of multi-task learning that constrains the encoder space is that the new objectives act as regularizers and improve near-field performance as well. Models trained only with far-field augmentation suffer a slight deterioration on near-field speech, as the support of input distribution to be modeled has increased but there is no mechanism to learn an efficient representation that exploits commonalities in the input. We also report that adding Gaussian noise didn't considerably help the encoder distance model, although there was some initial improvement during training. The WGAN enhancer model most likely benefited from input perturbations because it alleviates critic overfitting.

In our experiments, the encoder was never quite able to produce fully indistinguishable embeddings that can fool the discriminator. We suspect that the encoder's ability to generate invariant representations is limited by the lack of a specialized front-end or more flexible layer that can fully remove far-field effects. Grid LSTMs have been shown to better model frequency variations [19] than GRU or LSTM, and may further close the gap.

6. CONCLUSION

We showed that inducing invariance to noise at the encoder is a promising way to improve speech recognition robustness, and used the Wasserstein distance to train a robust seq-to-seq ASR model. Because this loss does not require alignments, the proposed method can be applied to problems where there are unpaired and unsupervised audio data. Although we were not able to completely close the performance gap between near- and far-field speech, we anticipate that augmenting our framework with hand-engineered or more expressive layers will significantly enhance robustness.

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] 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.

[8] 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.

[9] 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.

[10] 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.

[11] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2096–2030, Jan. 2016.

[12] Yusuke Shinohara, "Adversarial multi-task learning of deep neural networks for robust speech recognition," in *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, 2016, pp. 2369–2372.

[13] Dmitriy Serdyuk, Kartik Audhkhasi, Philemon Brakel, Bhuvana Ramabhadran, Samuel Thomas, and Yoshua Bengio, "Invariant representations for noisy speech recognition," *CoRR*, vol. abs/1612.01928, 2016.

[14] 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.

[15] 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.

[16] Eric Battenberg et al., "Exploring neural transducers for end-to-end speech recognition," .

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

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

[19] Bo Li et al., "Acoustic modeling for google home," in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, 2017, pp. 399–403.